Modular network comparison with application to labour flow networks

Mattie Susan Landman
Wolfson College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2022

Acknowledgements

I would like to take this opportunity to thank my supervisors Neave O'Clery and Peter Grindrod for their unwavering support and exceptional guidance over the past couple of years. Furthermore, I wish to thank all the members of the O'Clery research group for their feedback and insights after many presentations on this work.

This work has also greatly benefited from collaboration with Daniel Straulino, Stephen Kinsella and Sanna Ojanperä. I would like to thank Alexandre Bovet, Samira Barzin, John Fitzgerald and Nils Rochowicz for their helpful insights and/or proofreading of this thesis.

I also wish to express my gratitude to the Skye Foundation, the Oppenheimer Memorial Trust and the Oxford Mathematical Institute for their financial support.

Many thanks also goes to my parents, Gys and Martie Landman, for their love, support and continual investment in my education. I am also deeply thankful for the St Ebbe's Thesis community for their fellowship throughout my time in Oxford. I would also like to thank my husband, Christiaan van der Walt, for always standing by my side and believing in me.

Finally, this work is in honour of my Lord Jesus Christ. Soli Deo Gloria!

Abstract

Evolutionary economic geography emphasises the role of locally embedded skills and knowledge in regional economic development processes. A region grows into new economic activities by combining locally available skills. This is typically modelled as a branching process on an inter-industry labour flow network, where a region's probability of entering an industry is a function of the related industries available. In this thesis, we develop new methodologies that support the analysis of these labour flow network models.

In the first part of this thesis, we develop new metrics to compare the modular structure of different networks. The motivating application is the comparison of different countries' inter-industry labour flow networks. More specifically, we present a novel approach for the global modular comparison of two node-aligned networks. The measure quantifies the difference in the quality of the partition of two graphs by projecting the community structure of each graph onto the other. The technique is advantageous as it captures the influence of the underlying network structure and applies to a broad range of networks as it is agnostic to the quality function adopted.

Building on this framework, we also present a fine-grained modular comparison technique, which compares the modular structure of a single community across two networks. The measure compares the expected escape probability of a random walker, initially starting in the community, across the two networks. The measure is advantageous as it captures the impact of edge density and connectivity on the modular structure of the community.

In the second part of the thesis, we investigate two economic questions. First, we investigate whether the presence of related multinational enterprises (MNE) results in knowledge spillovers that foster new related domestic industries in Irish regions. We propose a new dynamic-based

cohesion measure for an inter-industry labour flow network that captures the impact of higher-order linkages between an industry and the industrial portfolio of a region. We find that industries containing both domestic and MNE firms enhance the entry and survival of related domestic industries within a region, while industries dominated by MNEs have the opposite effect.

Finally, we consider the indirect impact of an affirmative action policy on inter-industry gendered labour mobility patterns in South Africa. By harnessing a regression discontinuity design with an entropy-based centrality measure, we investigate how the policy has impacted the sectoral diversity of new employees in firms. We find that the policy widened the sectoral diversity of female workers, particularly the related variety of skilled female workers. However, unlike results in the literature, this did not translate into any significant impact on firms' performance.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Outline of papers 6

1.3 Organization of the thesis 6

2 Preliminaries **8**

2.1 Notation 8

2.2 Linear dynamics on a graph 9

2.3 Graph models 10

2.3.1 Random graph models 10

2.3.2 Scale-free graph models 11

2.3.3 Stochastic block models 12

2.4 Community detection 12

2.4.1 Modularity 13

2.4.2 Markov stability 14

2.4.3 The Louvain algorithm 15

2.5 Skill-relatedness network (SRN) 16

3 Global modular comparison **19**

3.1 Introduction 19

3.2 Literature review 20

3.2.1 Community detection 20

3.2.2 Network comparison methods 22

3.2.3 Inter-industry labour flow networks 25

3.3 Results 26

3.3.1 A bi-directional distance metric 26

3.3.2 Synthetic networks 30

3.3.3 Application to inter-industry labour flow networks 33

3.3.4 Different optimization functions . . . 36
3.3.5 Multi-scale optimization functions . . . 38
3.3.6 Application to inter-industry labour flow networks (cont.) . . . 42
3.4 Discussion . . . 44

4 Single community comparison **47**
4.1 Introduction . . . 47
4.2 Terminology: edge density and connectivity . . . 50
4.3 Literature . . . 52
4.3.1 Adopting a network comparison technique . . . 52
4.3.2 Adopting a community quality function . . . 54
4.3.2.1 Decomposing a global quality function . . . 54
4.3.2.2 Adopting a local quality function . . . 55
4.3.3 The use of dynamics to explain network structure . . . 57
4.4 Methodology . . . 60
4.4.1 The retention function . . . 60
4.4.2 Properties of the retention function . . . 62
4.4.3 The eigenvalue decomposition of the retention function . . . 64
4.5 Results . . . 64
4.5.1 The maximum retention distance . . . 65
4.5.1.1 A toy example . . . 66
4.5.1.2 Understanding the shape of the curve . . . 68
4.5.1.3 Distance confidence interval . . . 70
4.5.2 Properties of the maximum retention distance . . . 70
4.5.2.1 Capturing edge density and connectivity . . . 71
4.5.2.2 Adherence to normal network distance behaviour . . . 75
4.5.3 Maximum retention time . . . 77
4.5.3.1 Maximum retention time for a special case community 77
4.5.3.2 Maximum retention time for general community structure . . . 79
4.5.4 Application to inter-industry labour flow networks . . . 81
4.5.4.1 Comparing inter-industry labour flows across European countries . . . 81
4.5.4.2 Comparing gendered inter-industry labour flows . . . 85
4.6 Conclusion . . . 88
4.6.1 Limitations of the maximum retention distance . . . 89

4.6.2 Future Work . . . 91

5 An Application: The Irish Labour Flow Network 93

5.1 Introduction . . . 93
5.2 Literature . . . 96
5.2.1 MNEs as agents of structural change . . . 96
5.2.2 Regional resilience . . . 98
5.2.3 Industrial cohesion . . . 100
5.3 Data and definitions . . . 103
5.3.1 Industry data . . . 103
5.3.2 Industry presences, entries and exits . . . 103
5.3.3 Skill-relatedness matrix . . . 106
5.4 Measuring the cohesiveness of an industry . . . 108
5.4.1 Weighted closeness . . . 108
5.4.2 Strategic closeness . . . 108
5.4.3 Cohesion to domestic and MNE industries . . . 111
5.4.4 Correlation analysis . . . 112
5.5 Econometric framework . . . 112
5.6 Results . . . 114
5.6.1 Domestic industry entrance . . . 114
5.6.2 Domestic industry exit . . . 116
5.7 Conclusions and policy implications . . . 120
5.7.1 Policy implications . . . 121
5.7.2 Limitations and future work . . . 122

6 An Application: The South African Labour Flow Network 124

6.1 Introduction . . . 124
6.2 Literature . . . 127
6.2.1 Employment Equity Act . . . 127
6.2.2 Inter-industry labour mobility and skill diversity . . . 129
6.2.3 The evaluation of labour market regulations . . . 132
6.3 Data . . . 134
6.4 Network and metric construction . . . 135
6.4.1 Network construction . . . 135
6.4.2 Constructing a new industry classification . . . 136
6.4.3 Constructing the various inflow diversity measures . . . 140
6.4.3.1 Total inflow variety . . . 141

6.4.3.2 Related inflow variety . . . 142
6.4.3.3 Unrelated inflow variety . . . 143
6.5 Methodology . . . 143
6.5.1 Firm-level regression model . . . 143
6.5.2 Firm-level regression discontinuity design . . . 144
6.5.3 Industry-level regression model . . . 146
6.6 Results . . . 147
6.6.1 The impact of labour inflow diversity on firm performance . . . 147
6.6.2 Impact of the EE Act on firm labour inflow diversity . . . 149
6.6.3 Impact of the EE Act on industry labour inflow diversity . . . 152
6.7 Conclusion . . . 155
6.7.1 Limitations and future work . . . 156
6.7.2 Policy implications . . . 158

7 Conclusion **160**
7.1 Future work . . . 163

A SI: Global modular comparison **166**
A.1 Technical details regarding SBM construction . . . 166

B SI: Single-community comparison **168**
B.0.1 The evaluation of other summary statistics . . . 168
B.0.1.1 The retention distance at time = 1 . . . 168
B.0.1.2 Comparing the largest eigenvalue . . . 169
B.0.1.3 Taking the integral . . . 171

C SI: Ireland **174**
C.1 Data descriptive . . . 174
C.2 Extended economic model results . . . 175

Bibliography **178**

List of Figures

3.1 The comparison of toy networks using the Jaccard-, the Spectral- and NMI distance 23
3.2 The comparison of toy networks using the BiDir distance 28
3.3 The comparison of various families of SBMs using the BiDir distance 31
3.4 The comparison of four countries' SRNs using the BiDir distance . . . 35
3.5 An illustration of how the BiDir distance varies when using different optimisation functions 37
3.6 The comparison of toy networks with nested modular structure using the BiDir distance with a multi-scale optimisation function 40
3.7 The comparison of the Irish and German SRN using the BiDir distance with a multi-scale optimisation function 43

4.1 An illustration of a community's edge density and connectivity 50
4.2 The comparison of a single community across a set of toy networks using metrics in the literature 53
4.3 An illustration of the Q matrix of a community in a toy network . . . 61
4.4 An illustration of the workings of the maximum retention distance . . 67
4.5 The different shapes of the retention distance function 69
4.6 The impact of a change in community edge density on the maximum retention distance 72
4.7 The impact of a change in community connectivity on the maximum retention distance 73
4.8 The impact of edge density and connectivity on the maximum retention time 80
4.9 The comparison of the modular structure of three communities in the Irish SRN to their structure in different European countries' SRNs . . 83
4.10 The comparison of the modular structure of each of the communities in the Irish SRN to their structure in other European countries' SRNs 85

4.11 The maximum retention distance between the modular structure of different sectors in the South African male and female SRNs 87

5.1 The degree of structural change shown through industry entry and exit in Ireland 105
5.2 The number of new domestic industry entries into industries in which only MNEs are active in all regions in Ireland within the 2006-2019 period. 106
5.3 The percentage of MNE employment within each industry visualised on the Irish SRN 107
5.4 A toy network showing the workings of the weighted closeness and strategic closeness cohesion metrics 111

6.1 The division of firms within our dataset into a treatment and control group 135
6.2 Toy examples showing the limitations of the official South African industry classification 137
6.3 A visualisation of the South African SRN 139
6.4 A regression discontinuity plot showing the impact of the EE Act on the male and female inflow variety 150
6.5 A regression discontinuity plot showing the impact of the EE Act on the related and unrelated variety of newly hired skilled male and female workers 151
6.6 A regression discontinuity plot showing the impact of the EE Act on the change in labour productivity of firms who hired skilled employees 152
6.7 The relationship between the percentage of male employment and the female inflow diversity for firms who comply and those who are exempt from the EE act 153
6.8 The relationship between the percentage of male employment and the male inflow diversity for firms who comply and those who are exempt from the EE act 154

B.1 Comparing networks using different summary statistics for the retention distance function 169
B.2 Comparing a community across toy networks using the integral and the maximum to summarise the retention distance function 173

Chapter 1

Introduction

1.1 Background

The tenth sustainable development goal, set by the United Nations in 2015, is to reduce inequality within and across countries. Hence, a significant challenge facing governments worldwide is to reduce the unequal distribution of economic activities, wealth and opportunities within their countries. Most countries typically contain a few large thriving cities - which largely contribute to the country's economic growth. At the same time, their more peripheral towns and smaller cities experience much slower growth or even economic decline. These regions are economically falling further and further behind. Therefore, fostering economic activities and transforming these left-behind regions and cities is an economic imperative. However, this is no easy task.

Foundational theories on regional economic development processes, emerging from the fields of economic complexity and evolutionary economic geography, focus on the role of locally embedded skills and know-how [150]. From this perspective, regions or cities grow by combining existing skills and knowledge to create new economic activities [101, 104]. This is because it is costly to develop new economic activities that require locally unavailable skills. Hence, the industrial diversification of a region is modelled as a path-dependent process, where a region's probability of entering an industry is a function of the related skills available [83, 103]. A region's economic and industrial development is therefore constrained by the skills and knowledge embedded in its workforce.

However, skills and know-how are inherently slippery concepts and very tricky to measure. Several methods have been developed to infer the presence of skills and the skill-distance between economic activities (see [103] for a review). Many of these

rely on network science techniques. More specifically, the construction and analysis of inter-industry labour flow networks.

A common approach is to model the underlying economic landscape as a network where nodes represent industries and edges as the degree of skill overlap between related industries. The skill overlap is inferred by investigating the degree of labour flow between the corresponding industries. This assumes that if skilled individuals find alternative employment in another industry, the production processes of the old and new industries draw on similar skills and are thereby related. The degree of labour flows, compared to what we would expect at random considering the size of the industries, then proxies the degree of skill-overlap between the two industries. These inter-industry labour flow networks are typically referred to as skill-relatedness networks (SRN), or the industry space [149].

Typically, the industrial basket of a region is projected onto its country's SRN. How the region is positioned within the network then determines its potential development paths; regions positioned at the network's core have many diversification opportunities compared to those positioned on the periphery. Governments need to consider the current industrial basket of a region and where it is positioned in the SRN when choosing a future development path. Since specific paths will foster (or hinder) future development.

These SRNs are characterised by modular structure [161]. In other words, the networks contain communities defined as groups of densely inter-connected industries. A community in these networks represents an industrial cluster or skill-basin where workers can more easily transition between industries within the community than those outside it. These communities can also be seen as labour pools for both workers and firms. As these clusters allow for knowledge and skill sharing, they promote firm learning [131, 41], and are crucial to innovation [179, 60]. However, these modules can also constrain industrial diversification opportunities, as it is harder for regions to move and diversify outside these communities. It is therefore vital to understand and compare these structures across different economies to design good industrial diversification strategies and related policies.

Despite the growing and widespread adoption of network science techniques to study regional industrial diversification, there still exists gaps in the network science toolbox when it comes to answering some key questions. For example, *how do we compare the modular structure of these SRNs across different countries?*

In the network science literature, there exists a wide variety of community detection methods [186]. However, the comparison of modular structure across differ-

ent networks has been less well studied. The current state-of-the-art method, the Normalised Mutual Information [57], is an information theoretic-based metric that compares sets. It does not include information on the underlying structure of the two networks and therefore loses key information when comparing communities. In the first part of this thesis, using the comparison of different countries' SRNs as our primary motivation, we develop two novel techniques for the pairwise comparison of modular structure across node-aligned networks.

First, we present a network comparison technique that compares the *global* modular structure of a network: the *Bi-directional distance*. This distance quantifies the difference between two graphs' modular structures by swapping their partitions (representing their modular structure) and evaluating how well the new partition describes the modular structure of the graph in relation to its own original partition. By quantifying the quality of the two partitions on both of the graphs, the measure captures differences in the underlying connectivity of both graphs. It can also detect when one network in the comparison is a sparser or nested version of the other. Another advantage of the technique is that it can adopt various quality functions, making it widely applicable to compare many different types of networks.

This framework only evaluates differences in the global modular structure and is unable to show which parts of the two networks are more similar or different. To address this problem, we develop a more fine-grained or *local* network comparison technique that compares the modular structure of a single community across two node-aligned networks, the *Maximum retention distance*. In the literature, only ad hoc methods have been applied to this problem, and none consider the impact of both inner- and outer-community edge density and connectivity simultaneously.

The Maximum retention distance uses a dynamic approach and evaluates the largest difference in the expected escape probability of a random walker, initially starting in the community, across the two networks. The measure is advantageous as it is able to capture the modular structure of a single community without the use of a null model - thereby making it suitable for comparison across different networks. Furthermore, the measure captures both the impact of edge density and connectivity on the modular structure of both graphs.

In the second part of this thesis, we apply network techniques more broadly to labour flow networks to investigate two related economic questions.

First, consider that some regions' industrial baskets and their position within their country's SRN provide limited industrial diversification and growth opportunities. Governments have therefore started to consider whether they can 'import' skills

and know-how unavailable locally to these regions and use them as 'stepping-stones' to enhance their industrial diversification potential. One way to do this is to attract multinational enterprises (MNEs). The literature has shown that MNEs can benefit their host economy through transferring financial resources, creating new market opportunities and enhancing the productivity and innovation of co-located domestic firms [52, 107]. However, this effect may not always materialise as MNEs actively protect their skills and knowledge to prevent competition or the capability gap between MNEs and domestic firms may be too large for learning to occur [4, 26]. Hence, it remains unclear whether the skills introduced into a region by MNEs enhance the region's industrial diversification opportunities. In other words, *do MNE industries enhance the entry of related domestic industries?*

Here, we investigate whether the presence of related MNEs stimulates the entry and survival of new domestic export industries through knowledge spillovers in Irish regions. To do this, we construct a novel dynamic-based centrality measure, *the strategic closeness* for the Irish labour flow network. The measure captures the potential for higher-order linkages and knowledge spillovers to occur between a focal industry and other industries present in the region. The measure is modelled using a 2-step random walker process, where the initial state of the walker corresponds to the current industrial portfolio of the region on the labour flow network. The probability distribution of the walker after two steps is then used as a centrality measure of the various industries. The measure improves upon current cohesion measures as it not only considers directly related industries but also their connectivity to each other and other industries in the region. It therefore, captures the impact of higher-order linkages from a cluster of related industries.

We find that domestic industries are both more likely to enter and survive in a region if they are related to so-called 'overlapping' industries containing both domestic and MNE firms. In contrast, we find a negative impact on domestic entry and survival from cohesion to 'exclusive MNE' industries, suggesting that domestic firms cannot 'leap' and thrive in MNE-proximate industries, likely due to a know-how gap.

Besides modelling and predicting regional industrial diversification paths, labour flow networks can also be used to understand the general structure of labour markets and the patterns of labour mobility. Understanding these patterns is key as labour mobility is the primary mechanism through which knowledge diffuses within an economy and by which firm learning and innovation occur. It thereby characterises the degree of growth and expansion of both firms and industries in an economy.

Within the innovation management literature, emphasis is placed on the skill diversity of a firm's workforce, where a higher diversity is attributed to greater creativity and innovation [23]. However, if the skill diversity is too broad, communication problems can hinder innovation and reduces a firm's performance [157]. Authors have shown that newly hired employees with a prior background in an industry related to the firm's industry positively contribute to the firm's performance. However, employees from unrelated industries or the same industry either have no effect or reduce a firm's performance. This is due to either a struggle to integrate employees with too different skills or competition arising between employees with similar skills [30]. It is therefore essential to consider the skill diversity of newly hired employees to ensure that they complement the current knowledge base of a firm.

Governments often implement various hiring policies to ensure a fair and well-functioning labour market and control labour mobility levels. These include a national minimum wage, regulations around the hiring and firing of workers, bargaining powers by trade unions, unemployment protection grants and affirmative action (AA), amongst many others [75, 176]. Here, we are interested in investigating *if and how affirmative action policies influence the sectoral diversity of newly hired workers*. Although the impact of group-based affirmative action policies on labour market outcomes (such as employment representation and remuneration patterns) has been thoroughly investigated, we are the first study to consider their indirect effect on inter-industry labour mobility patterns.

More specifically, we investigate how the Employment Equity Act of 1999 has impacted the sectoral diversity of newly hired male and female workers in South Africa. We adopt a regression discontinuity design to causally evaluate the act's impact on the labour flow network structure. We exploit a cutoff created by the act's adoption legislation that requires all firms with more than 50 employees to comply. We therefore compare firms with slightly less than 50 employees (who are exempt from the act) with those with slightly more (who comply with the act). We adopt various entropy centrality measures first introduced by Frenken *et al.* [84] to quantify the total sectoral diversity and the related- and unrelated-variety of in-flowing workers for each firm in the inter-firm labour flow network.

We find that the act increased both the overall sectoral diversity of newly hired female workers, as well as the related variety of skilled female workers. Furthermore, although the literature shows that an increase in the related variety of workers enhances firm performance, we find no significant impact of the act on a firm's labour

productivity growth. Hence these act-induced increases in female related-variety did not necessarily translate into enhanced firm performance.

1.2 Outline of papers

Most of the work in this thesis, apart from Chapter 4, has been accepted for publication. Details are given below:

1. Straulino D, Landman M & O'Clery N (2021) A bi-directional approach to comparing the modular structure of networks. EPJ Data Science 10: 13.

2. Landman M, Ojanpera St , Kinsella S & O'Clery N (2022) The role of relatedness and strategic linkages between domestic and MNE sectors in regional branching and resilience. Journal of Technological Transfer.

3. Landman M & O'Clery N (2020) The impact of the Employment Equity Act on female inter-industry labour mobility and the gender wage gap in South Africa. United Nations WIDER Working Paper 2020/52.

1.3 Organization of the thesis

This thesis has a dual focus. First, we aim to develop network comparison metrics that compare the modular structure of two node-aligned networks. Although these techniques can be used to compare any two node-aligned networks, our primary motivation is the comparison of labour flow networks. In the second part, we develop and apply network techniques more broadly to labour flow networks to investigate two economic questions.

The remainder of this thesis is organised into six chapters. First, in chapter 2, we briefly define our notation and present necessary preliminaries.

In chapter 3 and 4, we focus on the development of modular network comparison techniques. In chapter 3, we develop the Bi-Directional distance for comparing the global modular structure of two node-aligned networks. Using toy models, synthetic networks and real-world labour flow networks, we illustrate the properties of the measure and how it captures the underlying network topology in the comparison.

Building on this chapter, we present the Maximum Retention distance: a dynamic-based technique to compare the modular structure of a single community across node-aligned networks. In §4.4.1, we first illustrate the properties of the retention function. In §4.5.1, we then show how we adopt this function to create our distance measure.

Finally, again using toy networks, synthetic and real-world data, we demonstrate the measure's properties and implementation.

In the second part of the thesis, our focus shifts, and we consider the broader use of network techniques to investigate two economic questions concerning labour flow networks. In chapter 5, we investigate whether the presence of MNEs has influenced the dynamics of domestic exporting industries in Ireland through knowledge spillovers. In §5.4, we propose a new dynamic-based cohesion measure that captures the impact of higher-order linkages between an industry and a cluster of related industries. In §5.5 and §5.6 , we then use this cohesion measure in a regression model to show how related MNE industries have enhanced domestic industry entry and survival.

In chapter 6, we investigate how a group-based affirmative action policy has influenced a firm's sectoral diversity of newly hired male and female workers in South Africa. This involves first constructing a new skill-based industry classification in §6.4. We then use this classification to quantify the sectoral diversity and the related and unrelated variety of newly hired workers in firms. We employ various entropy centrality measures on an inter-firm labour flow network. Finally, in §6.5 and §6.6, we combine a regression discontinuity design with these entropy centrality measures to investigate the causal impact of the act.

Finally, in chapter 7, we provide a summary of our main findings, the broader implications of our research and potential avenues for future research.

Chapter 2

Preliminaries

In this chapter, we briefly introduce some basic notation and some preliminary results in network science. We also review the construction of the skill-relatedness network (a type of inter-industry labour flow network) used extensively throughout this thesis.

2.1 Notation

In this thesis, we adopt the following mathematical notions and conventions. We denote vectors as small bold letters, for example $\mathbf{x}$, and denote the vectors of ones as $\mathbf{1}$. Matrices are denoted by capital letters such as A, where I is the identity matrix. We also use $\text{diag}(\mathbf{x})$ to denote the diagonal matrix, in which the diagonal entries are defined by the elements of the $\mathbf{x}$ vector and 0 otherwise. All vector and matrix entries are non-bold and denoted as x_i or A_{ij}.

Furthermore, we denote a *network* or *graph* by $G = (V, E)$, where V is a set of nodes and E as the set of edges, where each edge connects a node pair i and j. The network size is given by $n = |V|$, and the number of edges is given by $m = |E|$. A network is called *weighted* if each edge is associated with a weight w_{ij}. A weighted network is also called *non-negative*, if the weight of all its edges are non-negative ($w_{ij} > 0, \forall i, j$). An *undirected* network, is one in which $w_{ij} = w_{ji}\ \forall i, j$. In this thesis, we mainly focus on undirected and non-negative networks.

An adjacency matrix A can represent the structure of a network, where $A_{ij} = w_{ij}$ for each edge and $A_{ij} = 0$ otherwise. If the network is unweighted, all edges are given a weight of 1. For undirected graphs the adjacency matrix A is symmetric ($A_{ij} = A_{ji}$)). The *strength* (or *degree* in the case of an unweighted graph) of a node i is defined as $k_i = \sum_j A_{ij}$. The strength vector is given by $\mathbf{k}$ and the strength matrix by $D = \text{diag}(\mathbf{k})$. An unweighted graph in which each node has the same degree is said to be a *regular* graph.

A network is called *complete* or a *clique* if every pair of nodes are connected to each other by an edge. A pair of nodes i and j are connected by a path, if there is a sequence of adjacent nodes that connect i to j. An undirected graph is called *connected* if there is a path between any node pair, and *disconnected* otherwise.

2.2 Linear dynamics on a graph

In this thesis, we focus primarily on diffusion dynamics on a graph. This is as it is the simplest case of dynamics, and various other linear dynamics can be deduced from it. We use random walks on weighted, undirected networks[1] to model this process. Our brief introduction on the workings of a discrete random walker models on a graph is similar to that of [123].

First, we start by defining an unbiased discrete random walk process on a graph G. A random walker starts from node i at time $\tau = 0$ and jumps at time $\tau = 1$ to a neighbouring node j with probability proportional to the edge weight A_{ij}/k_i. The walker successively jumps at $\tau = 1, 2, 3, 4, \ldots$, which defines a Markov chain or random walk on the graph. The probability of the presence of the walker on each node evolves according to the following equation

$$\mathbf{x}(\tau + 1) = \mathbf{x}(\tau)D^{-1}A = \mathbf{x}(\tau)T, \tag{2.1}$$

where $\mathbf{x}(\tau)$ is a $1 \times n$ probability vector and $x_i(\tau)$ is the probability that the random walker will be at node i at time τ. As $\mathbf{x}(\tau)$ is a probability vector, $\sum_{i=1}^{N} x_i(\tau) = 1$ for any τ. Furthermore, T is the transition matrix, where T_{ij} is the probability that a random walker will jump from node i to node j in one step. T is an asymmetric matrix unless the underlying graph is regular. It is also a row-stochastic matrix, where $\sum_{j=1}^{N} T_{ij} = 1$.

Given that the walker starts with a probability distribution $\mathbf{x}(0)$, the probability distribution at time τ can also be expressed according to this initial probability distribution by substituting it into Eq (2.1), and given as:

$$\mathbf{x}(\tau) = \mathbf{x}(0)T^{\tau}. \tag{2.2}$$

A Markov chain can be composed of different types of states. An *absorbing state* is a state in which the walker cannot escape once reached. In this case, $T_{ii} = 1$ and $T_{ij} = 0$ where $i \neq j$. An *ergodic set* is a group of states where the walker can move between any of the states in the set but cannot leave the set. An absorbing state is,

[1] We only consider fixed graphs (where the underlying graph structure does not change)

therefore, an ergodic set that is composed of a single state. On the other hand, a state is called *transient* if it is not a member of an ergodic set. We consider absorbing Markov chains in Chapter 4.

Now, given that G is undirected, connected, and aperiodic[2], any initial state $\mathbf{x}(0)$ will converge to a unique stationary state π, where $\pi = \lim_{\tau \to \infty} x(\tau)$, which is the solution of the fixed-point equation $\pi = \pi T$. If the graph also contains no absorbing states, the stationary distribution is given as $\pi_i = k_i/2m$.

The discrete random walk model can easily be translated into a continuous time; an interested reader is referred to [137, 123]. However, we primarily focus on the discrete case in this thesis.

2.3 Graph models

Next, we define four network models that we use throughout this thesis: an Erdős-Rényi (ER) random graph model [73], a configuration model [142], a scale-free graph model [16], and a stochastic block model [105].

2.3.1 Random graph models

When analysing the structure of empirical networks, it is crucial to understand how their properties compare to those of an appropriate reference point. We use random graph models to produce these reference points, which act as null models or benchmark networks. The choice of null model, constructed using specific assumptions, is highly dependent on how one believes the real-world system would behaviour in a controlled or random environment. We use these models to decide whether a measured observation in empirical data is significantly different. Furthermore, as these models have known mathematical and statistical properties, we also use them to test the behaviour of newly developed measures.

A random graph model defines an ensemble of stochastically equivalent graphs where the edges in the graph are random variables obeying certain probabilistic laws. The most classic example is the Erdős-Rényi (ER) model [73], where each pair of nodes is connected to an edge with a fixed probability. The model has two parameters: the number of nodes n and the probability q that an edge exists between any pair of nodes (where $0 < q < 1$). Note that each edge is an independent Bernoulli random variable that determines the presence of an edge. The model is denoted as $\mathcal{G}(n, q)$.

[2]A graph is aperiodic if there is no integer $z > 1$ so that all cycles comprise an exact multiple of z edges.

Note that each single graph instance of this model will be different. Therefore when using these graph models, the aim is to compare the average behaviour of network metrics over many graph instances. An ER graph's expected number of edges is given as $\overline{m} = qn(n-1)/2$, and the expected degree of the nodes is $\overline{k} = q(n-1)$.

In our thesis, we employ ER graphs to generate synthetic networks with homogeneous edge connectivity.

Another model we use as a benchmark to compare our labour flow networks is the configuration model [142]. This model is similar to an ER random graph model with a further constraint: each node i has a given degree k_i. To generate a graph instance of the configuration model from a given degree distribution, we first create half-edges on each node equal to its degree. We then randomly select two half-edges to join, ensuring no self-loops or multiple edges. Again when comparing these networks, we focus on the expected values. The expected probability of each edge between node i and j is given by $\overline{A}_{ij} = k_i k_j / 2m$.

This model allows for generating an ensemble of graphs where the degree distribution can be taken from either empirical data or a family of functions (e.g. power law). This model is often adopted as it is more appropriate as a baseline for real-world networks.

2.3.2 Scale-free graph models

In this thesis, we also construct scale-free graph models. These networks represent a graph with a power law degree distribution which results in a small number of highly connected 'hub' nodes and slow convergence properties. More formally, the fraction of nodes in the network that have a degree k follows the power law $P(k) = k^{\alpha}$. Scale-free networks have been widely used to describe diverse real-world systems such as the internet [40] and neural networks [201]. Various statistical analysis has questioned these claims and argued that a fat-tailed degree distribution is more representative of real-world networks [46]. We still adopt this model in this thesis as it is a simple and well-known model to investigate the impact of heterogeneous edge connectivity within a graph.

Adopting a general approach in the literature, we construct scale-free networks using the Barabási-Albert (BA) graph model [16]. This model was proposed as a mechanism to explain the presence of power-law degree distributions in real networks. Unlike the random graph models, this model shows how a graph is assembled over time. The model adopts the mechanism of preferential attachment where nodes are

added to a graph sequentially, and the probability that a node gains a connection is proportional to its current degree k.

The BA graph is typically constructed as follows. First, we start with a complete graph of n_0 nodes. We then sequentially add a node with $m \leq n_o$ half-edges to the existing network. The probability that each half-edge connects to an existing node is proportional to its degree. We again avoid multiple edges and self-loops. We continue this process until we have attached all nodes, resulting in a graph with n nodes. This model produces a graph with a power law degree distribution. Although nodes initially have similar degrees close to m, as soon as the degree distribution becomes somewhat heterogeneous, it will reinforce this heterogeneity through the preferential attachment mechanism.

We use this BA graph model to investigate how our metrics perform on graphs with heterogeneous edge connectivity.

2.3.3 Stochastic block models

Stochastic block models (SBM) provide a good generating model for networks with a clearly defined community structure. First introduced to study the structure of social networks [105], they have been widely adopted because they provide a straightforward way of generating modular networks as well as a simple benchmark for testing the statistical significance of communities. In a classical stochastic block model, n nodes are partitioned into a set of $\tilde{n}_c$ communities $C = \{C_1, ..., C_{\tilde{n}_c}\}$. A matrix P of $\tilde{n}_c \times \tilde{n}_c$ encodes inter-community connection probabilities (the probability that nodes $v \in C_i$ and $u \in C_j$ share an edge is given by P_{ij}). Hence, we can directly specify the number of communities and the likelihood of the connection within and across communities.

We adopt SBMs to generate synthetic networks with pre-defined community structures in both Chapters 3 and 4 to show the properties of our network comparison methods.

2.4 Community detection

Many networks, particularly inter-industry labour flow networks, display modular or community structure. Loosely speaking, community structure is the partitioning of nodes within a network into groups so that nodes of the same group are densely connected, and nodes in different groups are sparsely connected. Several algorithms and methods have been developed to find communities within a network. A subset of these methods are concerned with defining a quality function that quantifies the goodness

of a given partition. Typically community detection algorithms try to find network partitions that optimise these quality functions. Although we present a review of community detection techniques in Chapter 3, here we briefly outline the workings of two quality functions that we use extensively throughout this thesis, namely the Newman-Girvan modularity function [87] and the Markov stability function [61]. We also introduce the Louvain heuristic [27] used to approximate the optimal partition in optimising various quality functions.

We define the set of communities into which G is partitioned as $\mathbb{C}_G = \{C_1, C_2, \dots C_q\}$, where C_i denotes the set of nodes within community i. The number of communities is given by $n_c = |\,\mathbb{C}_G\,|$. Furthermore, the community to which node i is assigned is denoted by $c(i)$.

2.4.1 Modularity

Modularity [87] is one of the most popular quality functions in the community detection literature. Modularity defines a *community* as a set of nodes with a higher density of edges within the set compared to the number of edges that would be expected in a null model. Typically, the configuration model is chosen as the null model, although other choices such as the uniform null model have also been considered. Recall that under the configuration model, the probability of an edge existing between node i and j is given as $p_{ij} = k_i k_j / 2m$. Each community's contribution is then normalised to obtain the modularity quality function.

In the case of an undirected and unweighted network G, with a given partition $\mathbb{C}$, the Newman-Girvan modularity is given as

$$Q(\mathbb{C}) = \frac{1}{2m} \sum_{i,j=1}^{n} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c(i), c(j)), \tag{2.3}$$

where $\delta(c(i), c(j))$ is the Kronecker delta function which is equal to 1 if $c(i) = c(j)$ and 0 otherwise. This quality function is maximised to detect the community structure of a network. Modularity maximisation has many limitations, this includes: The resolution limit and the field-of-view limit, which defines an upper and lower limit on the number of communities that can be detected, respectively; generating a large amount of nearly degenerate local maxima [89, 187]; and displaying statistically significant issues associated with an optimal partition [125]. The first two limitations are further discussed in Chapter 3.

It is important to note that many networks include community structure at multiple scales or resolutions. By this, we mean that a community with dense interactions

may consist of subsets of nodes with even denser interactions. Hence, a set of partitions define the community structure of a network at different resolutions: from a coarser partition containing few but large-sized communities to a finer partition containing many but small-sized communities. The modularity function has been adapted to obtain partitions at various scales by introducing a resolution parameter $\gamma \geq 0$. The resolution parameter scales the null model and allows the network's community structure to be investigated at multiple scales. The multi-scale modularity function [181] is defined as:

$$Q(\mathbb{C}) = \frac{1}{2m} \sum_{i,j=1}^{n} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c(i), c(j)). \tag{2.4}$$

2.4.2 Markov stability

Next, we introduce a multi-scale dynamic-based quality function: the Markov stability function [61]. This quality function is based on a simple random walk model on a graph [61, 122, 124]. The key idea behind this method is that it sets a walker to roam on a network - jumping from node to node with probability proportional to the edge weight. Suppose the walker gets trapped in a region of the network (a group of nodes) for a prolonged period, this corresponds to a group of densely connected nodes which indicates a community. Note that a *community* is defined as a group of nodes where flow gets trapped. Here, the Markov 'time' is used as the resolution parameter. Intuitively, if we let a walker roam for longer periods on the network, the walker will detect larger and larger communities. Therefore, by varying the Markov times, we detect communities on a range of scales: from many small communities to a few large communities.

Recall that an unbiased discrete random walk model is governed by Eq (2.1). Furthermore, if the graph is undirected, connected, non-ergodic and non-bipartite the dynamics converges to a unique stationary distribution given by $\boldsymbol{\pi} = \boldsymbol{k}^{'}/2m$. We encode a given partition, $\mathbb{C}$, in an $n \times c$ indicator matrix H, where $H_{ij} = 1$ if node i belongs to community j and 0 otherwise. We define the clustered auto covariance matrix of the diffusion process above as:

$$R(\tau, H) = H^{'}[\Pi T^{\tau} - \boldsymbol{\pi}^{'}\boldsymbol{\pi}]H, \tag{2.5}$$

where $\Pi = \text{diag}(\boldsymbol{\pi})$. Note that $(\Pi T^{\tau})_{ij}$ represents the probability that the random walker starting from node i ends up at node j at time τ. $(\boldsymbol{\pi}^{'}\boldsymbol{\pi})_{ij}$ represents the probability that the random walker starting at node i arrives at node j at stationarity.

Given our partition matrix H, the diagonal entries of $R(\tau, H)$ represent the probability that a random walker will remain in the community in which it started after taking τ steps. We define the stability of a partition as the sum of the diagonal elements of $R(\tau, H)$, given by:

$$r(\tau, H) = \text{Trace}(R(\tau), H). \tag{2.6}$$

As we want the walker to remain in the community in which he started, we seek a partition matrix H that satisfies:

$$H = \text{argmax}_{\hat{H}} r(\tau, \hat{H}), \tag{2.7}$$

on the set of all the possible partitions (for a given time τ).

2.4.3 The Louvain algorithm

Optimizing any of the above quality functions is an NP-hard problem, so computational heuristics are adopted to obtain a partition close to the global maximum. Although there are various heuristics within the literature (e.g. simulated annealing [115], the spectral bi-partitioning heuristic [153] and the Leiden heuristic [198]), we adopt the greedy[3] Louvain heuristic [27] in this thesis. We use this heuristic to find a local maximum for the modularity and stability quality functions.

The Louvain algorithm consists of two phases which are repeated iteratively. Initially, the network is divided into n sets - where each node is placed into its own community. In the first phase, each node is considered (in some order) and placed in a neighbouring community (or on its own) depending on what results in the largest increase in the quality function. This phase is repeated until no more moves are available to increase the quality function. Hence, a local maximum has been reached. Phase 2 then consists of constructing a reduced network (where each community is collapsed into a single node). Each node within the reduced network is partitioned again into its own community. Phase 1 is repeated on the reduced network and continues until the heuristic converges. The output of the heuristic is then an approximation to the optimal partition.

[3]A greedy heuristic performs locally optimal moves at each update until a local optimum is reached.

2.5 Skill-relatedness network (SRN)

In this thesis, we focus on the modelling and analysis of inter-industry labour flow networks. These are networks in which nodes represent industries and edges represent the number of workers who flow between the two corresponding industries. We adopt a specific variation of these networks: the skill-relatedness network (SRN). The SRN was first introduced by Neffke *et al.* [146, 149], and stems from the evolutionary economic geography literature, where it is used to model regional industrial diversification and economic growth. This section elaborates on the methodology used to construct an SRN.

First, an SRN is a network in which nodes represent industries and edges a measure of the degree of skill-overlap or skill-relatedness between the two corresponding industries. The skill-relatedness infers the degree of skill and knowledge overlap between industries by measuring the number of workers who transition between these industries. It assumes that if two industries share a high degree of skills and knowledge, workers will more freely move between them. This is because a worker's skill set from one of these industries will also be highly valued within the other industry and thereby be most likely to switch to this industry (compared to others).

There has been a range of other industry networks that have also been proposed, for the same purpose, but which differ on how they capture the 'relatedness' between industry pairs. These include, amongst others, geographic clustering or co-location of industries as a proxy for general capability overlap [102], occupational similarity as a proxy for labour sharing [79], and collaboration on patents as a proxy for knowledge sharing [110]. We focus on 'skill-relatedness' which infers the relatedness through labour mobility. This measure has been shown to outperform other relatedness measures as it is less noisy, has a larger industry coverage and has been shown to better predict regional employment growth [168].

To quantify the skill-relatedness between industries, we follow the recipe of [149]. Let L_{ij} be the number of job switches between industries i and j (during a given period). We cannot use these raw flows as a relatedness measure alone, as the value is highly dependent on the size and flow rate (i.e., the fraction of employees switching jobs) of the two corresponding industries. To account for this, we compare the observed volume of labour flows to a baseline or null model: the number of job switches that are expected at random from industries' flow rates. This is modelled using a configuration model [142]. We assume that workers switch industries with probabilities proportional to the total outflow of the industry of origin ($\sum_j L_{ij}$) and the

total inflow into the destination industry ($\sum_i L_{ij}$). The expected labour flow between industry i and j is then given as $\tilde{L}_{ij} = \frac{\sum_j L_{ij} \sum_i L_{ij}}{\sum_{ij} L_{ij}}$. The ratio of the observed to the expected flows is then:

$$\tilde{SR}_{ij} = \frac{L_{ij} \sum_{ij} L_{ij}}{\sum_j L_{ij} \sum_i L_{ij}}. \tag{2.8}$$

The values of $\tilde{SR}_{ij}$ lie between $[0, \infty]$. Values between $[1, \infty]$ indicate that the labour flows are greater than the null model, while those between $[0, 1]$ indicate that the labour flows are smaller than the null model. As the distribution is right-skewed, we transform the measure as follows:

$$\tilde{A}^{SR}_{ij} = \frac{\tilde{SR}_{ij} - 1}{\tilde{SR}_{ij} + 1}. \tag{2.9}$$

This maps the values of $\tilde{SR}$ between $[0, 1]$ to $[-1, 0]$, and those $\tilde{SR}$ values between $[1, \infty]$ onto $[0, 1]$.

Furthermore, to obtain an undirected skill-relatedness value that indicates the skill and knowledge shared between two industries, we average the $\tilde{A}^{SR}$ matrix with its transpose. Hence,

$$A^{SR}_{ij} = \frac{\tilde{A}^{SR}_{ij} + \tilde{A}^{SR}_{ji}}{2}. \tag{2.10}$$

Finally, following the common revealed advantage approach in the economic literature, we only conserve positive values of this matrix ($A^{SR}_{ij} > 0$) and hence the proportion of flows that is larger than would be expected at random. We use this matrix as our network adjacency matrix when construct the SRN. For a more detailed discussion of these methodological considerations, see [149].

Recently, O'Clery *et al.* [161] showed that these networks are characterised by modular structure. Hence, they contain communities (*i.e.*, groups of industries) within which workers switch much more easily (relative to switches between industries in different communities). The authors refer to these communities as 'skills-basins' in reference to the degree of skill and knowledge sharing within industry groupings. These skill-basins represent the industrial clusters of an economy. As we model regional industrial diversification paths on these networks, the presence of modular structure can constrains these industrial diversification processes. The modular structure of these networks therefore needs to be carefully considered and used to improve network-based models. This will allow for the development of industrial strategies that enhance regional diversification patterns and fosters economic growth.

Skill-relatedness networks have been applied to model and predict the diversification paths of regions in a large number of contexts [147, 161]. They have also

been used in various other applications, such as models for urban formality growth [159, 160], employment resilience and adaptability [68] and knowledge spillovers from MNEs [56]. In the first part of this thesis, we compare the modular structure of different countries' SRNs. In the second part, we focus on the Irish SRN, where we develop a new centrality measure that measures the cohesion of an industry to its region's industrial basket. Furthermore, we also construct the South African SRN and use it to construct a new skill-based industry classification which we then use to quantify the sectoral diversity of newly hired employees.

Chapter 3

Global modular comparison

Here we develop a network comparison measure to compare the global modular structure of two node-aligned networks. The measure is advantageous as it captures the underlying networks' topology in the comparison. We show the implementation and properties of the measure through toy networks, synthetic networks, and real-world inter-industry labour flow networks.

3.1 Introduction

In this chapter, we aim to develop a novel method for the pairwise comparison of the modular structure of two node-aligned networks. Community structure is a key network feature that is prevalent in many real-world applications. This is because communities are typically linked to the presence of some kind of higher-order organisation in a network, and is often related to functional or societal structure [169]. It is also particularly useful in obtaining a simplified description of a complex system's behaviour. Comparing the modular structure of networks is therefore vital to understand the differences between two networks. For example, it is key to understanding differences in brain functions in two neural networks [5], or understanding how partisanship has changed over time in the United States Congress [178]. Since the presence of communities in these networks lies at the core of the question of interest, it is natural to use them as the basis of comparison.

Although there is a vast community detection literature, the modular comparison of the network is less well-studied. Perhaps the most widely-used community comparison measure is *normalized mutual information* (NMI) [57]. This and other similar measures are derived from an information theoretic approach to comparing sets. By focusing exclusively on partitions, and ignoring all other features of the network (for

example, the strength of the connections between different communities), these methods lose information. Furthermore, while mutual information is symmetric, there are frequent cases where one network provides useful information about the other but not vice-versa. For example, if the communities of one network are nested inside the communities of the other (meaning that one community of the second network contains several communities of the first) then the second network provides information about the structure of the first but not the other way around.

We propose a methodology that addresses both of these issues, and that is agnostic to the use of a community detection algorithm within a large class of commonly used approaches. We call it the *bi-directional distance*, and it is based on a simple idea: we swap the partitions representing the communities of the two networks and evaluate whether the re-assigned communities are a good fit. In the sections below, we demonstrate this method for a range of toy and synthetic networks, as well as a real-world example concerning inter-industry labour flows relevant for industrial policy. These networks are the main motivation for our community comparison methodology. Since their communities have a natural interpretation as industrial clusters [161], understanding the way in which they differ provides valuable insight for industrial diversification strategies.

The chapter is structured as follows. We begin by giving a short overview of both the field of community detection and current network comparison methodologies, while also illustrating the shortcomings of current community comparison techniques. We then present our new methodology and the results of experiments on toy and synthetic networks. Second, we demonstrate the flexibility and wide applicability of our method using three distinct community detection algorithms. Third, we apply our methodology to compare the inter-industry labour flow networks of 4 European countries: Germany, Sweden, the Netherlands and Ireland. Finally, we extend our methodology to the more complex case of multi-scale community detection.

3.2 Literature review

3.2.1 Community detection

Given the diversity of community detection algorithms, it is useful to classify approaches according to the nature of the problem at hand. Schuab *et al.* [186] propose four broad categories of community detection objectives: community detection as the minimisation of some constraint function, community detection as clustering into densely connected groups, community detection aimed at identifying stochastically

equivalent nodes and community detection as the simplified description of dynamical flows on the network. We follow this categorisation to briefly introduce some of the most popular community detection algorithms.

In the first category, we find cut-based approaches which aim to minimise the number or weight of the edges between communities In other words, the constraint to minimise is calculated as the number of edges that must be deleted (or cut) to achieve the partition, *e.g.*, "ratio-cut" [97]. Note that this constraint is independent of the communities' intrinsic structure. Graph partitioning is heavily used in parallel computing, in the layout of circuits, and in the design of serial algorithms to solve partial differential equations [81].

In the second category, we find, among others, modularity optimisation [155][1] which is one of the most popular community detection algorithms. This technique partitions a network to obtain groups of nodes with high edge density within groups and low edge density across groups. A key difference with cut-based approaches is that modularity does not a priori set the number of clusters nor constrains them to be of similar size. Although not without its limitations, this approach has seen widespread use and given rise to a large number of variations and adaptations [81]. It has been used to investigate protein interaction networks, functional brain networks, and ecological networks, among other applications.

The structural category groups nodes that exhibit similar connectivity patterns, that is, they are similar if they connect to similar nodes with equal probability. One popular technique for identifying groups in this manner is the stochastic block model (SBM) [105]. This approach is based on the assumption that nodes can be divided into certain classes, and that the likelihood of a connection between two nodes is determined by their respective classes. This gives rise to a probabilistic model for the network such that, via identification of the best fitting parameters, latent groups in the network can be recovered. Originating in the social networks literature, this approach is frequently used to uncover sub-groups in a social network.

The last category deploys dynamics on networks to uncover modular structure [153, 137]. Here a community is defined as a group of nodes with similar dynamical function compared to other nodes in the rest of the network (*i.e.* a set of nodes that trap or channels flow in a specific direction and allow for a reduced description of the dynamics). In this category we find Markov stability [61] and InfoMap [183]. Markov stability, for example, exploits random walker dynamics on a network to

[1]Note that metrics can be part of various categories. For example, modularity can also be seen as a flow-based metric.

detect the presence of node communities. If a walker, which jumps from node to node with probability proportional to edge weight, gets 'trapped' in a region for a period of time, this indicates a cluster of nodes with high internal connectivity. This approach has been applied to dynamical processes on networks such as consensus and synchronisation (*e.g.*, [17, 159]).

In general, as we have seen, different types of applications call for different approaches to detecting community structure. Furthermore, no particular methodology consistently outperforms others [172]. Hence, any network comparison technique that focuses on the comparison of modular structure will need to accommodate a variety of underlying community detection techniques.

3.2.2 Network comparison methods

There is a rich literature in network comparison [70] spanning from global methods (*e.g.*, spectral distances) to more localised approaches (*e.g.*, graph editing). Since we are interested in comparing labelled networks, where each node represents a particular instance (an industry, a person etc.), we focus on methods where there is a one-to-one correspondence between the nodes of both networks.

Perhaps the simplest approach is to count the proportion of edges that need to be added and deleted to transform one graph into the other, a measure which originated in the field of information theory and is known as the Hamming distance [99]. While easy to interpret, the Hamming distance is highly sensitive to the density of the graphs, a shortfall that is addressed by the Jaccard distance [109]. The simplicity of these methods has seen them become very popular, not just in the network comparison literature, but also in the fields of information theory and machine learning [70]. Nevertheless, they are very local in scope, looking only at the direct neighbours of each node, with all deletions and insertions given equal weight regardless of changes in properties such as node importance as captured by various centrality measures. Therefore, these methods are not well suited to capture similarities at medium or large scales.

Motif based methods, which compare the frequency of certain sub-graphs (motifs) such as triangles or stars, have proved very popular in applications to protein interaction networks [141]. Despite being relatively local in scope, they are able to capture patterns beyond immediate neighbourhoods. This technique is particularly useful when the motifs represent functional components of the network. However, just as for graph editing distances, these methods fail to capture similarities in large scale connectivity or organisation patterns across networks.

On the other end of the spectrum we have spectral distances. These are global measures based on the eigenvalues of either the adjacency matrix or some version of its Laplacian. Since the eigenvalues are related to topological properties of the network, such as connectivity, they capture global features of the network. Nevertheless, these approaches are permutation invariant, and thus ignore the labels of the nodes which renders them impractical for many applications (consider a network of friendships, the spectral distance does not differentiate between the original network and one obtained by randomly exchanging individuals).

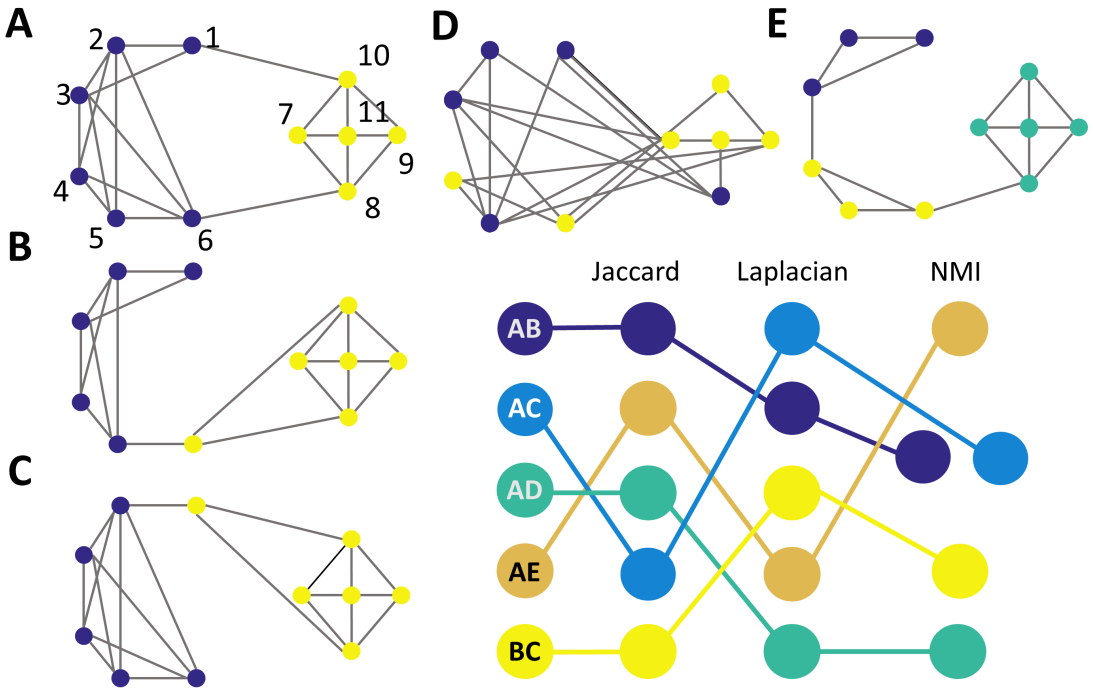


Figure 3.1: Here we show five networks whose nodes are coloured according to their communities found using modularity. We illustrate the different behaviours of three network comparison methodologies: the Jaccard distance (graph editing), the Spectral distance (eigenvalues) and the NMI distance (partitions). (F) shows the ranking of these similarity metrics for network pairs from closest (top) to furthest (bottom). Notice that standard community comparison metrics like NMI do not differentiate between pair A-B and pair A-C since they ignore the underlying structure and consider exclusively the partitions.

There are other approaches, both global and local, beyond those that we have outlined above but there is comparatively little work comparing networks at a mesoscopic or modular scale. By far the most widely-used community-based comparison measure is *Normalized Mutual Information* (NMI) [57]. NMI, which has been borrowed from the field of information theory, defines the similarity of two partitions as the mutual information of the two partitions, normalized by a combination of the

entropy of each partition. It therefore measures how easily one can infer one partition, given the other. Modifications of NMI have been proposed to address some of its limitations, namely that it can do very poorly when the number of communities in two networks differ [165, 154].

Figure 3.1 illustrates the differences between some of these key network comparison methodologies by comparing different toy networks. We construct five toy networks A-E) which differ in their community structure shown via the colouring of the nodes in the figure. Network A represents a network with two communities. Network B and C differ from network A with only a single node changing communities. However, the two displaced node has very different connectivity in their community in graph A. Network E represents a network with a nested or embedded community structure of network A. Finally, network D is constructed by randomly shuffling the edges in network A, which represents a network with no community structure. We construct and compare these toy networks as they illustrate examples where current network comparison methodologies either obtain different network rankings or fail to detect differences in the networks' modular structures.

We have chosen three popular network comparison methodologies: the Jaccard distance, the Spectral distance, and NMI. As seen in sub-figure F, while the Jaccard distance considers B to be the most similar to A, the Laplacian ranks C as more similar, and NMI indicates that E is the closest network to A. It is clear that these metrics capture different features of the networks.

We are specifically interested in comparing the community structure of the various toy networks. We observe that E is mostly embedded in A and therefore quite similar. This is picked up by NMI. Comparing B and C to A, it appears that C is 'closer' since the node that changes community is relatively peripheral to both the blue and yellow clusters. However, since NMI only uses information on partitions - and ignores the strength of the community as well as the role of the nodes - it does not differentiate when comparing A to B or C.

Although the above methodologies for comparing networks based on community structure have proven to be useful, the fact remains that there are a number of shortcomings. Perhaps the most prominent issue is that NMI and its modifications rely exclusively on partitions of network, and ignore the quality of these partitions. By reducing the community structure to a partition, these methodologies discard important information about the network [167]. There have been some efforts to address this issue [43], but they rely on correction factors (usually weighting the nodes by their degree). Furthermore, we described earlier how different community

detection algorithms differ in their objectives, but partition based methodologies do not account for these differences.

In this chapter we propose an alternative framework to comparing networks: the bi-directional distance measure. This approach presents three marked advantages: it is simple to understand and compute, it incorporates information beyond the partition of the network, and it is flexible, adapting to different community detection algorithms. Its simplicity comes from the underlying idea: for a pair of networks, we assess the fit of each node partition with respect to the other network's connectivity structure. Since the quality function takes into account the edge weights in order to evaluate the fit, we include information about the network topology. And by changing the quality function we immediately adapt to the application at hand.

3.2.3 Inter-industry labour flow networks

One of the initial motivations for the development of a new modular network comparison method was to compare the community structure of several inter-industry labour flow networks, more specifically the skill-relatedness networks (SRN) [146], introduced in §2.5, of different countries. Recall that these networks are key in the modelling of regional industrial diversification paths [83, 147]. Furthermore, a key feature of these networks is their modular structure [161], which represents industrial clusters. These communities can constrain knowledge diffusion and thereby regional diversification paths.

One might expect that these networks exhibit a near-universal structure, with skill-based industry clusters conserved across countries. On the other hand, differences in structure may illuminate potential growth paths and opportunities unseen in models based solely on data from a single country. Within the related diversification literature, the SRN has been assumed to be identical across both space and time. Neffke *et al.*[146, 149] showed that there was little variation between the SRN when constructed for East and West of Germany, as well as for Germany compared to Sweden. The authors estimated the similarity between the two networks by calculating the Spearman correlation between edges. Based on these results, various studies have used the SRN of a different country within their analysis when the SRN of the country under consideration was not available.

However, we believe that these networks may be different, particularly on a mesoscopic scale (*i.e.*, with respect to their modular structure). As an SRN is constructed from all inter-industry labour flows within a country, the network reflects the intricacies of the structure of the local labour market. We expect that the degree of

inter-industry labour flows is highly dependent on the historical economic progress of a country and its institutional labour market structures. We are interested in both similarities and differences in the modular structure of these networks. Similarities uncover universal structure in inter-industry skill-sharing and portability of networks across contexts. Differences may provide insight into hidden growth opportunities: linkages and clustering patterns present in one context may suggest potential unseen paths in another.

3.3 Results

3.3.1 A bi-directional distance metric

A large class of community detection algorithms are based on optimizing an objective or quality function Q that measures the goodness of fit of a partition according to some desired property, whether structural (for example modularity from [151]), dynamic (see [61]) or other (see [183]). We propose to compare the modular structure of two networks, say A and B, by computing the ratio of A's quality-score (Q) under B's optimal partition to its quality-score under its own optimal partition, and vice versa. In a sense, our two dimensional method tells us how well network A describes modular structure of B as well as the other way around.

More formally, let $A, B \in M^{n\times n}$ be the adjacency matrices of two networks of size n, where the networks are node-aligned (*i.e.*, there is a one-to-one correspondence of the nodes in each network given by the identity function). Let $\mathcal{P}$ be the space of partitions of $[n]$, the set of integers up to n. Without loss of generality, let $Q : M^{n\times n} \times \mathcal{P} \mapsto [0, 1]$ be an objective function [2]. We then denote

$$P_A := \arg\max_{P\in\mathcal{P}} Q(A, P),$$

$$P_B := \arg\max_{P\in\mathcal{P}} Q(B, P),$$

the partitions that maximize the objective function for each network. As we may not be able to obtain the optimal partition in polynomial time, computational heuristics are often adopted. These heuristics may only result in a near-optimal partition.

We are now ready to define our (non-symmetric) distance score:

$$d(A, B) = 1 - \frac{Q(A, P_B)}{Q(A, P_A)}. \tag{3.1}$$

[2]Any bounded quality function can be linearly transformed to match this range. For unbounded functions, it is sufficient to shift them to $[0, \infty)$ as the ratio will remain in $[0, 1]$

The ratio on the right tells us how well partition B (representing the community structure of network B) describes the modular structure of network A - a high value indicates that the near-optimal partition of B is indeed also a good partition of network A. By construction, $d(A, B)$ ranges over $[0, 1]$, and is 0 if and only if $Q(A, P_A) = Q(A, P_B)$.[3] By swapping A and B we can obtain a second distance score. In this way, for every pair of networks, we have a pair of distance scores that reflect how well their respective partitions capture each others' community structure. We propose a two dimensional distance, the *bi-directional distance* (BiDir):

$$\text{BiDir}(A, B) = (d(A, B), d(B, A)). \tag{3.2}$$

Hence, for any pair of networks A and B, we compute a two dimensional or bi-directional distance score in order to compare their modular structure.

When we calculate each of the distance coordinates, we compare the value of the quality function on an alternative partition to its value on the optimal partition. This ratio will therefore depend on the shape of the quality function but notice that both the numerator and the denominator are calculated on the same network. So even if networks A and B have very different densities (which in some cases might affect the range of the values that the quality function takes), $Q(A, P_B)$ and $Q(A, P_A)$ depend only on network A and can thus be compared.

Nevertheless, notice that if $Q(A, P_A)$ is very close to 0, indicating that network A is not modular, the ratio will not be very informative, as arbitrary partitions will produce similar scores. Therefore, our framework will be best suited to networks which obtain relatively high quality-scores. Furthermore, it is necessary that the quality function correctly discriminates good partitions from bad ones, and thus the cases in which it fails to do so will result in non-informative BiDir scores.

This method has three distinct advantages over previous community or partition comparison methods. First, it can be adapted to any community detection algorithm that is based on optimizing an objective function. Second, it goes beyond simply comparing the resulting partition arising from a community detection algorithm but accounts for the underlying network structure, specifically the strength of the connections within and between communities. This means, for example, that our method can capture differences between networks with similar partitions but where one of them might have a more defined or robust community structure. Third, it can also identify

[3]Theoretically, we can obtain $d(A, B) < 0$. This can occur when partition P_B better describes the modular structure of network A than its own partition P_A. Here, we suggest redefining $P_A = P_B$ as its new near-optimal partition. This then results in $d(A, B) = 0$.

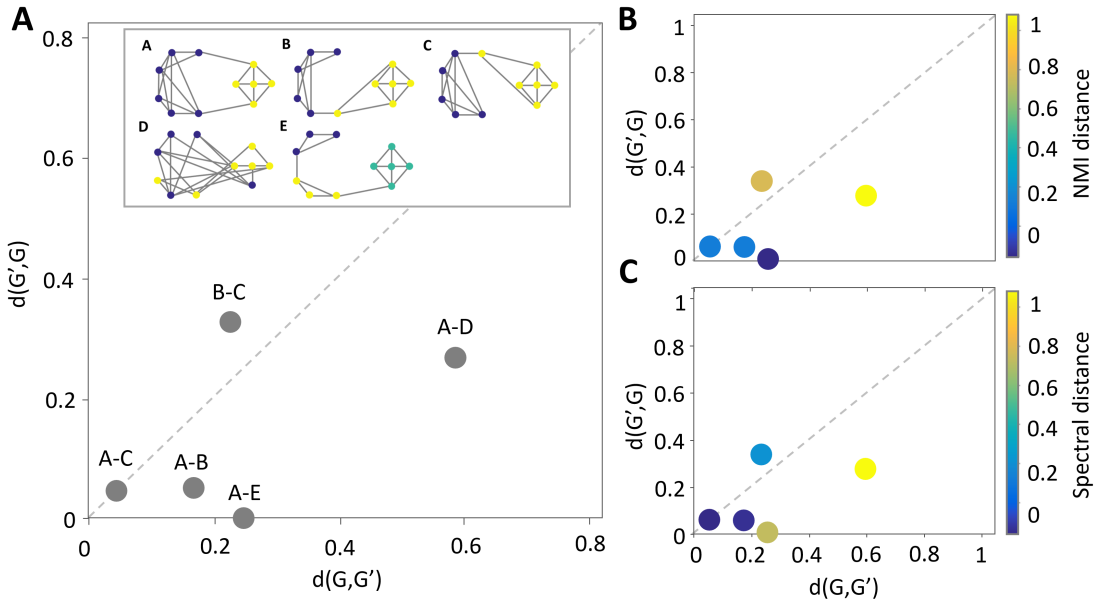


Figure 3.2: Using the same networks as Figure 1 (shown in the inset of (A)), we compare the same pairs using the BiDir distance, shown in (A). We notice that distance between the pair A-C is small, as demonstrated by a BiDir value very close to the origin and identity line. Similarly, if we project the communities of A onto network E, we obtain a very good fit and a near zero distance (as shown by the point on the x axis). This is a consequence of the fact that the communities of E are a refinement of the ones in A. This does not hold true in the other direction, *i.e.*, if we project the communities of E onto A, as shown by a non-zero distance of the point to the y axis. On the right, in (B) and (C), we can see how rankings derived from other metrics, such as NMI and spectral distance, do not fully capture these features. For NMI, the A-E pair is similar, but not so for the spectral distance.

cases in which very dissimilar partitions hide the fact that the community structure of both networks is relatively similar. A network might be well suited to multiple distinct partitions - for example, when there are nested communities - all of which will score highly in the objective function. Unlike other methods, our framework will pick up this signal.

In Figure 3.2 we illustrate the key features of the BiDir distance using the toy networks from Figure 3.1. We use modularity optimisation as the community detection algorithm. As outlined above, we obtain not one but two values for each pair of networks, which are plotted against each other in two-dimensional space in subfigure A.

The first thing to notice is that the two components of our distance convey distinct pieces of information. Points that are positioned away from the identity line (diagonal) correspond to asymmetric behaviour, whereby one of the bi-directional scores is higher than the other. We can interpret this as a case where the optimal partition of

the first network is well-described by the modular structure of the second. However, the optimal partition of the second network is not as compatible with the first. For example, we can see that the point corresponding to pair A-E is very close to the x-axis. This indicates that if we project the communities of A onto network E, we obtain a very good fit. In fact, since the distance is near zero, the partition found on A is virtually as good a description of E's community structure as the communities found on E. Looking at the other direction, whereby we project the communities of E onto network A, do we not find as good a fit. By eyeballing the relevant networks, shown in the inset, we notice that the communities of E are a refinement of the communities of A, and thus we can think of them as being nested in the communities of A. This explains why the distance $d(E, A)$ is so small.

The added information provided by the two components of the BiDir distance comes at a price. While single value distances are relatively easy to interpret, the two-dimensional score we obtain requires one to carefully consider the directionality of each dimension. Therefore, while BiDir is well suited to investigate hierarchical or nested similarities, it does not provide a unique way of ranking the networks in terms of similarity. Different choices on how to collapse the two dimensions into one, *i.e.*, geometric or arithmetic mean, will produce different results, and the application of interest should guide this choice.

The second key observation is that our approach implicitly considers the network's connectivity patterns, and not just the partition, by using the objective or quality function. Notice that, unlike NMI (or any other purely partition-based methodology), we differentiate between pairs A-B and A-C.

They are both equally distant from the x-axis, so the optimal partition of A, when projected onto B or C, results in more or less the same modularity score. The community structure of C, however, scores better than B when projected onto network A. Again, looking at the two networks, we can understand why this is the case.

In the former case, when we project the partition of A onto networks B and C, the partition misplaces a single node (relative to the optimal partition of B and C). In the latter case, when we project the partition of B and C onto network A, we see that both the partitions derived from B and C differ from A by a single node. However, in the first case, a node central to the blue community in A changes groups, while in the second, it is the most peripheral node. Hence, the distance between A and C is smaller.

This simple example shows how BiDir, by exploiting the objective function of the community detection algorithm, captures information on connectivity structure

underlying the community structure of both networks and is therefore capable of distinguishing cases in which NMI and related methodologies would not be able to discriminate.

At the same time, this dependence on the objective function comes with some trade-offs compared to metrics agnostic on the community detection algorithm. While NMI and other similar metrics can be used to compare the performance of different community detection algorithms by comparing them on benchmark datasets, BiDir uses a different function to calculate the distance for each community detection algorithm, and thus results for each algorithm are not directly comparable.

While the motivation for the BiDir distance was investigating the differences between skill-relatedness networks, the framework can be applied more generally. The key assumptions of our method are:

- The networks must be node-aligned (this implies that both networks have the same size).
- For each network, we have a partition representing its modular structure.
- This partition is near-optimal for a given quality function.

We emphasise that by modular structure we mean that it can be divided into subgraphs, where each subgraph is more tightly connected within itself than it is connected to the rest of the graph. Various community detection algorithms operationalise this concept differently, which is why they all define different quality functions. The quality function can add some restrictions to the applicability of our framework (*e.g.* modularity does not work on directed networks). But as long as the networks being compared are node aligned and feature modular structure, our framework can be applied.

3.3.2 Synthetic networks

Having introduced the BiDir distance, we will now construct several families of networks drawn from a stochastic block model (SBM) to further illustrate its features and test its performance. Specifically, we probe the effect of the strength of the community structure, and nestedness and overlap of communities, on the BiDir distance using modularity.

We generate three families of SBM, as illustrated in Figure 3.3. Full details on the parameters used to generate these ensembles are provided in Appendix A.1. The

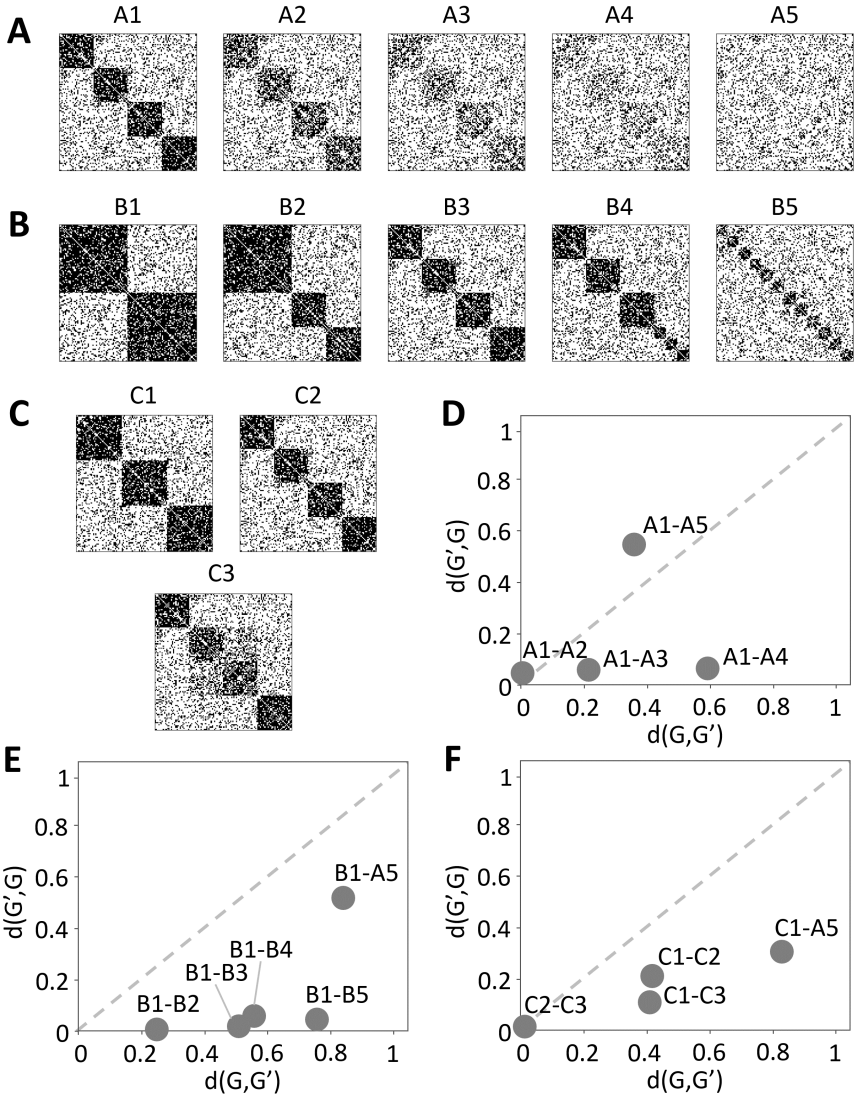


Figure 3.3: We illustrate our framework using three different families of SBMs. In family A all five networks contain the same communities, but the strength of in-group ties is sequentially decreasing. In (D) we see that the distances are zero whenever the original communities are detected, and increase with the weakening of the community structure. In family B each successive SBM is a refinement of the previous one, with blocks splitting along the way. So while B1 has two large communities, B3 has 4 communities, two for each of B1's communities. Notice that in (E), the community structure of B1 describes well the structure of B2-B5, but that doesn't hold in the other direction. In the last family of SBMs C1 has three communities while C2 and C3 have four identical communities. But for C2, two of these communities can also be merged into one larger community. In (F) we can see that, while C2 and C3 have an identical optimal community structure, our method picks up that C1 is closer to C3 than to C2 because C3 can also be partitioned into 3 communities.

distances shown correspond to the average BiDir distances over 1000 instances of each SBM.

The first family has four SBMs (A1-A4), all with the same community structure

but with a successively decreasing probability of in-group ties. As long as the communities can be recovered, their structure remains intact, and thus the distance between them will be 0. Once the communities can no longer be adequately identified by our community detection algorithm, the distances increase. We see in subfigure D that when we project the community structure of A1 onto A2, A3 and A4, we obtain values close to zero (*i.e.*, the points lie along the x-axis). As expected, A1's community structure is very close to optimal for A2, A3 and A4. However, if we instead project the community structure detected for A2, A3 and A4 onto A1, the distance progressively increases as the community detection algorithm struggles to find the underlying or generating structure. The final SBM A5, where intra- and inter-community connection probabilities are equal, does not have a community structure and is thus most distant from A1.

The second family comprises five SBMs (B1-B5), each a refinement of the previous one. In other words, B1 has two communities, B2 has three communities, obtained by splitting one of the communities of B1 into two, B3 has four communities and so on. We compare B1 to the other four SBMs. As we would expect, we observe that as we split the communities further, we increase the distance from the original network. More specifically, if we project the community structure of B1 onto B2-B5, we again obtain values close to zero (*i.e.*, the points lie along the x-axis). Hence the original partition B1 remains informative even after a few splits since the new communities are nested inside the original one. Looking in the other direction, however, the more disaggregate partitions describe the structure of B1 progressively less well.

Our last family of SBMs is slightly more complicated. C1 has three communities, but none of them maps onto the communities in C2 or C3. C2 and C3 are both split into four identical communities, but for C3, two of these communities are also strongly connected between them, and thus C3 can also be reasonably partitioned into three communities. In subfigure F, we can see that the $\text{BiDir}(C2, C3) = (0, 0)$ - since the near-optimal partition of C3 divides the network into four communities and is equivalent to the partition of C2. But when comparing these partitions to C1, we see a different picture. C3 is closer to C1 because the middle community of C1 is close to the combined middle block of C3. C2 is further from C1 as the middle community of C1 overlaps two very distinct blocks in C2. We see that the BiDir distance rightly identifies that although C2 and C3 have the same optimal partitions, their underlying community structure is markedly distinct.

Hence, even when the optimal partitions of two networks appear to exhibit little overlap, our method will generate a low distance score when there exists an alternative

and almost optimal partition that is similar. It is well known that for a flat modularity landscape, very different partitions might obtain similar near optimal scores [81]. Therefore looking exclusively at the resulting partitions is likely to miss important structural similarities that are nevertheless captured by the BiDir distance whenever these two distinct partitions produce similar scores for the optimization function.

3.3.3 Application to inter-industry labour flow networks

In this section, we compare the modular structure of four European Skill-Relatedness Networks (SRN) using the BiDir distance. Recall that, although it has previously been argued that the SRN is universal, we expect to find variation across the modular structure of different countries' SRN. Understanding this variation will unveil how structural characteristics of labour markets impact industrial diversification paths.

Here we compare the SRNs for Germany [148], Ireland [161], the Netherlands [68], and Sweden [146]. All the SRNs have been previously constructed according to the methodology of Neffke *et al.* [146] previously outlined in §2.5. Recall that in the SRN, a node represents an industry and an edge the skill-relatedness between its two corresponding industries. In all cases, the industry classification corresponds to the 4-digit NACE 1.1 industry classification. The graph intercept of these networks is used to ensure that all networks have the same size and are node-aligned - a requirement of the BiDir metric. Hence, each of the networks consists of 383 nodes.

In Figure 3.4 (A) we show a visualisation of the Irish SRN from O'Clery *et al* [161]. The node layout is based on a spring algorithm called 'Force Atlas' in Gephi, where more skill-related industries are positioned closer together. We have added labelling to indicate the general position of sectors in the network. We observe that service-orientated industries and government activities tend to be located on the left-hand side of the network. In contrast, heavy goods, construction, manufacturing and agricultural sectors dominate the right. Retail (bottom) and business (top) activities lie in between.

The modular structure of the network is shown via node colouring. The communities were detected using modularity [151]. Specifically, the partition maximising the modularity function from 10000 iterations of the Louvain algorithm is adopted. Using the same layout, we visualise the communities found for Germany, the Netherlands and Sweden on the right.

Figure 3.4 (E) highlights the overlap of the community structure of each country with the official industrial sector classification. Industries are ordered and grouped by their 1-digit NACE 1.1 sector along the x-axis. Each country's industries are

coloured according to community membership, corresponding to the node colouring in subfigures (A) to (D). Hence, blocks of colour within a sector indicate a close correspondence between a community and a sector. Conversely, a variety of colours within a sector indicates a community structure quite distinct from the official classification.

This figure provides insight into how the modular structure varies across the different SRNs. First, we observe that the Irish SRN has more communities than the other three networks. The Irish SRN has nine communities, the Netherlands and Germany have seven communities, while Sweden only has five communities. We can also see that industries within certain sectors are always clustered together across all four networks (as indicated by solid colour blocks), such as financial intermediation, public administration and social security, agriculture and the hotel and restaurant sectors. These communities show a universal structure in inter-industry skill-sharing. Furthermore, insights such as the increased subdivision of the manufacturing sector in Ireland compared to other SRNs can also be observed.

We now quantify the distance between these networks using our BiDir distance. The pairwise comparison of the various SRNs is shown in Figure 3.4 (F). We observe that Germany and the Netherlands are the closest in terms of modular structure with the smallest distance, and both are similar to Sweden. Ireland appears to have a slightly more different modular structure compared to the other countries. A comparison of the Irish SRN and the German SRN uncovers an asymmetric relationship. If we project the German communities onto the Irish SRN, we obtain a good fit - but not vice versa. Visual inspection of subfigures (A), (B) and (E) confirms that the communities of Ireland are somewhat nested inside the communities of Germany. For example, the manufacturing and social service sectors consist of many small communities in Ireland compared to just a few larger communities in Germany. Hence, inter-industry flows, and thus knowledge diffusion and skill-sharing that constrains development paths, are more fragmented in Ireland. From an Irish policy perspective, the German SRN might be informative in the design of smart interventions. For example, policies could be developed to facilitate and encourage worker mobility between sectors in, for example, manufacturing that is highly interconnected in Germany but not yet in Ireland.

While it is generally accepted [148] that the skill-relatedness networks vary little across space and time, our analysis shows that there are important differences across the European countries we analyse. While the modular structures of the Northern countries (Germany, the Netherlands and Sweden) are relatively similar, those of Ireland are less so. Furthermore, the directionality of our distance metric enables us to

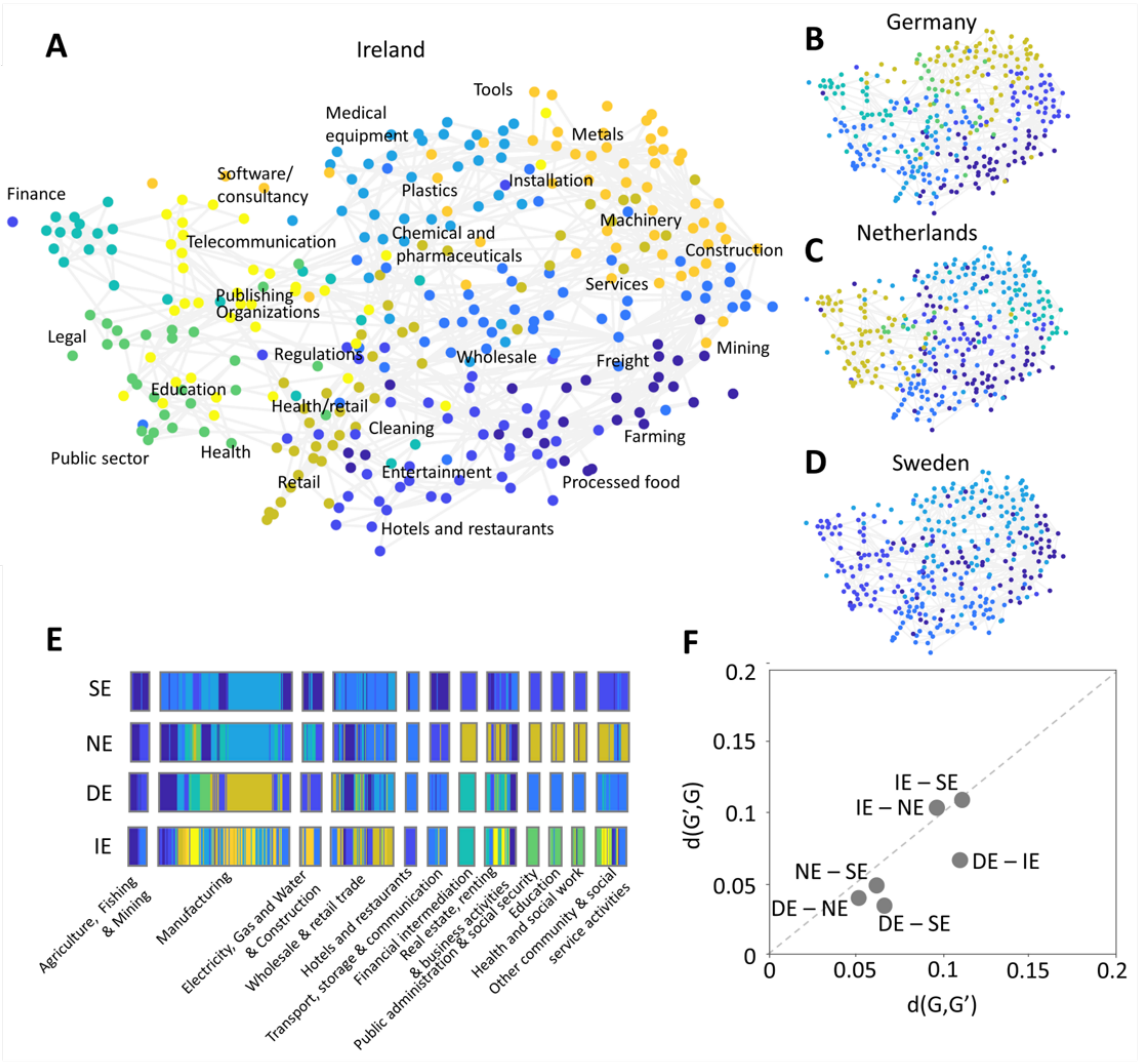


Figure 3.4: The pairwise comparison of the modular structure of four countries' SRNs, namely: Ireland (IE), Germany (DE), the Netherlands (NE) and Sweden (SE). In (A) we show the Irish SRN from O'Clery *et al.* [161]. The nodes represent industries and edges the skill-relatedness between the two corresponding industries. Nodes are coloured according to their community (found using the modularity optimisation algorithm). The node layout is based on a spring algorithm called 'Force Atlas' in Gephi. Labelling is added to indicate the general position of sectors within the network. Using the same layout, we also show communities detected for Germany, the Netherlands and Sweden in (B-D). In (E) we highlight the overlap of the community structure of each country with the official industrial sector classification. Industries are ordered and grouped by their 1-digit NACE 1.1 sector along the x-axis. For each country, industries are coloured according to their community membership. In (F), we show the resulting BiDir distance from the pairwise comparison of the different SRNs. Notice that the German, Swedish and Dutch SRNs have a relatively similar modular structure, while the Irish SRN's modular structure is less similar.

pick up nuanced differences between SRNs (specifically in terms of nested community structure, like in the case of Germany and Ireland), and uncover unseen potential linkages, which are key to policy efforts to generate regional industrial growth and

diversification potential.

3.3.4 Different optimization functions

The flexibility of BiDir, in terms of its adaptability to a class of community detection algorithms (with the use of an objective function), is advantageous in that it facilitates the comparison of networks across a wide variety of applications. This is because different optimisation functions define communities differently and will therefore often detect different partitions even on a single network. Thus far, we have illustrated the use of BiDir when using the modularity optimisation function [152] as our objective function. Here, we illustrate the BiDir distances obtained when comparing the modular structure of two networks using three different optimisation functions. Specifically, we consider InfoMap [183], Markov stability [61] and modularity functions. Specifically, we illustrate the degree to which properties of an optimisation function are inherited by and reflected in our distance measure.

We compare the modular structure of networks A and B illustrated in Figure 3.5 (A) and (B), respectively. Network A is a ring-of-rings and is an example of a network with non-clique-like communities (communities that have a high average diameter). The network is taken from Schaub *et al* [187]. Network B is a ring-of-circulant subgraphs $C_n\langle 1,2\rangle$ [4]. It has a similar community structure as network A, however, the clique-like structure of each community is enhanced (the average diameter of each community is decreased) by the addition of intra-community edges.

For modularity and InfoMap, the partition is obtained by applying the Louvain- and internal InfoMap heuristic 1000 times, respectively, and choosing the partition that either maximises the modularity or minimises the InfoMap function. Similarly, in the case of the multi-resolution stability algorithm (where time is the intrinsic resolution parameter), for each time a resulting partition is obtained by applying the Louvain heuristic 1000 times. For each time resolution, we identify the partition that maximises the stability function. In order to choose the partition corresponding to the optimal resolution, we calculate the mean variation of information[5] (VI) [139] at

[4] A circulant graph is a graph of order n in which the i^{th} vertex is adjacent to the $(i+j)^{\text{th}}$ and $(i-j)^{\text{th}}$ graph vertices for each j in a so-called connection set. A circulant graph of order n and with connection set $\{1,2\}$ is denoted as $C_n\langle 1,2\rangle$.

[5] The variation in information is a metric used to measure the robustness of our partitions. The variance of information calculates the amount of information (in an information-theoretical sense) two partitions share. Two partitions that are similar will exhibit a low variation of information. For each resolution, a pairwise variance of information is calculated for each partition obtained from the Louvain algorithm. The average of these is then the mean variation of information at the resolution. If this value is low, it shows that the detected partitions are similar and therefore more robust [139].

each resolution, and choose the partition corresponding to the resolution with the lowest VI. We consider the full range of partitions obtained at different resolutions in the next section.

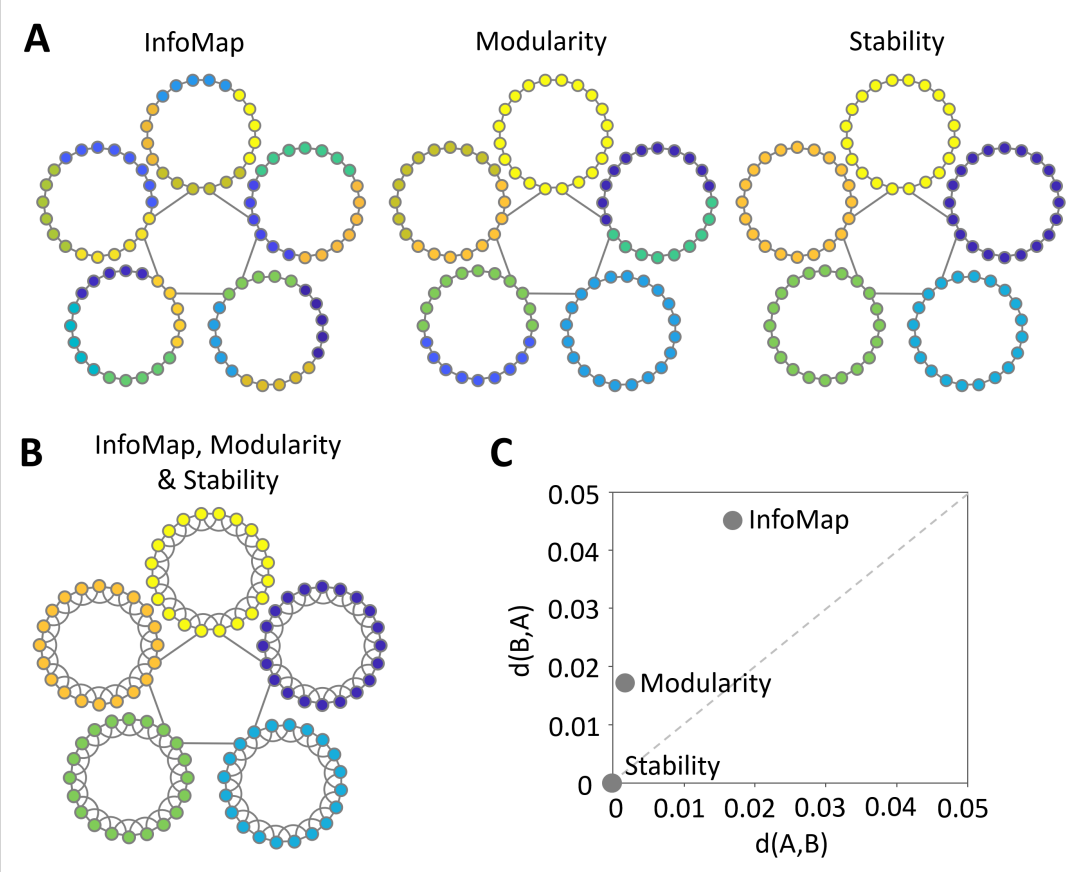


Figure 3.5: We illustrate how the BiDir distance varies when using three different optimisation functions, namely: Modularity, Infomap and stability. In (A) and (B), we illustrate networks A and B, respectively, with node colouring corresponding to the communities detected. In (C), we show the BiDir distances obtained when comparing the modular structure of networks A and B using each of the different optimisation functions. A larger distance is obtained when using the modularity and Infomap function compared to the stability function. This is because the first two functions suffer from a 'field-of-view' limit and, therefore, over-partition network A. In general, the BiDir distance inherits the features of the optimisation function deployed.

The resulting partitions are visualised via node colouring. As noted by Schaub *et al.* [187], we observe that both InfoMap and modularity over-partition network A (each ring is partitioned into multiple communities). This can be explained by the 'field-of-view' limit of which both of these functions suffer [89]. The 'field-of-view limit' is an upper limit on the effective diameter of the communities that can be detected with one-step dynamical community detection techniques [187]. As the rings in network A have high diameters (non-clique-like structure), both these algorithms

over partition the rings, creating communities with smaller diameters. However, this over-partitioning is not observed in the case of network B. This is because each of the circulants in network B has smaller diameters, and the optimisation functions are able to partition them into their own community. As the Markov stability function is a multi-step dynamical community detection algorithm, it does not suffer from the field-of-view limit [187], and obtains a community structure consistent with what we would expect for both networks A and B.

In Figure 3.5 (C), we illustrate the BiDir distances (comparing networks A and B) obtained when adopting each of these optimisation functions. As expected, Infomap and modularity result in the largest distances. Both of these distances are also highly asymmetric. We find that if we project the communities of B onto network A, we obtain a good fit - but not vice versa. The BiDir distance, therefore, recognises the subdivided communities found using modularity and Infomap. As InfoMap results in the most severe over-partitioning of the network, it results in the largest and most asymmetric distance. As the Markov stability algorithm does not suffer from the 'field-of-view limit', it obtains the same community structure for networks A and B, and results in a distance of $(0, 0)$. The different distances obtained using the different optimisation functions illustrate that the BiDir distance inherits the properties of the community detection algorithm used.

3.3.5 Multi-scale optimization functions

Although most well-known community detection algorithms seek to obtain a single node partition, in many cases, it is more natural to analyse a range of partitions - from many small communities to a few large communities - when investigating the modular structure of a network. Such a hierarchical structure can be informative, revealing layers of the organisation. For example, consider a university friendship network. Larger communities may reveal institutional structures such as departments or colleges, while smaller communities might capture friendship or social circles.

Hence, when comparing the modular structure of two networks, one may want to consider not a single partition but a range of partitions for each network. This approach can reveal, for example, distinct distances corresponding to particular scales, and consequently uncover the resolution at which the two networks are most similar. For example, two networks may be quite similar in their modular structure at a coarse scale but distinct at a finer or more dis-aggregate scale. Taking further advantage of the flexibility of our method, here we illustrate the use of the BiDir distance for a multi-resolution optimisation function.

To illustrate how we adapt the BiDir distance to this case, consider two networks, A and B. We compute the ratio of A's quality-score Q under B's optimal partition found at resolution β to its quality score Q under its own optimal partition found at resolution α, and vice versa. For the comparison, we calculate the quality of both partitions using the resolution of the partition of the base network. Formally,

$$d(A_\alpha, B_\beta) = 1 - \frac{Q_\alpha(A, P_{B_\beta})}{Q_\alpha(A, P_{A_\alpha})}, \tag{3.3}$$

where P_{A_α} is the optimal partition of network A obtained at resolution α and P_{B_β} is the optimal partition of network B obtained at resolution β. Furthermore, Q_α shows the resolution at which the quality function is calculated. Note that, as before, if we obtain $Q_\alpha(A, P_{A_\alpha}) < Q_\alpha(A, P_{B_\beta})$ we redefine $P_{A_\alpha} = P_{B_\beta}$ and obtain $d(A_\alpha, B_\beta) = 0$. Now, the final BiDir distance corresponds to:

$$D(A_\alpha, B_\beta) = (d(A_\alpha, B_\beta), d(B_\beta, A_\alpha)). \tag{3.4}$$

In practice, we calculate this distance for a range of α and β values such that small values of α or β correspond to fine partitions with many small communities. In contrast, larger values correspond to fewer larger communities[6].

To illustrate this approach, we compare three families of networks generated by hierarchical SBMs that share modular structures at three different scales. All three SBMs consist of 300 nodes (see Appendix A.1 for more details). In Figure 3.6 A–C, we visualise the adjacency matrix for a single network generated by each SBM. A and B share a coarse-level modular structure of three equally sized communities. However, at a more granular level, their community partitions are more dissimilar. On the other hand, networks A and C share a granular-level modular structure of 12 equally sized communities.

Next, we use the multi-resolution Markov stability community detection algorithm to detect a range of partitions for each network. Recall that this algorithm is based on a simple random walk model [61, 122, 124]. The key idea behind this method is that it sets a walker to roam on a network - jumping from node to node with probability proportional to the edge weight. If the walker gets trapped in a region of the network (a group of nodes) for a prolonged period, this corresponds to a group of densely connected nodes which form a community. Here, the Markov 'time' is used as the resolution parameter. Intuitively, if we let a walker roam for longer periods on the network, the walker will detect larger and larger communities. Therefore, by

[6]The range of α and β is network specific and chosen by the user.

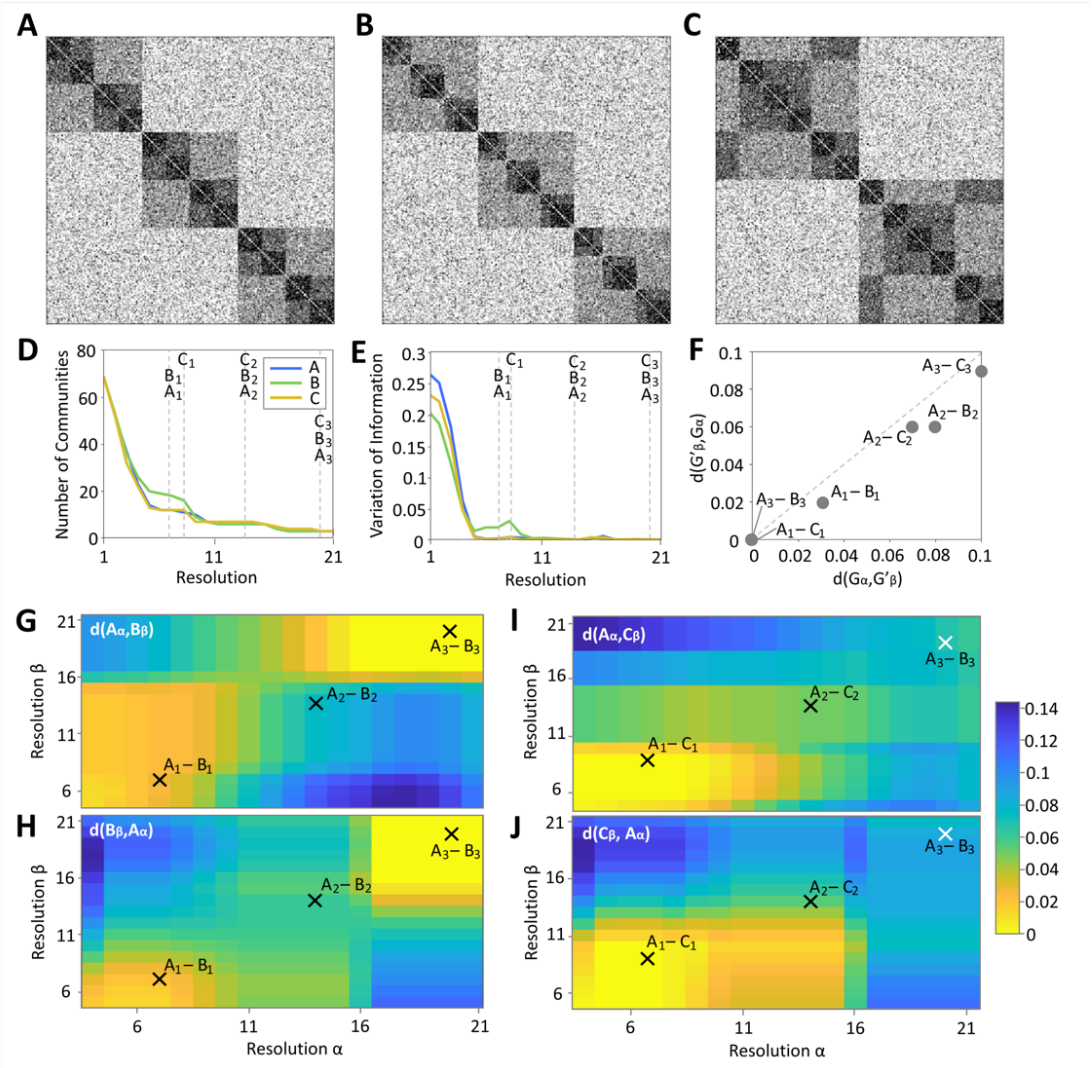


Figure 3.6: We illustrate the use of the BiDir distance to compare the modular structure of three nested SBMs. (A-C): We visualise a single network realisation for each SBM. Network A and B share a similar coarse-level community structure with three equally sized communities, while A and C share a similar granular-level community structure with 12 equally sized communities. In (D and E): We show the number of communities and the mean variation of information obtained using the Markov stability algorithm for each of the SBMs for a range of resolutions. For each network, we highlight three robust partitions corresponding to distinct scales. (F): We show the BiDir distances obtained when comparing these partitions. (G and H): We illustrate the BiDir distances obtained when comparing networks A and B and networks A and C across the full range of resolutions. In general, we observe that networks A and B's modular structure is most similar (e.g., the distance is smallest) at a coarse scale with few large communities, while networks A and C are most similar at a dis-aggregate or finer scale.

varying the Markov times we detect communities on a range of scales: from many small communities to a few large communities. More detail on this algorithm can be found in Lambiotte *et al.* [124].

We adopt the Louvain algorithm [27] as our search algorithm to find the optimal

node partition with respect to our objective function (the stability function [61]) at each resolution. We use the variance of information (across a large number of realisations of the Louvain algorithm) to assess which of these partitions is the most robust. In subfigures D and E, we display the number of communities and the mean variation of information for each resolution[7]. We highlight three resolutions that are both robust, with low variation of information, and correspond to three distinct scales of community structure.

First, we use the BiDir distance to compare these chosen partitions in the standard manner, as shown in (F). We observe that the distance varies depending on the resolutions chosen. For example, partitions A_3 and B_3, corresponding to larger scale coarse partitions of both A and B, are much more similar than A_1 and B_1 (fine) or A_2 and B_2 (moderate).

Next, we compare the modular structures of networks A and B across the full range of resolutions. Figure 3.6 (G and H) shows values obtained for different (α, β) resolution pairs. As our distance is two-dimensional, (G) corresponds to direction $d(A(\alpha), B(\beta))$ and (H) corresponds to direction $d(B(\beta), A(\alpha))$. Consistent with our example above, we observe that A and B's modular structure is most similar (e.g., the distance is smallest) at a coarse scale with few large communities. As we increase the community structure's granularity, the networks' modular structure becomes more distinct (at a moderate scale) and then more similar (at a fine scale).

We also observe an asymmetric relationship for the modular distance. For small α and moderate β we observe that $d(A(\alpha), B(\beta)) < d(B(\beta), A(\alpha))$. Hence, if we project the communities of network B onto network A, we obtain a good fit - but not vice versa. In this region, network B has fewer and larger communities (9 communities) than network A (12 communities). Thus, our metric captures that both partitions' stability is similar regarding network A.

Conversely, when we compare networks A and C (in I and J), we observe that the smallest distance is obtained at a fine resolution corresponding to the detection of smaller communities. As the resolution increases and larger communities are detected, the modular structure of the networks becomes less similar. Here we observe a stronger symmetric relationship between the two dimensions of the BiDir distance.

[7]Note that for all results, we calculate the average values obtained from generating $100,000$ network instances of each SBM and running the Markov stability algorithm with $1,000$ iterations of the Louvain heuristic on each network instance generated.

3.3.6 Application to inter-industry labour flow networks (cont.)

Using the BiDir distance with a multi-resolution optimisation function we now further investigate how the modular structure of the Irish and German SRNs differ. Recall that in §3.3.3, we observed an asymmetric distance, where through further inspection we found that the communities of Ireland where a somewhat nested version of the communities of Germany. We now investigate this relationship more thoroughly by examining their similarity at different resolutions.

In Figure 3.7 A and B, we visualise the Irish and German SRN adjacency matrices. We observe that both networks show clustering along the diagonal of the matrices. Furthermore, we see that Germany has an overall higher edge density than Ireland.

We use the multi-resolution Markov stability community detection algorithm to detect a range of partitions (from many small communities to only a few large communities) for each network. In subfigures C and D, we show the number of communities and the mean variance of information obtained for the partitions found at different scales for both graphs. For each network, we also indicate three robust partitions (with a low VI) that describe the network's modular structure well.

We now compare the modular structure of these two countries' SRN using the full range of partitions. The heat maps in subfigures E and F show the X and Y BiDir distances obtained for different (α, β) resolution pairs. The X distance corresponds to $d(IE(\alpha), IE(\beta))$ and the Y distance corresponds to $d(DE(\beta), DE(\alpha))$. We also indicate our chosen robust partitions on the heatmaps.

We observe that our distances are most similar at small resolutions (at (P1,Q1)), where the German and Irish SRN are clustered into many small communities. However, as we increase the resolution and allow for fewer but larger communities (at and around (P2,Q2)), we see that the X distance remains small (yellow) while the Y distance becomes large (blue). This shows that the German partition can well describe the Irish SRN's modular structure, but the Irish partition cannot describe the German SRN's modular structure. Hence, Germany has more well-defined mid-ranged communities (with a higher density of edges joining smaller communities into larger communities) compared to Ireland's mid-range communities which are more sparsely connected.

We also observe that as we increase the resolution parameter β keeping α constant (thereby increasing the size of German communities but keeping the Irish communities at a finer partition), we continue to see this same behaviour (small X distance with a corresponding larger Y distance). This again shows that Ireland's communities are

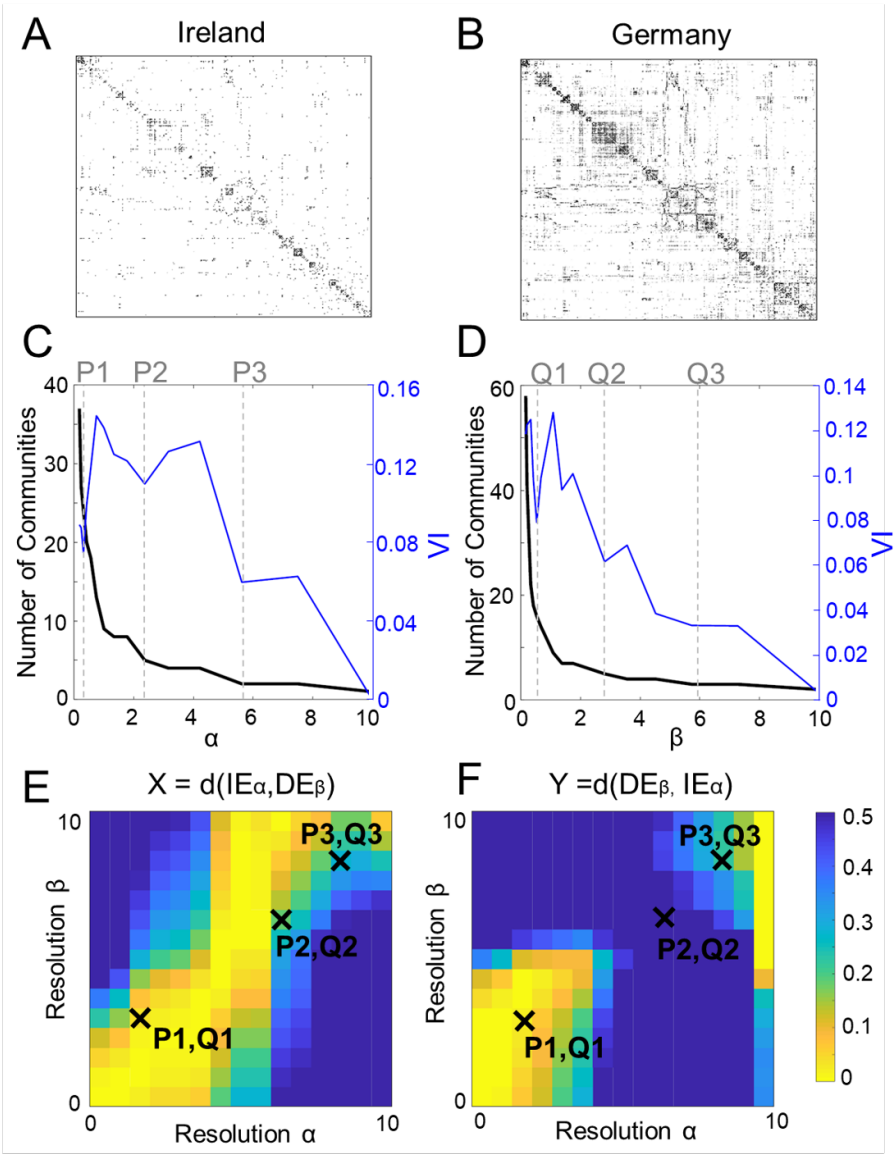


Figure 3.7: We illustrate the use of the BiDir distance to compare the modular structure of Irish and German SRN at different resolutions. (A and B): We visualise the adjacency matrices for Ireland and Germany's SRN, respectively. In (C and D): We show the number of communities and the mean variation of information obtained using the Markov stability algorithm for each network for a range of resolutions. For each network, we highlight three robust partitions. (E and F): We illustrate the BiDir distances obtained across the full range of resolutions. In general, we observe that Ireland and Germany's SRNs modular structure is most similar (e.g., the distance is smallest) at a fine scale with many small communities.

a nested version of German communities which are more strongly connected at more granular scales.

This result confirms that Ireland and Germany have similar small clusters, but Germany has more strongly joined various smaller clusters into larger clusters. This

shows that it has successfully widened smaller labour pools and allowed for less restraining of knowledge flows. This has important implications for industrial policy. Irish industrial policy-makers can use the German SRN to indicate which smaller sectors can be joined to create larger, more resilient labour pools.

Finally, at a larger resolution (at (P3,Q3)) when the community is described by only two or three very large communities, we see a weakly-similar community structure between the two SRNs.

3.4 Discussion

Comparing community structure across networks remains a challenging task. Most current methodologies equate the community structure to the graph partition, but this approach ignores key topological information. While we also compare two partitions, by assessing the fit of one partition to the topology of the other network (and vice versa), the bi-directional distance we propose incorporates information on the network structure. This approach enables us to identify similarities and differences in the underlying network topologies not captured in the partitions alone. The bi-directional nature of the measure is also advantageous, as it allows us to identify cases in which one network provides useful information about the other but not vice-versa. This is often the case when one network is a nested or sparser version of the others. Furthermore, as the measure is agnostic about the community detection quality function used, this approach can be deployed across a wide range of applications.

The BiDir distance assumes that the partition assigned to each network is optimal (under some chosen community detection quality function). Although this is also true for NMI, it may appear like a strong assumption. However, the BiDir distance is well suited to handle cases where there exist alternative partitions that are close to optimal. This is as the measure evaluates the underlying network structure, and will assign similar scores by the quality function. The measure is thereby able to identify cases in which dissimilar partitions hide the fact that the community structure of the two networks is relatively similar.

The multi-dimensionality of BiDir means that it is neither a metric nor a quasi-metric in a formal sense. While the directionality provides more information on the similarities and differences in modular structure, there are cases where it will be ambiguous. If we are interested in identifying which networks are the most similar (or different), an extra step is required to reduce the BiDir to a single score, which can be done by, for example, taking the sum of the squares. Even then, it is important

to keep in mind that there can be no clustering comparison methodology that can be generally used in all applications - the implications of the summary statistic will need to be carefully considered with the application at hand.

A key aspect of community detection is that the specific algorithm is chosen according to whether structural, dynamic or other features are of interest. For example, in the inter-industry flows, we are interested in flows within the network; we should, therefore, adopt a dynamic-based community detection algorithm. Similarly, our measure is well suited for comparing the community structure of networks when the application of interest is related to said function. In such a case, as we use the quality function in our method, we can retain this focus on the features of interest. However, the BiDir distance also inherits the properties of the quality function, which can constrain its use. Furthermore, unlike NMI, it cannot be used to benchmark different quality functions (or their corresponding community detection algorithms) using ground truth data. This is because we are unable to compare the scores of different quality functions.

Often partitions, found through a community detection algorithm, are assigned a level of robustness. This is often measured through variation of information (VI) and determines how often the algorithm finds the near-optimal partition. A partition with a particularly low VI, is assumed to well describe the modular structure of the network. It would be interesting for future work to investigate how we can incorporate information on community robustness into the Bi-directional distance framework.

Furthermore, when adopting a heuristic to optimise a quality function, a landscape of potential network partitions is considered. This landscape can have different properties, for example, it can be particularly flat or be characterised by various peaks. It would be interesting to investigate whether we could compare more of the partition landscape using our metric.

To construct a confidence interval for the Bi-Dir distance, an intuitive approach would be to adopt the parametric bootstrapping approach of [184] used to assign a level of significance to changes in the community structure of temporal networks. This approach uses parametric bootstrapping to create an ensemble of networks for each of the networks in the comparison. The ensemble consists of networks that have been slightly perturbed, thereby adding some noise. Each of the networks in each of the ensembles can then be compared to each other using the BiDir framework. One can then obtain a confidence interval by using the 5th and 95th percentile of the BiDir score distribution. This approach is one possibility for quantifying the significance of our distance.

Currently, our BiDir distance is solely designed for descriptive community detection algorithms that use the optimization of a quality function. Descriptive community detection algorithms attempt to find communities based on some context-dependent notion of what constitutes a good division of the network into groups. A further research endeavor could extend our BiDir distance framework to include other community detection techniques, like inferential community detection techniques. Inferential community detection methods describe a precise generative model which includes the notion of community structure, and attempt to fit it to the network data. Here, a distance between the posterior distributions for each network's own partition and the second graph's obtained partition can be obtained. This can then potentially be compared using a posterior odds ratio or a descriptive length as the distance function.

Furthermore, we have focused exclusively on comparing associative modular structure. However, another future research endeavour could include extending our partition swapping framework to compare other meso-structures (such as the core-periphery structure of networks). In this case, the appropriate quality function, measuring this structure, should be adopted. Similarily, it would be interesting to investigate how we can extend our measure to compare the degree of assortative structure across two networks. Assortativity, or homophily, in a network is the tendency of nodes of the same type or metadata to link to each other. This property is often found in social networks, as people of the same age, race or political belief prefer to link to each other compared to other with different attributes. In the literature, the degree of assortativity is measured by the assortativity coefficient, or modularity in the case of categorical metadata. Note that we can, and do, adopt this same function in our network comparison framework. Hence, it would be interesting to consider the implications of using our framework to compare the degree of assortativity across two networks, as we consider the underlying group ties in the comparison.

Finally, our measure identifies the presence of similar modular structure between network pairs. However, when it comes to applications, often more information is required to understand 'where' in these networks similarities (or differences) occur. This was clearly illustrated in the comparison of the inter-industry labour flow networks, where a finer investigation of community overlap was used to shed light on sectors with either conserved or distinct patterns of labour flows across countries. In the next chapter of this thesis, we investigate a further research endeavour to develop a more fine-grained community-level distance.

Chapter 4

Single community comparison

In this chapter, we develop a dynamic-based network comparison technique to compare a single community across two node-aligned networks. Unlike techniques in the literature, we show how the measure is able to capture both the influence of community edge density and edge connectivity in the comparison. We again, illustrate the workings and properties of the distance through toy networks, synthetic networks, and inter-industry labour flow networks.

4.1 Introduction

In the previous chapter, we developed a metric to compare the *global* modular structure of a pair of node-aligned networks. Here, we consider a more fine-grained approach: the comparison of a single community across a pair or set of node-aligned networks. As the difference in the modular structure is not always homogeneous across two networks, we may want to investigate which of the communities are more similar or different to each other? As a community can hold functional importance within a network, we can also be interested in whether a specific community has the same function properties across different networks? Furthermore, in the case of temporal networks, we may wish to investigate whether particular communities change more rapidly in their modular structure than others? All these questions render the development of a single community comparison measure.

These question arises in many real-world applications. For example, consider investigating how a disease might impact the function of a specific part of the brain. To investigate this, we could compare a community (representing a functional part of the brain) in a brain network across healthy and sick patients. On the other hand, we may wish to investigate whether a specific community in a partisan network is become more modular over time - thereby polarising the political landscape.

We are particularly interested in comparing the structure of specific communities in the skill-relatedness network (SRN) across different countries. Recall that a community in an SRN represents a skill basin: a group of industries in which workers can more freely move between them than to other industries in the network [161]. The more isolated a community, the less resilient it is to economic shocks. This is because, in the case of an economic shock, workers are unable to be easily absorbed into the rest of the labour market, and therefore have a higher chance of becoming unemployed [161].

Now, consider that we are interested in designing industrial policy for a specific community such as the financial sector. It is vital to understand the structure of this community, as well as how it compares across different countries. Often, policymakers turn to other countries to either adopt their data for evaluation or learn from their implemented policies. However, to what extent is this community similarly dis(connected) to the broader labour market across different countries? Or to which country should Irish policy-makers turn to learn from their implemented policies? To investigate this, we need to compare the modular structure of this sector (a single community) across different countries' SRNs.

Within the network comparison literature, there are only a few measures that compare the modular structure across node-aligned networks (e.g. NMI [57] and BiDir [193]). However, all these measures compare the global network modular structure. According to the author's knowledge, none take a single-community perspective. We are also unable to decompose one of these global modular network comparison techniques by considering a single community's contribution. This is because these measures evaluate the differences in the network partitions. As we are comparing the same single community across the networks - these measures are futile.

This chapter aims to construct a single-community modular comparison measure. To do so, we first need to quantify the modular structure of a single community in a way that allows it to be compared across different networks. We, therefore, turn to the community detection literature to find a suitable quality function that is comparable across different networks. We find that we cannot decompose a global quality function (such as modularity [151]) into its constituent communities and compare these values across different networks due to different underlying null models that arise when using the same quality function on different networks. Furthermore, current local quality functions, such as the most widely used conductance and local modularity [44], only considers the inner- and outer-community edge density (i.e. the number of edges inside the community or the number of edges that connect a node inside

the community with a node outside the community) but neglect the connectivity of these edges. By connectivity, we refer to the distribution of these inner- and outer-community edges. Considering connectivity enables the 'role' of nodes to also be considered. For example, differentiating cases in which nodes in the core (or the periphery) of the community are well connected outwards to the rest of the network.

Finally, we adopt the retention function (which forms part of the multi-resolution severability community detection algorithm [206]). This measure is based on a diffusion model on a network. The retention of a community measures the probability that the dynamics will stay within a community for a given time. If a community has high retention, it can well 'retain' the dynamics and is therefore isolated from the rest of the network.

Using this function, we propose a new dynamic-based single community comparison distance: *the maximum retention distance*. The distance measures the largest difference in the retention function of a community across two graphs at any point in time. The intuition is that if a community within two networks conserves the dynamics for a similar length of time, it has a similar degree of modular structure - considering both edge density and connectivity of the community in both networks.

The key advantage of this measure is that, unlike other measures in the literature, it captures the impact of a community's edge density and connectivity when comparing its modular structure across a set of graphs. The measure also does not use a null-model comparison or contain global network information, making the measure mathematically coherent for comparison across different node-aligned networks. Furthermore, the measure abides to normal distance behaviour (as defined by [195, 118]) which makes it easily interpreted.

In the next section, we more precisely define what we mean by edge density and connectivity. We then review the relevant literature - focusing on choosing a suitable quality function. This follows with a further examination of the properties of the retention function. We then introduce our maximum retention distance. In this section, we also discuss how the shape of the curve on which we take the maximum value can also provide deeper insight into how communities compare and how to obtain a confidence interval for our distance. We then illustrate its ability to capture both the impact of community edge density and connectivity as well as adhere to normal network behaviour. Next, the time at which the maximum retention distance is obtained is investigated. We then illustrate its implementation on real-world networks. Finally, we discuss its limitations and future research avenues to enhance its usability.

4.2 Terminology: edge density and connectivity

In this section, we define and illustrate two properties of a community: A community's edge density and its edge connectivity. We use both of these properties extensively throughout this chapter to investigate and compare the modular structure of a single community across different graphs.

First, we define a community's *edge density* as the number of inner- and outer-community edges. Inner-community edges refer to those edges that are incident to two community nodes, while outer-community edges refer to those that are incident to one community node and one node outside of the community. A community with a strong modular structure is characterised by many inner-community edges and few outer-community edges.

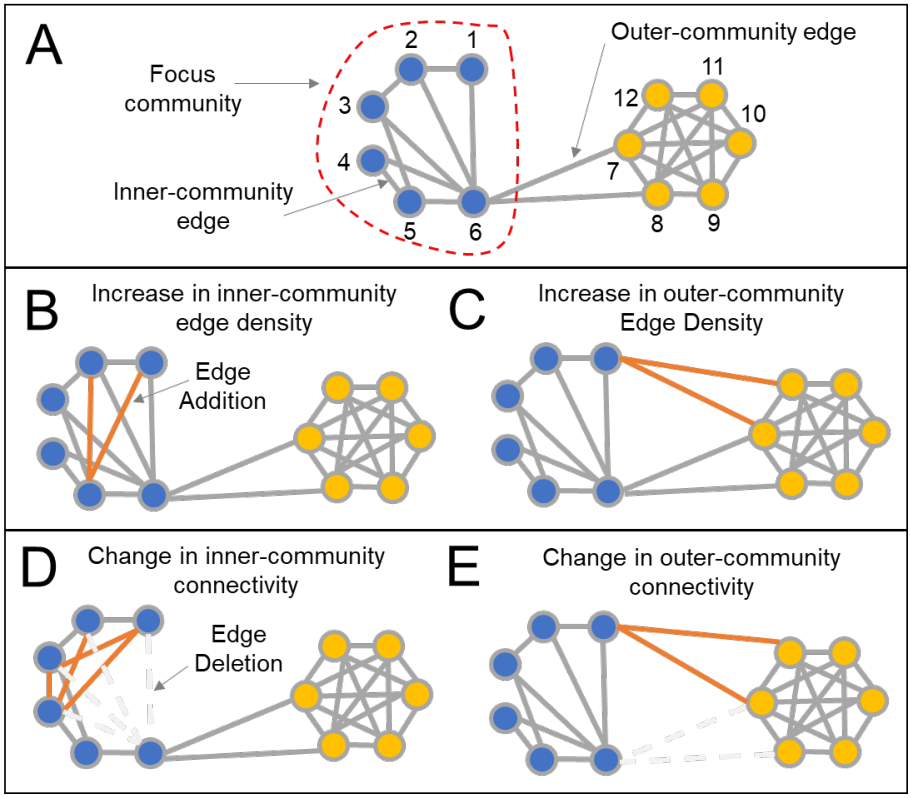


Figure 4.1: Toy networks illustrating an increase in a community's edge density and connectivity. (A) We show a toy network where we focus on evaluating the modular structure of the blue community. In (B&C) we illustrate the increase in inner- and outer-community edge density, respectively. In (D & E) we show a change in the inner- and outer-community edge connectivity, respectively.

We illustrate these concepts in Figure 4.1. First, in sub-figure A, we show a toy network with two communities illustrated through node colouring. We focus on evaluating the modular structure of the blue community, which is encircled in red.

We also label the inner- and outer-community edges on the figure. In sub-figure B, we illustrate an increase in the inner-community edge density by adding two inner-community edges. Note that the blue community has a stronger modular structure in B than in A. Similarly, we illustrate an increase in the outer-community edge density in C. Here, the blue community in A has a stronger modular structure (and is more isolated) than in C.

Next, we define a community's *connectivity*. This refers to how the inner- and outer-community edges are distributed amongst the community nodes. More specifically, inner-community connectivity refers to how the inner-community edges are distributed amongst the community nodes. Edges can be evenly distributed across all nodes (displaying a uniform edge distribution) or highly concentrated where most edges are connected to only a few nodes (displaying a power law edge distribution). Similarly, outer-community connectivity refers to how the outer-community edges are distributed amongst the community nodes. For example, the edges can be connected to only a few nodes in the community or to many nodes in the community.

Considering the inner- and outer-community connectivity together enables us to consider the influence of node roles on the community's modular structure. For example, this enables us to differentiate between cases where nodes in the community's core (or the periphery) displaying high (or low) inner-community connectivity are connected outwards to the rest of the network. An isolated community, with a strong modular structure, is characterised by edge connectivity, where peripheral nodes are strongly connected to the rest of the network and core nodes are weakly connected to the rest of the network.

We illustrate the change in a community's connectivity in sub-figure 4.1 (D) and (E). In D, we show a change in the inner-community connectivity. Here we reshuffle (remove and add) four edges. Note that the change in connectivity results in node 6 changing from a core to a periphery node in the community. As node 6 is connected outward to the rest of the network, we observe that the community becomes more isolated (and a stronger community) in D compared to A. In E, we illustrate a change in the outer-community connectivity. Here we see that the outer-community edges shift from node 6 to node 1. As node 1 is more peripheral in the community, this change results in the blue community having a stronger modular structure.

4.3 Literature

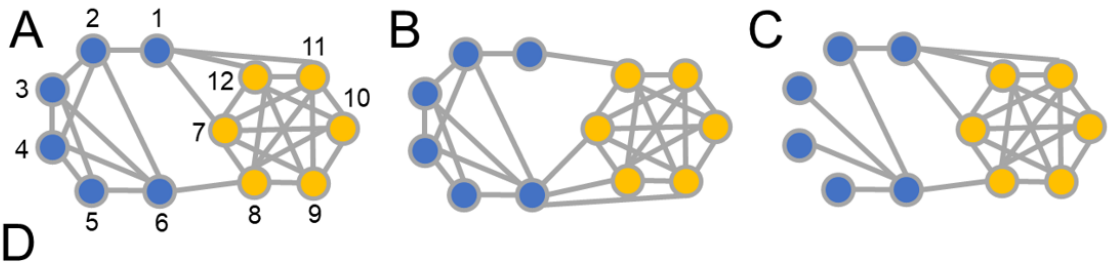
4.3.1 Adopting a network comparison technique

We reviewed an array of node-aligned network comparison techniques in the previous chapter. Amongst the ensemble of techniques, there were only a few that capture differences in the modular structure of node-aligned networks (*e.g.* NMI and BiDir). However, these techniques pick up differences in the two networks' community partitions (NMI in the degree of overlap in the partition sets and BiDir in the difference in the partitions' quality within each of the networks). If the partitions are equivalent, both distances are zero. As we are interested in comparing the same single community across two networks - it is unclear how we could adopt these measures to pick up any modular structure differences.

According to the author's knowledge, there are no modular network comparison techniques that focus on comparing a single community across different networks. However, various authors [59, 144] have adopted more ad-hoc approaches by applying local network comparison techniques to address this problem. For example, a common ad-hoc approach is to only consider the community (its nodes and inner-community edges) as an independent sub-graph. This community sub-graph can then be compared across the network pair using a local network comparison metric (*e.g.* the graph edit distance - where the number of edges that need to be removed or added to transform the community in the first graph to its structure in the second graph is counted). This approach captures the differences in the inner-community structure across the network pair. However, as the outer-community edges are ignored, the metric doesn't capture the impact of how many and which nodes are connected to the rest of the network.

A second, ad-hoc approach is to collapse the community (all its nodes and inner-community edges) into a single community node. The centrality of this community node (*e.g.* using the degree centrality) can then be investigated and compared across the network pair. This approach captures how well connected the community is to the rest of the network. However, it ignores the inner-community structure.

We could evaluate both of these ad-hoc measures together, and compare both values. This would give us an indication of how the inner- and outer-community structure differs. However, evaluating these structures separately ignores how they influence each other. For example, it does not consider which nodes within the community (*e.g.* those in the core or periphery of the community) are connected outward to the rest of the network.



	A-B	A-C
Edit distance of isolated community	0	-5
Degree distance of collapsed community	0	0
Δ Conductance	0	-0.1587
Δ Local modularity	-0.0214	0

Figure 4.2: A comparison of the modular structure of the blue community across three toy networks using various distance measures in the literature. (A-C) Visualisation of three toy networks where the networks' communities are indicated through node colouring. (D) A table showing the difference in the modular structure of the blue community in network A to its structure in B and C using different distance measures in the literature. We observe that no measure can capture that A is both more modular than B and C.

In Figure 4.2 we illustrate how the above two ad-hoc approaches fail to capture the difference in modular structure between two toy networks. In sub-figure A-C, we show three toy networks. We first focus on comparing the modular structure of the blue community in A to its' structure in B. When closely examining the blue community in these two networks, we see that it differs in which node in the blue community is strongly connected to the yellow community. In graph A, node 1, which is on the periphery of the community, is connected to the yellow community. In B, node 6, which is at the core of the community, is connected to the yellow community. Therefore, the blue community in A has a stronger modular structure and is more isolated than in B. As the above measures consider the inner- and outer-community structure separately, they do not capture their interplay. They are unable to differentiate between the node roles. We observe in Table D that both measures obtain a distance of zero and are unable to capture any difference in the community's modular structure of A and B.

4.3.2 Adopting a community quality function

Next, we turn to the broader community detection literature to investigate if we can adopt a quality function to evaluate the modular structure of a single community and then compare it across different networks.

4.3.2.1 Decomposing a global quality function

Recall that a community detection algorithm aims to find a node partition that best describes the modular structure of the network. Inherent to each of these algorithms is a quality function that measures how well a given partition describes the modular structure of the underlying network. This function also defines what is meant by a community. The algorithm, using a heuristic, then tries to find a node partition that maximizes this quality function.

There is a range of quality functions that differ in how they describe a community. In §3.2.1, we reviewed a range of community detection approaches which differ in their perspective, and thereby their quality function, used to define a community (cut-based-, clustering-, stochastic equivalent - or dynamic-perspective [186]). When choosing a quality function (within a corresponding community detection algorithm), the context is critical to ensure that the community defined is meaningful for the complex system which the network represents. As we are interested in considering labour flows (a particular dynamical process) on an industry network - a dynamic perspective is most well-suited to our motivating problem.

Now, most global quality functions are a summation of each individual community's measured quality. This is true in the NG-modularity quality function [151], where we sum the number of edges within each community compared to the number of edges we expect at random. Similarly, in the Markov Stability's (dynamical-perspective) quality function [61], we sum the probability of a random walker to remain in a community compared to at random. In each of these cases, more modular communities result in higher quality scores.

Naively, we could consider using each community's contribution to the quality function to evaluate and compare the modular structure of the community across different networks. However, we can't compare the values obtained by these quality functions across different networks. This is because the measure is a difference from a null model. The null model differs across different networks as it contains topological information about the underlying network (in the case of Newman-Girvan modularity, the underlying network's degree distribution is considered in the configuration null

model [142]). We can only meaningfully use each community's contribution to its quality function to compare its modular structure *within* the same network. This, therefore, renders the use of a community's contribution to a global quality function non-viable for comparison across different networks.

4.3.2.2 Adopting a local quality function

Next, we turn to the local community detection literature to consider whether we can similarly adopt a local quality function to evaluate and compare the modular structure of a community across different networks. A local community detection algorithm sets out to find a single community of nodes concentrated around a given seed node in a localised way. These algorithms are especially beneficial in large-scale real-world networks where it is computationally intractable to use global information or where only partial information may be available [128, 206]. There is a vast array of these algorithms which differ in their expansion strategy and optimisation heuristics (see [204, 202]). However, most adopt conductance as their quality function to be optimised, as it is currently deemed the best scoring function [205].

Conductance captures how strongly a set of nodes is connected to the rest of the network; sets of nodes that are isolated from the rest of the graph have low conductance and make good communities. Formally, the conductance of a community C measures the fraction of the total community edge volume that points outside of the community and is given as:

$$\text{Conductance}(C) = \frac{\sum_{i \in C} \sum_{j \notin C} A(i,j)}{\min(\sum_{i \in C} \sum_{j} A(i,j), \sum_{i \notin C} \sum_{j \notin C} A(i,j))}. \tag{4.1}$$

Note that conductance takes a local cut-based community detection perspective.

Conductance has been criticised in that it favours communities with quasi-cliques [112] or communities of large size which may have irrelevant sub-graphs [7]. Note that our use of a quality function is slightly different. Unlike this literature, we are not trying to detect a local community around a seed node but instead aiming to evaluate and compare the modular structure of an already defined community. When adopting the conductance for this case (by taking the difference across a network pair), it still, however, falls short in that it merely captures the difference in community edge density (i.e. the number of edges) but ignores their connectivity (i.e. how they are connected amongst the community nodes).

We illustrate this in Figure 4.2. Again, consider the comparison of the blue community's modular structure across networks A and B. As conductance evaluates the

number of outer-community edges (4 edges in both graphs) and the total community volume (13 edges in both graphs) independently, it does not take into account which of the community's nodes (either node 1 in A or 6 in B) are connected outward to the rest of the network. Hence, the conductance of the blue community in A and B are equivalent (calculated as 4/13 for both graphs), and the resulting distance is zero.

Various other local quality functions have been considered within this literature (see [205]). Many of these either capture inner- or outer-community density separately. For example, the well-known *density* quality function considers the sum of edges inside the community relative to the communities size. As these functions consider inner- and outer-community density separately, they also fail to consider their interplay. Various authors have also proposed slight variations of the conductance measure. For example, Luo *et al.* [130] proposed a measure that directly compares the ratio of inner- and outer-community edges. This is merely the inverse of the conductance (provided that the volume of the community is smaller than the volume of the rest of the network). Hence, these measures similarly only focus on community density and ignore its' connectivity.

Chen *et al.* [44] proposed a variant of conductance, the *local modularity* function, that captures some of the community's connectivity. Here, they measure the quality of a community by the "sharpness" of its' boundary. A community's boundary B is the set of nodes in the community that are directly connected to any other node outside of the community. More specifically, the local modularity function measures the quality of community C by the fraction of the boundary nodes' edges that point inwards to other nodes in the community. Formally, the measure is given as:

$$\text{Local modularity}(C) = \frac{\sum_{i \in B} \sum_{j \in C} A(i,j)}{\sum_{i \in B} \sum_{j} A(i,j)}. \tag{4.2}$$

Although this measure does consider the connectivity of the nodes on the boundary of the community, it still ignores the structure of the rest of the nodes within the community that are not in the boundary.

We illustrate how this fails to capture differences in modular structure in Figure 4.2. We now focus on comparing the blue community in network A to its' modular structure in C. Observe that the blue community's boundary (nodes 1 and 6) has the same structure in both networks. However, the other nodes (nodes 2,3,4 and 5) in A are more well-connected to each other than in C. The blue community in A is, therefore, more modular than in C. We observe in Table D, that the difference in the local modularity function results in a distance of zero. As the measure only

considers the structure of boundary nodes it is unable to capture the difference in modular structure of the community in these two networks.

More recently, Yu *et al.* [206] proposed a dynamics-based local community detection algorithm - the severability algorithm. In this paper, we adopt the retention function, which forms part of this algorithm, to compare the modular structure of a single community across different networks. However, before introducing this function, as well as the severability algorithm, we first briefly review some key results on the use of dynamics to explain network structure. This is to provide the reader with a deeper understanding of how dynamics can be used to investigate network structure and why this perspective is well-suited to our problem.

4.3.3 The use of dynamics to explain network structure

There is a vast literature investigating how network structure influences the dynamics that act on it. For example, the spectral properties of the matrix encoding the graph structure (e.g. the adjacency matrix or the graph Laplacian) can describe various aspects of linear dynamics acting on the network. However, one can also take the reverse perspective, and use the dynamics acting on a graph to characterise the structural features of the underlying graph [137]. Here, we only consider a fixed underlying network, where the dynamics is simulated on this fixed network.

Taking a dynamics-based approach to probe graph structure is advantageous compared to other static methods, as it naturally reveals the structural properties of a graph at different scales (local, meso and global). This approach is also particularly advantageous when considering community structure (meso-scale structure), as unlike a static approach, this perspective allows for the coarse-grain description of the behaviour of a system - not only its structure [124]. This perspective has also gained traction within the community detection literature, both in defining and detecting communities within networks [123]. We briefly outline some key results on how linear dynamics can unveil network structure at different scales but primarily focus on its role in uncovering modular structure.

First, we focus on how a dynamic process can uncover local structural properties of a network. Consider a diffusive dynamical process on a graph modelled by a discrete unbiased random walker. As introduced in §2.2, this process is governed by Eq (2.1). If the underlying graph is assumed to be un-directed, connected and non-bipartite (aperiodic and ergodic Markov process), the asymptotic dynamics are characterised by an unique stationary state given by $\pi = \mathbf{k}/2m$. Hence, the stationary state of this particular dynamical process reveals the degree distribution of the graph - a local

property of the graph structure. One can therefore use the stationary state of the dynamics to uncover each node's degree. Using the same approach, but with different dynamical processes, we can also detect other network centrality measures such as the PageRank [36] or Eigencentrality of each node in the graph.

Secondly, we can also harness the dynamics acting on the graph to unveil its overall connectivity - a global network property. More specifically, we can look at the random walk relaxation time [80]. The random walker relaxation time is a measure of the speed of convergence of the dynamical process to its stationary state. It measures the effective size of the system in terms of dynamics. This quantity also reveals the overall connectivity of the underlying graph. It is well-known that this measure is associated with the spectral gap, which is the difference between the first two dominant eigenvalue of the graph Laplacian $(1 - \lambda_2)$ [80].

So far, we have shown how the asymptotic dynamics unveil both local (through the stationary state) and global (through the relaxation time) structural features of the graph. Before, turning to the meso-scale features, we first also highlight how the dynamics can show the presence of structural symmetries within a network. For example, the dynamics can potentially show the presence of an (External) Equitable Partition ((E)EP) [88, 188] on a graph. An EP divides a network into cells so that the nodes inside each cell have the same out-degree pattern with respect to every cell. An EEP similarly divides a network into cells, but the requirement is relaxed so that nodes only have to hold the same number of connections between different cells. Each of the various cells within the graph have equivalent diffusive dynamics. The dynamics can therefore be reduced and exactly described via a quotient graph (where we collapse groups of nodes into a single node). Hence, by identifying the presence of a reduced dynamic description (through equivalent node dynamics) we can uncover structural graph symmetries.

Finally, and most important to our chapter, we can also use dynamics to study and detect the modular structure of a network. This is because the modular structure of a graph can be revealed when time scale separation is observed in a diffusive process on the graph. Time scale separation is a phenomenon first explored in detail in the framework of system dynamics by Ando, Fisher and Simon [9, 190], who considered the existence of a partition of states into components with sufficiently low dynamical influence between them. For short time horizons, there is limited cross-influence between components, and thereby we can accurately describe the dynamics using the dynamics of the disconnected components. For long times, the states inside the components reach equilibrium and thereby, the dynamics of the system can be

accurately approximated by the aggregated system (*i.e.*, where all components are collapsed into a single state). This phenomenon, therefore, allows for a simplified description of complex system behaviour.

Now, to understand how we can harness a random walker model to detect communities in a modular network, consider a random walker positioned on a node in a community within a modular network. For a particular period of time, the random walker will reach other nodes within the same community with a greater likelihood than those in the rest of the network. In a sense, it will be trapped within the community. Hence, the dynamics observed are essentially decoupled for a time interval. Therefore, the modular structure of a network characterises the time-scale separation of the dynamics acting on it [185]. This essentially underpins the workings of several popular dynamical-based community detection algorithms, such as the WalkTrap algorithm [177], the Map Equation [183] and the Markov stability method [61]. The Markov stability also provides a framework to understand various other well-known community detection algorithms such as the Newman-Girvan modularity [151] and the Potts model [181] from a dynamical lens. This approach is advantageous as it defines the behavioural communities that allow for a reduced description of the network function. Furthermore, it also naturally extends to a multi-resolution community detection where the Markov time of the random walker is a natural resolution parameter [187]. It unveiled the network community structure at a range of scales (from many small communities to only a few large communities).

The above-mentioned dynamic-based community detection methods all take a global perspective. They focus on finding the global community structure. In certain contexts, only partial information about the network may be available, or the network may be sufficiently large that these approaches are computationally intractable. This originally motivated the local *severability community detection algorithm* [206]. This algorithm aims to find coherent dynamical structures at different time scales (*i.e.* modular structures that display time scale separation). The algorithm uses the severability quality function, which is made up of the retention and mixing time functions from Markov chain quasi-stationarity [58].

Intuitively, drawing an analogy from energy landscapes, the framework defines a community as a set of nodes with high barriers (long escape times of the random walker to the rest of the network measured through the retention function) and low roughness (short mixing time of the random walker within the component). More formally, retention is defined as the conditional probability that a walker, starting uniformly across all nodes in the community, will not escape after t time steps. Mixing

is defined as the total variation distance between the probability distribution of the random walker at time t and its distribution reached at long times, given that the walker is restricted to only the community states (is never leaves the community). Note, that if the walker position at t and $t = \infty$ are strongly correlated, the walker poorly mixes. The severability of a community at time t is then defined as the average of the retention and mixing at time t. The function can therefore be understood as a balance of mixing and retention. As time increases, walkers can better mix but also have more opportunities to escape. The function will ultimately peak at a given time, below which walkers are poorly mixed and beyond which retention is degraded. The severability measure's theoretical underpinnings lie in the local adaptation of the Simon-Ando-Fisher time scale separation theorem, as it measures the degree of dynamical influence between sets by only considering local interactions. For a more detailed description of the severability function and its use to unveil community structure, the author is referred to [206].

Note that the algorithm aims to find a local partition (around a seed node) that maximise the severability function. Although this measure takes a dynamic approach, it is less clear how we can adopt this quality function to evaluate and compare the modular structure of a pre-defined community across different networks. Naively, we could consider calculating the maximum severability of a pre-defined community (the value at which the severability curve peaks). However, this value is not comparable across different networks. This is as the curves (mostly) peak at different times, making the values obtained asynchronous and non-comparable.

In this chapter, we adopt the retention function to evaluate and compare the modular structure of a single community. In the next section, we further elaborate on its properties and why it is a suitable function to compare the modular structure across different graphs.

4.4 Methodology

In this section, we more formally introduce the retention function and discuss its properties that motivate our adoption of this function in our distance measure.

4.4.1 The retention function

The retention of a community is the expected probability that a random walker, starting uniformly distributed across all nodes within the community, will not escape the community after a fixed time. The intuition is that if a subset of nodes is particularly

isolated the walker will tend to stay within the community for longer periods, as it is hard for it to escape, and result in a high retention value. Hence, the function defines modular structure as a set of nodes where dynamical flow is isolated from the rest of the network.

More formally, we define a graph $G(V, E)$ with a set of n nodes V and a set of non-negative and non-directed edges E. We assume that G is connected and aperiodic. Recall, that a discrete unbiased random walk on a graph G is governed by Eq (2.1), again given as:

$$\mathbf{x}(t+1) = \mathbf{x}(t)T = \mathbf{x}(0)T^t,$$

where $\mathbf{x}(t)$ is a $1 \times n$ probability vector representing the probability that a walker is present on a node at time step t and T the transition matrix ($T = D^{-1}A$).

Now, given a connected community $C \in V$ with n_c nodes, which has at least one outer-community edge (the community is not completely isolated), let Q be a $n_c \times n_c$ sub-matrix of T corresponding to the nodes in C.

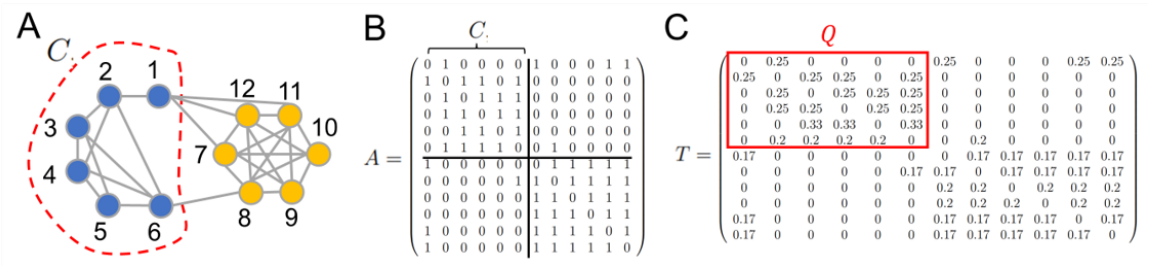


Figure 4.3: An illustration of the Q matrix of a community in a toy network.(A) A toy network where we choose the blue community as our focus community. (B & C) We show the network's adjacency matrix and transition matrix, respectively. We highlight the blue community's Q matrix on the transition matrix.

In Figure 4.3, we show a toy network A with its corresponding adjacency matrix (A) and transition matrix (T) in sub-figure B and C, respectively. We choose the blue community in A as our focus community C, and highlight its corresponding Q matrix on T in the figure.

Now, the retention of community C of graph G at time t, is defined as:

$$\rho(C, G, t) = \frac{1}{n_c}\mathbf{1}Q^t\mathbf{1}'. \tag{4.3}$$

The retention $\rho \in [0, 1]$, where perfect retention ($\rho(C, G, t) = 1$) is where no walker has left C at time t, and no retention ($\rho(C, G, t) = 0$) is where all walkers have left C at time t.

4.4.2 Properties of the retention function

In this section, we elaborate on the properties of the retention function that motivates it to be adopted into our single-community network comparison framework.

The first key property is that the measure focuses on how isolated the dynamics acting on a set of nodes are. This well matches the community structure we wish to compare in our motivating example.

Second, the function considers the behaviour of an unbiased random walker within the community. Hereby, it inherently considers both the edge density and the connectivity of the community. To further elaborate, note that the retention function is merely the row-sum of the community's Q^t matrix. As this matrix is a sub-matrix of the transition matrix ($T = D^{-1}A$) it contains the full information regarding the structure of the community.

We can also see how it picks up both edge density and connectivity by considering what it is measuring at different times. First, we consider the retention at $t = 1$. This measures the difference in the conditional probability that the walker will remain in the community after one step. This is simply the average fraction of the boundary node's edges that point to other nodes within the community ($k_{in}(i)/k(i), \forall i \in B$); equivalent to the local modularity measure [44]. Recall, that the disadvantage of this measure was that it only considers the structure of boundary nodes and ignores the structure of all other nodes in the community. Now, as t increases the retention measure considers probability paths of increasing longer lengths and allows the walker to explore more (and deeper parts[1]) of the community. For each increasing time, the measure now evaluates both the boundary node's $k_{in}(i)/k(i)$ ratio but also this ratio for the nodes that are connected to the boundary nodes, and the ratio for nodes connected to these nodes *etc.*. Hence, to obtain high retention at larger times we want communities with many nodes with high $k_{in}(i)/k(i)$ ratios, however, we also want these nodes to be strongly connected to other high ratio nodes. Note, that we are considering both community edge density and connectivity.

The third key property of the retention function is that because it only considers the conditional probability of the walker remaining in the community, it is not influenced by the topology of the rest of the network (i.e. the structure of nodes outside the community). This is an advantage as it allows the comparison of the community across different graphs and ensures that only the differences in the community's topology is captured. To elaborate, consider the opposite case, in which we also consider

[1] As the time increases, walkers from deeper in the community can escape and thereby the influence of the structure of nodes in deeper parts of the community is considered.

the structure of nodes outside the community. Here, we would consider the probability of a random walker over the whole network (with a similar initial probability distribution only inside the community). Although evaluating the full probability distribution will include interesting information on the dynamic influence of the community on the network, the probability distribution will include information on the structural topology of the rest of the network. It will therefore be difficult, especially at larger Markov times, to entangle to what extent the measure is picking up the structure of the community or rather that of the greater network. This merge of information makes it difficult to interpret this value, and compare it across different networks.

We can formally see, that the retention function, only captures the community's topology, as retention only uses the sub-matrix Q of P. Essentially, we can transform the graph into an absorbing Markov chain ($G \rightarrow \tilde{G}$) and still obtain the same community retention value. We transform the graph by collapsing all nodes in the rest of the network into a single absorbing node (which we call "RON" standing for the rest of the network) and making all edges from the community to node RON directed (pointing towards RON). Equivalently, we are transforming T into $\tilde{T}$ given as:

$$\tilde{T} = \left(\begin{array}{c|c} Q & \mathbf{r}' \\ \hline \mathbf{0} & 1 \end{array} \right), \tag{4.4}$$

where $\mathbf{r}$ is a $1 \times n_c$ vector indicating the fraction of each community node's edges that are incident to RON and $\mathbf{0}$ is a $1 \times n_c$ zero vector.

Transforming the graph into an absorbing Markov Chain also enables us to redefine the retention function of a community as the probability that a walker has not been absorbed by RON at time t ($\rho(C, G, t) = 1 - \tilde{x}_{RON}(t)$), where $\tilde{x}_{RON}(t)$ is defined as x in Eq (2.1).

Finally, the fourth advantage of using the retention function is that it compares the raw dynamics. It, therefore, does not use a null-model comparison to quantify the modular structure of a community (often adopted in global quality functions such as Modularity or Markov stability). These properties render the measure mathematically suitable for comparison across different node-aligned networks.

Before introducing our maximum retention distance, we first show how the retention function can be described using eigenvalue decomposition. We use this function, in the next section, to better understand the factors that influence our metric and its properties.

4.4.3 The eigenvalue decomposition of the retention function

Here, we show how we can use the eigenvalue decomposition of the Q matrix to redefine the retention function in terms of its eigenvalues and eigenvectors.

Recall that the Q matrix, as defined in the retention function in Eq. (4.3), is non-symmetric. Therefore to find its eigenvalue decomposition we first define a symmetric version: $Q_{sym} = D^{1/2}QD^{-1/2}$. This ensures that the matrix has real eigenvalues λ_i and orthogonal eigenvectors $\sigma_{\mathbf{i}}$, such that $Q_{sym}\sigma_{\mathbf{i}} = \lambda_i\sigma_{\mathbf{i}}$.

We now use this expression to redefine the retention function as follows:

$$\begin{aligned}
\rho(C,G,t) &= \frac{1}{n_c}\mathbf{1}Q^t\mathbf{1} \\
&= \frac{1}{n_c}\mathbf{1}(D^{-1/2}Q_{sym}D^{1/2})^t\mathbf{1} \\
&= \frac{1}{n_c}\mathbf{1}D^{-1/2}(\sum_i^{n_c}\lambda_i\sigma_{\mathbf{i}}\sigma_{\mathbf{i}}{}^T)^tD^{1/2}\mathbf{1} \\
&= \frac{1}{n_c}\mathbf{1}D^{-1/2}(\sum_i^{n_c}\lambda_i^t\sigma_{\mathbf{i}}\sigma_{\mathbf{i}}{}^T)D^{1/2}\mathbf{1} \\
&= \frac{1}{n_c}\sum_i^{n_c}\lambda_i^t(\mathbf{1}D^{-1/2}\sigma_{\mathbf{i}}\sigma_{\mathbf{i}}{}^TD^{1/2}\mathbf{1}) \\
&= \frac{1}{n_c}\sum_i^{n_c}\lambda_i^t\alpha_i
\end{aligned} \tag{4.5}$$

As $\sum_j Q_{sym}(i,j) \leq 1$, we know that $\lambda_i \leq 1 \forall i$. Hence, the retention function is mainly dominated by the decay of the largest eigenvalue. Note, that as our matrix is not row-stochastic, the largest eigenvalue is not necessarily 1. If, it is 1, the community is completely isolated from the rest of the graph. On the other hand, the smaller its value, the faster the rate of decay and therefore the more connected the community is to the rest of the network.

4.5 Results

To compare the modular structure of a single community across node-aligned networks, we propose the *maximum retention distance*. This measures the largest difference in the retention of a community across a pair or set of networks. In this section, we present our distance, and show its working on various toy models, synthetic- and real-world networks.

4.5.1 The maximum retention distance

We propose the *maximum retention distance* of a community as the maximum (or minimum) difference in the retention distance of a community in two different networks. This measures the largest difference in the probability that a walker will remain within a community across two networks at any given time.

Formally, let $G1 = (V, E1)$ and $G2 = (V, E2)$ be two weighted undirected graphs with the same vertex set, where we wish to compare the modular structure of community $C \in V$ which is connected in both graphs. The retention function is given as $\rho(C, G1, t)$ and $\rho(C, G2, t)$ for each of the graphs, respectively. The retention distance is now defined as:

$$d_{ret}(C, G1, G2, t) = \rho(C, G1, t) - \rho(C, G2, t). \tag{4.6}$$

The maximum retention distance is then given as:

$$d_{max\_ret}(C, G1, G2, t) = \begin{cases} \max_t\{d_{ret}(C, G1, G2, t)\} & \text{if } d_{ret}(C, G1, G2, t) \geq 0, \\ \min_t\{d_{ret}(C, G1, G2, t)\} & \text{if } d_{ret}(C, G1, G2, t) \leq 0. \end{cases} \tag{4.7}$$

As the maximum retention distance is a difference in probability, the distance is bounded: $d_{max\_ret} \in [-1, 1]$. This is a key advantage as it makes the distance **easily interpreted**. A value closest to 0 indicates a small distance and shows that the two communities have a similar modular structure. On the other hand, a larger (absolute) distance value indicates a greater difference in the communities modular structures. As all walkers start uniformly distributed on all nodes within the community in both graphs: $d_{ret}(C, G1, G2, 0) = 0$. Similarly, as all walkers eventually leave the community in both graphs (provided that the community is connected to RON in both graphs): $d_{ret}(C, G1, G2, \infty) = 0$. The retention distance curve therefore both starts and ends at 0. Our distance evaluates the point in between where d_{ret} has the greatest absolute value.

Another advantage is that the distance can also **show which community has the strongest modular structure**. If a positive value (a maximum value) is obtained, the first graph has a stronger modular structure than the second. On the other hand, if a negative value is obtained (a minimum value), the second graph has a stronger modular structure. In some cases, the retention function curve can display both a minimum and maximum retention value. In these cases, we take both the minimum and maximum values and the order in which they occur. We further discuss these cases in Section 4.5.1.2.

By taking the maximum (or minimum) value to summarise the retention curve, we create a single-value distance. The advantage of this summary statistic is that it retains the properties of the retention function. By retaining these properties, our measure is able **to capture the influence of a community's edge density and connectivity** when comparing its modular structure across two graphs. This is unlike the other measures in the literature. Furthermore, it also **doesn't consider the structural topology of nodes outside of the community**, which enables the sole modular comparison of the community of interest. This also increases the use of the distance measure, as we do not need information on the connectivity of nodes outside the community to compare the modular structure of the community. For example, we can adopt this measure to compare a single community in a large network where only partial information is available. Finally, it also **doesn't use a null-model** and can be compared across different networks. This makes the maximum retention distance value mathematically coherent.

4.5.1.1 A toy example

Here we illustrate the workings and advantages of the maximum retention distance using two toy networks. We adopt the same two toy networks as shown in Figure 4.2 (A and B). Recall that existing modular comparison distance measures were unable to capture any differences in the community's modular structure across these two networks.

In Figure 4.4 A and B, we show the same two toy models as in Figure 4.2, but using the graph transformation, we have collapsed all nodes outside the community into an absorbing state "RON". We focus on comparing the modular structure of the community (circled in red) across graph A and B. Recall that the communities have the same inner- and outer-community edge density and only differ in their edge connectivity - which nodes in the community is more strongly connected outwards. In A, node 1 positioned more in the periphery of the community is strongly connected (with an edge of weight 3) to RON. In B, node 6, positioned at the core of the community, is strongly connected to RON. Therefore, the community in A is more isolated than in B.

Next, we show how the probability distribution of the walker changes after each time step by the colouring of the nodes. Initially, we observe that all nodes in the community of both graphs have a similar colour. This is as both graphs start with walkers uniformly distributed amongst all nodes within the community with probability $1/n_c$. In contrast, RON initially has a dark blue colouring representing a probability of zero

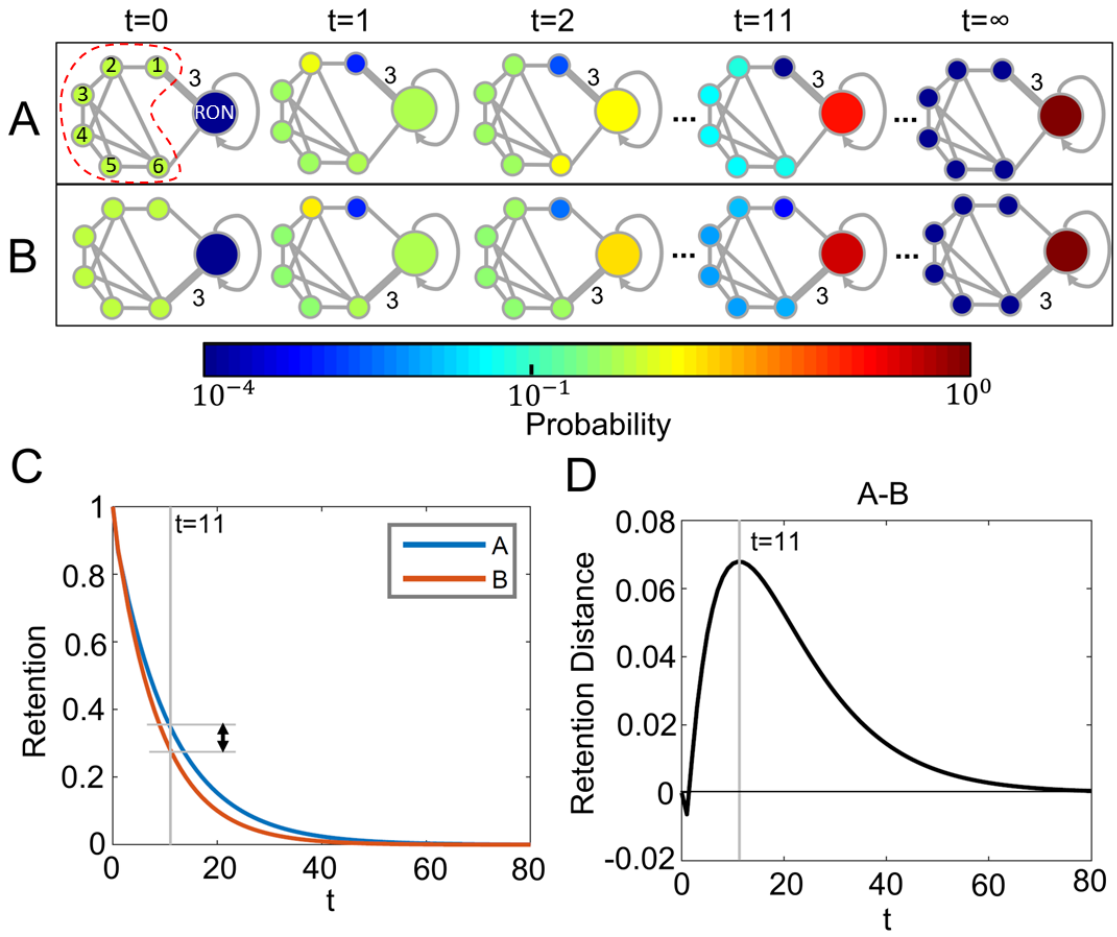


Figure 4.4: The workings of the maximum retention distance shown by comparing a community across two toy networks.(A & B) An illustration of the random walker dynamics on two toy networks at various points of time. The probability of a walker being present on a node is shown via its' colour. We observe that the distributions differ and that the dynamics consider both the community's edge density and connectivity. (C) The retention function for the community in each of the toy networks. The maximum distance between the curves is illustrated at $t = 11$. (D) The retention distance, with the maximum retention distance value, for the comparison of the modular structure of the community across the two networks.

in both graphs. Now, as time increases, we see that the distribution of walkers differs across the nodes in both graphs. First, at $t = 1$, where only the connectivity of the boundary nodes (nodes 1 and 6) are considered, we see that RON has a slightly larger probability distribution in A compared to B. This is as the k_{out}/k ratio of graph A's boundary nodes is greater than those of B. However, as t further increases and we consider more of the community structure, we see that the probability distribution of RON becomes larger in B compared to A. Walkers are therefore able to more easily escape the community in network B compared to network A. The reason is that a core community node (node 6) is connected outwards in network B, while a periphery

node (node 1) is connected outwards in A. We find that the largest discrepancy in RON's probability distribution is at $t = 11$. Eventually, as $t = \infty$ the probability distribution is the same again across the two graphs: where all walkers are absorbed by the absorbing state. Note, that by considering the probability distribution of walkers within the community we are capturing the influence of its edge density and connectivity.

In sub-figure C, we show the corresponding retention function for graphs A and B. Observe that for both graphs, the retention starts at 1 (all walkers are within the community) and ends at 0 (all walkers are absorbed). As the retention curves are different, at intermediate time steps, the retention is capturing the difference in the connectivity of the community across the two networks. This is unlike conductance which only considers edge density, and is unable to capture any difference in the modular structure of the community across these networks.

Finally, in Figure 4.4 D, we show the resulting retention distance obtained when comparing the community across networks A and B. We observe that the retention distance curve has two humps. This shows that initially, the community in graph B is more modular than A as it is able to better retain the walkers at very small times. The boundary nodes in B have a higher k_{in}/k ratio. However, as the walkers consider more of the community connectivity, we see that A becomes more modular than B shown by the positive retention distance values.

Finally, we can see that the maximum retention value lies at $t = 11$. As the value is larger than zero, we see that the maximum retention distance captures the difference in the communities connectivity. Furthermore, as the value is positive, we see that network A has a greater modular structure than network B.

4.5.1.2 Understanding the shape of the curve

As we have seen in our toy example, the retention distance curve can have multiple humps (both a minimum and maximum value). In this case our distance will obtain both a minimum and a maximum value. In this section, we show how differences in the community's modular structure across two graphs results in different shaped curves.

A retention distance curve can either have a single hump (a single maximum or minimum value) or multiple humps (a minimum and a maximum value). To illustrate which community structure produces which type of curves, we construct and compare toy networks. Consider the four toy networks illustrated in Figure 4.5 A-E. Where we compare the modular structure of the blue community in network A to its' structure

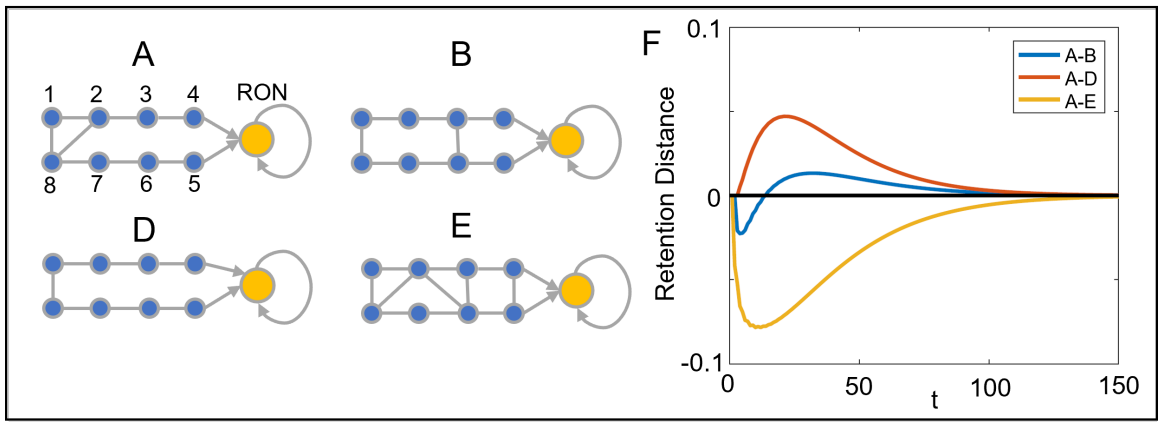


Figure 4.5: The comparison of toy networks resulting in retention distance functions with different shapes. (A-E) Four toy networks where we compare the blue community in graph A to its structure in each success graph. The rest of the network is represented by an absorbing state. (F) A graph showing the resulting retention distance for each pairwise comparison.

in each successive network. First, we compare the community across networks A and D. We can spot that the community has one less inner-community edge in D than in A, and is therefore less modular. The corresponding retention distance curve (in sub-figure F) has a convex single-hump shape with a positive maximum - showing that the community in A is more modular than in D. Next, we compare the community across networks A and E. Here we see that the community is more modular in E than A - as it has three extra inner-community edges. In F, we observe a concave single-hump shape with a negative minimum - showing that the community is less modular in A.

Finally, we compare the community across networks A and B. We observe that both graphs have a similar density of edges. However, B has an edge $(3, 6)$ which is much closer to RON, while A's edge $(2, 8)$ is much deeper in the community. Therefore, A traps more flow than B and is more modular. In F, we see that the corresponding retention distance has a multi-hump shape. First, we observe a concave hump with a local minimum and then a convex hump with a local maximum. The measure is therefore picking up that at small times (when the walker has not yet explored the whole community) the community in B is more modular than A. This is as it seems to have an extra inner-community edge. However, at larger times, when the walker explored all of the community we see that it is now more modular in A than in B.

The shape of the curve can therefore inform us how the structure of the communities differs. Therefore, when calculating the maximum retention distance it is important to consider the shape of the retention curve. If we have a single-hump

curve we can take the global maximum or minimum value. However, if we have a multi-hump curve, we need to consider the local maximum and minimum values of the curve and the order in which they occur.

4.5.1.3 Distance confidence interval

Before showing the properties of our distance measure using synthetic networks, we first introduce the methodology we follow to obtain a distance confidence interval. A key property for the usability of any distance measure is its confidence interval. This is because it measures whether the distance is actually picking up a significant difference within the community structure across the graph set or merely noise.

To obtain a confidence interval for the maximum retention distance we adopt a similar approach as [184]: a parametric bootstrapping to create an ensemble of bootstrap networks for each network in the comparison. We then use the two ensembles to construct a distribution of maximum retention distances on which a confidence interval is obtained.

More specifically, we adopt the following steps:

1. For each of the graphs in the comparison, we use parametric bootstrapping to obtain a large ($N_b = 1000$) sample of bootstrap networks. We uses the network's edge weights to parameterise and re-sample. We, therefore, treat each edge as an independent event and re-sample the edge weight from a Poisson distribution with the original network's edge weight as the mean. Although we adopt this re-sampling technique, other techniques can also easily be adopted (for example see [51]). Note that at the end of this step we obtain two ensembles of N_b bootstrapped networks for each of the graphs in the comparison.

2. We then calculate the maximum retention for each pair of graphs obtained from each set. We therefore obtain N_b^2 distance values.

3. Finally, to obtain a confidence interval for our summary statistic, we assign a 95% bootstrap confidence interval spanning the 2.5th and 97.5th percentiles of the bootstrap distribution [51].

4.5.2 Properties of the maximum retention distance

In this section, using synthetic networks we show how the maximum retention distance is able to capture both the influence of a community's edge density and edge connectivity when comparing its' modular structure across two graphs. This is the

key feature which is currently ignored by current distance measures in the literature and the key feature we wish to compare across our skill-relatedness network. We also show how the distance adheres to normal network distance behaviour which makes it interpretable and more widely applicable.

4.5.2.1 Capturing edge density and connectivity

Here, we use synthetic networks to illustrate how the maximum retention distance is able to capture both the influence of edge density and edge connectivity when comparing the modular structure of a community across a pair or set of networks. To do this we conduct experiments in which we apply successive perturbations on an original graph and measure, after each perturbation, the maximum retention distance between the obtained new graph and the original one.

We start by evaluating how the maximum retention function performs when we change the community edge density (the number of edges of a community). Here we apply the following two perturbations: (1) decreasing the **inner-community edge density** by randomly removing edges (which reduces the community's modular structure) and (2) increasing the **outer-community edge density** by randomly adding edges (which reduces the community's modular structure).

We start by constructing a modular base network. We construct an SBM with a pre-defined community of size $n_c = 100$ (which we will compare across different perturbed graphs). The rest of the network consists of 100 nodes, making the graph size $n = 200$. Note, that we do not specify the structure of the nodes in the rest of network as it is not captured by our measure (the measure collapses these nodes into an absorbing state RON). The nodes within our focus community have an ER graph structure, with a constant inner- ($p_{in} = 0.8$) and outer-edge probability ($p_{out} = 0.2$). We illustrate an instance of the base network in Figure 4.6 (A), where we mark the community we wish to compare in red.

First, we investigate how our measure performs when we decrease the inner-community edge density. To do this we randomly remove edges within the community. Thereby, reducing the community's modular structure. Note that we are reducing the inner-community edge density (by reducing the community's inner-edge probability p_{in}) but keeping the community's connectivity (the ER graph structure) constant. In Figure 4.6 (B), we show the resulting maximum retention distance obtained when comparing our base network to successive perturbed networks. Note

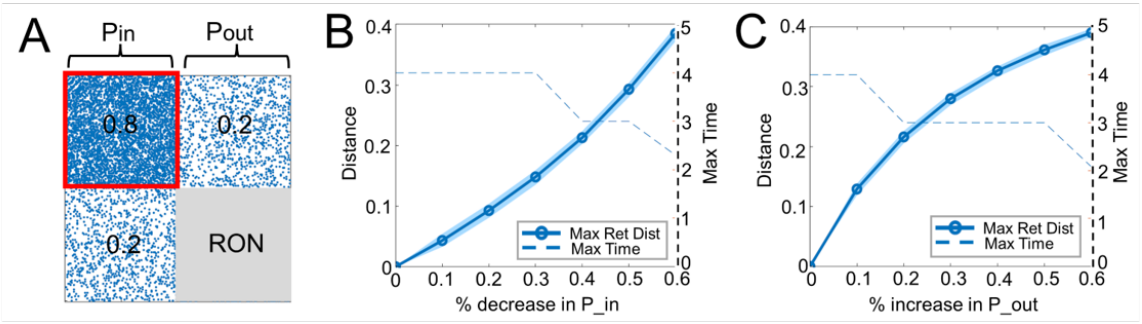


Figure 4.6: The impact of a change in inner-community and outer-community edge density on the maximum retention distance. (A) Our base network, where we compare the modular structure of the community circled in red to its structure in successive graphs of similar structure. (B-D) The resulting maximum retention distance when comparing the network in A to a second network with the same structure but where we (B) reduce p_{in}, (C) increase p_{out}. We observe that the distance can capture the influence of changes in density on the community modular structure and complies with normal distance behaviours.

that our maximum retention distance is obtained from 1000 network instance comparisons. We observe that as we reduce the inner-community edge density, the maximum retention distance increases. Therefore, our distance can capture the influence of inner-community edge density.

Next, we test how our distance performs when we increase the outer-community edge density. Again we compare our base network A to successive networks in which we randomly add additional outer-community edges. Here, we are increasing the community's outer-edge density (p_{out}), without changing its connectivity (it continues to have an ER graph structure). In sub-figure (C) we show the resulting maximum retention distance. We observe that as the outer-community edge density increases, our distance also increases. Therefore, our measure is also able to capture the influence of edge density when comparing a community across networks.

Next, we turn to investigate how the maximum retention distance performs when we change the community's connectivity (the distribution of edges). Here we apply the following two perturbations: (3) we change the **inner-community connectivity** and (4) change the **outer-community edge connectivity**. We apply both of these changes by randomly shuffling the edges, thereby randomising the community's connectivity.

To evaluate the **inner-community connectivity**, we start with a base SBM network with a pre-defined focus community of 100 nodes. However, unlike our normal SBM, our community has a BA inner-community structure (characterised by a power law degree distribution). The outer-community structure is that of an ER graph with a constant edge probability ($p_{out} = 0.1$). We illustrate this base network as A1 in

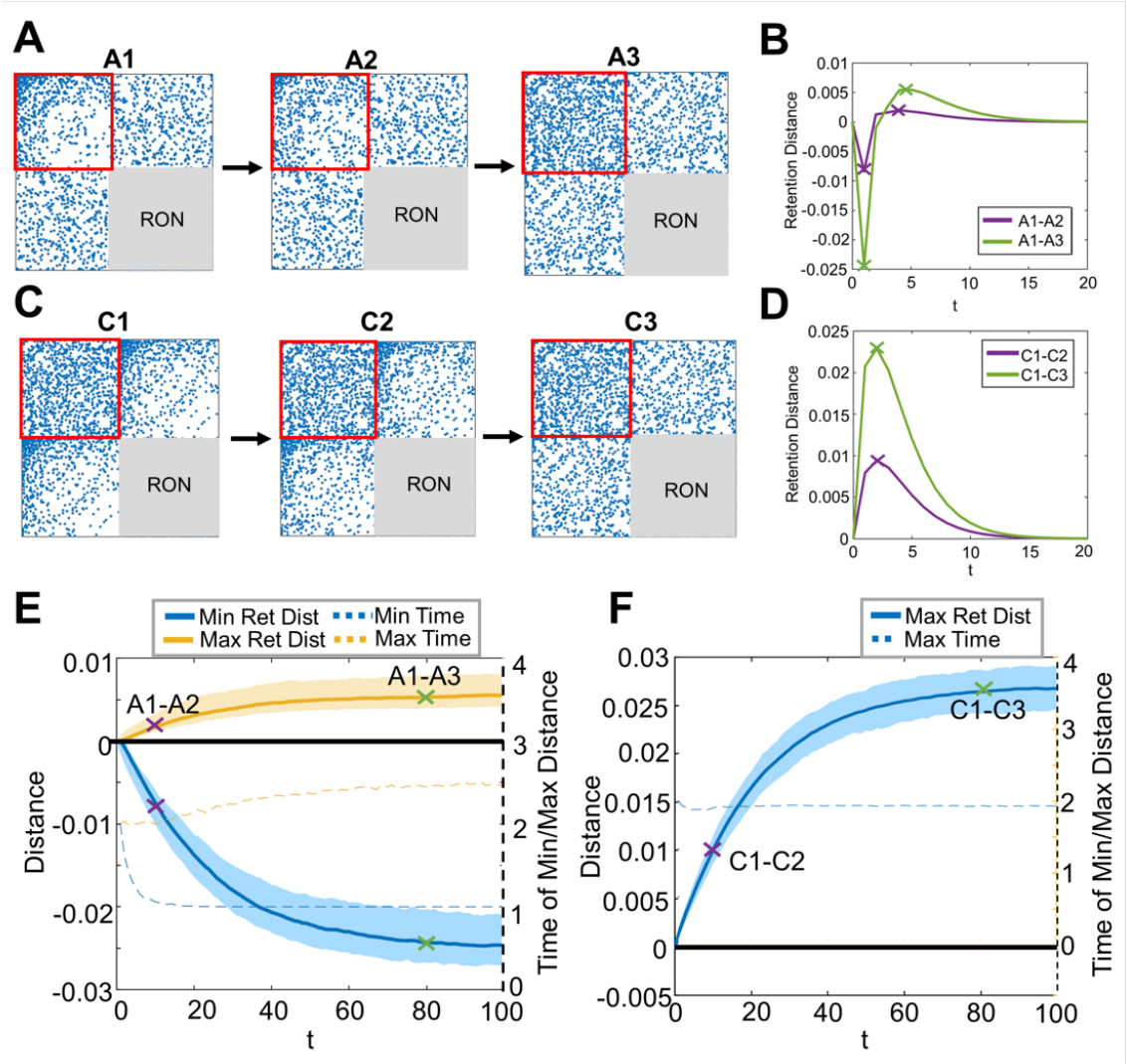


Figure 4.7: The impact of a change in community connectivity on the maximum retention distance. (A) we show three SBM networks, where we compare the modular structure of the community circled in red in A1 to its structure in successive networks. A1 is characterised by a BA inner-community structure and an ER outer-community structure. A2 and A3 are constructed by 100 and 800 random inner-community edge switches of A1. (B) we illustrate the retention distance curve of A1-A2 and A1-A3. (C) We show three SBMs, where we again compare the modular structure of the community circled in red in C1 to its structure in successive networks. C1 is characterised by ER inner-community structure and BA outer-community structure. C2 and C3 are constructed through 100 and 800 random outer-community edge switches of C1. (D) We illustrate the retention distance curve of C1-C2 and C1-C3. In (E) & (F), we illustrate the maximum (and minimum) retention distance obtained for successive random edge switches of the graphs in A and C, respectively.

Figure 4.7 (A). We then successively randomly shuffle the inner-community edges to produce a community with a more ER edge structure. Note that we are changing the inner-community connectivity from heterogeneous to more homogeneously distributed across the community nodes. We also illustrate network instances after randomly

shuffling 100 (A2) and 800 (A3) inner-community edges. As A3 is fully-randomised, it has ER inner-community structure. Note that as the community becomes more randomised, it loses its' core-periphery structure and becomes less modular. Furthermore, as the community edge density remains constant (we are not adding or removing any edges), we are only changing the community's edge connectivity.

Before investigating the impact of these changes on the resulting maximum retention distance, we illustrate the retention distance curve obtained when comparing A1-A2 and A1-A3. We observe a multiple-hump-shaped curve with an initial negative concave shape (with a local minimum) followed by a positive convex shape (with a local maximum). This shows that the more randomised graph is initially more modular (as boundary nodes have an expected higher percentage of inner-community edges). However, at larger times, we see that the randomised graph is less modular (as the walker explores more of the community, it picks up that the first graph has a core-periphery structure and is more modular).

Next, in sub-figure (E), we show the resulting maximum retention distance (both the local minimum in blue and the local maximum in yellow). We indicate the comparison of A1-A2 and A1-A3 on the curve with a purple and green cross, respectively. We see that as we randomise the inner-community edge connectivity, the distance increases (both for the initial minimum and in the latter maximum value). However, after many perturbations, the distance starts to saturate at an asymptotic value. This is because the graph has become fully randomised and remains so for further perturbations. We again highlight, that unlike the distance measures in the literature, our measure is able to capture how changes in the inner-community connectivity changes the community's modular structure.

We now repeat a similar investigation but focus on how the maximum retention distance performs when we change the **outer-community connectivity**. Here, we again use an SBM with a pre-defined community of 100 nodes. The community has an ER inner-community structure (constant p_{in} value) and a BA outer-community structure. We illustrate an instance of this network as C1 in Figure 4.7 (C). A heterogeneous outer-community structure, therefore, characterises the community. Similar to our previous experiment, we successively randomly shuffle the outer-community edges and investigate their impact on the maximum retention distance. We illustrate two network instances obtained after randomly shuffling 100 (C2) or 800 (C3) outer-community edges. As C3 is fully randomised, it displays ER outer-community structure. Note that as we are not adding or removing any edges the outer-community edge density remains constant, and only its connectivity is changing.

In sub-figure D, we again show the resulting retention distance curves (and where the maximum retention lies on the curve) obtained when comparing the community in C1-C2 and C1-C3. In this case, we see a single hump with a maximum value. We show the maximum retention distance obtained after each perturbation in sub-figure F. We observe that as we randomise the outer-community connectivity of the community, the distance increases. Therefore, our distance measure is able to capture the impact of a more significant change in the community's outer connectivity. The distance also displays a similar asymptotic behaviour, showing that further perturbations only produce similar and fully randomised graphs.

Through our experiments (and toy networks), we have shown that the maximum retention distance can capture the influence of both edge density and connectivity when comparing the modular structure of a community across a set of networks.

4.5.2.2 Adherence to normal network distance behaviour

Another key property of our distance measure is that adheres to "normal" network distance properties. [195] and [118] proposed a list of axioms that all network distance measures should conform to.

The properties that apply to our community comparison measure include:

1. The distance between very similar graphs should be small. Hence, the distance should tend to zero when network perturbations tend to zero.

2. When applying successive network perturbations to the graph, the distance should increase monotonically with the number of perturbations. Hence, any fluctuations in the distance should be limited.

3. When applying perturbations that tend to randomize the structure of the graph, the distance should saturate to an asymptotic value after a larger number of perturbations. This is because after many perturbations the graph will be fully randomised. Hence, the graph will only remain so after any further perturbations.

4. A perturbation on a sparse graph (with only a few edges) is more important than a similar perturbation in a denser graph of equal size (this property is also known as the property of edge-"sub-modularity").

Using the results from our previous perturbation experiments we are able to evaluate how well our distance adheres to these properties. Consider the first property.

We can see in both Figure 4.6 and Figure 4.7, that our distance starts at zero and then increases in magnitude as we apply our various perturbations. This shows that few perturbations result in small distances. We also observe, that our distance increases monotonically and shows no large fluctuations as we increase the number of perturbations. Thereby also adhering to the second property.

Next, we focus on the perturbation experiments in which we randomised the inner- and outer-community edges. These results are shown in Figure 4.7. Here, we observe for both experiments, as we increased our perturbations our distance increased. However, after many perturbations, the distance starts to saturate at an asymptotic value. This is because the graph is now fully randomised and further perturbations do not change its connectivity - the graph remains fully randomised with an ER graph structure. Our results show that our distance also adheres to the third property.

To evaluate the fourth property, we again consider the first two experiments where we decreased and increased the inner- and outer-community edge density, respectively. For both of these experiments, we observe that the maximum retention distance gradient is steepest when the inner- or outer-community edge density is lowest. For the inner-community edge density (shown in Figure 4.6 (B)), this is when p_{in} is smallest and thereby when the most perturbations have occurred. For the outer-community edge density (shown in Figure 4.6 (C)), this is when p_{out} is smallest and when the least perturbations have occurred. This shows that the sparser the community, the larger the increase in the distance value and therefore the more significant the impact of a change. Our distance measure, therefore, also complies with the "Edge-submodularity" property.

We have shown that the maximum retention distance is both able to capture the impact of changes in a community's edge density and connectivity, as well as, adhere to normal network distance behaviours. This is advantageous as it shows the measure is able to capture the community's topology and can easily be used and interpreted as a distance measure.

In the construction of our method, we also considered the usage of other summary statistics on our retention curve (including using the retention distance value at $t = 1$, using the integral of the retention curve and comparing the largest eigenvalue of the system matrix Q). This is instead of using its maximum value. However, each of these methods where either unable to capture the impact of changes in both edge density and connectivity or did not adhere to these normal distance behaviours. We illustrate where these metrics fail in Section B.0.1 of Appendix B.

Next, we show how we can harness the time at which the maximum retention distance occurs to gain further information on how communities differ.

4.5.3 Maximum retention time

A key property of the maximum retention distance, is the time at which it occurs. In this section, we investigate what influences this value and whether it provides any additional information on how the community's modular structure differs across a set of graphs. To do this, we start by deriving an exact algebraic expression for the maximum retention time for a special case community (a community with a constant inner- and outer-community strength across two graphs). We use this expression to gain insight into some of the factors that influence the maximum retention time. We then use toy networks to investigate the impact of a community's edge density and connectivity on the maximum retention time of other community structures more generally.

4.5.3.1 Maximum retention time for a special case community

First, we construct and compare a special case community C across two graphs ($G1$ and $G2$). Each node within the community has a constant inner- (k^1_{in} in $G1$ and k^2_{in} in $G2$) and outer-community strength (k^1_{out} in $G1$ and k^2_{out} in $G2$). Note, that these special case communities have the same structure as the expected structure of a community in an SBM with an ER structure defined with a constant inner- (p_{in}) and outer edge probability (p_{out}). This also represents a community with a homogeneous edge structure (uniform degree distribution) and is a structure that is often used as a modular network benchmark.

For these special case communities, we can derive an exact expression for their retention functions. Focusing on the first community in $G1$, we know its retention function is given as $\rho(C, G1, t) = \frac{1}{n_c}\mathbf{1}Q^t\mathbf{1}'$ by Eq. (4.3). For this community, the row-sum of its Q matrix is $\frac{k^1_{in}}{k^1_{in}+k^1_{out}}$. Therefore, $Q\mathbf{1} = (\frac{k^1_{in}}{k^1_{in}+k^1_{out}})\mathbf{1}$. Notice that Q, therefore, has eigenvalue $\lambda_{max} = \frac{k^1_{in}}{k^1_{in}+k^1_{out}}$ with corresponding eigenvector $\mu = \mathbf{1}$. For this graph, we can now exactly define the retention function in terms of the largest

eigenvalue. This is given as:

$$\begin{aligned}\rho(C, G1, t) &= \frac{1}{n_c}\mathbf{1}Q^t\mathbf{1} \\ &= \frac{1}{n_c}\mathbf{1}\lambda_{max}^t\mathbf{1} \\ &= \lambda_{max}^t \\ &= (\frac{k_{in}^1}{k_{in}^1 + k_{out}^1})^t\end{aligned} \tag{4.8}$$

For this case, we see that the largest eigenvalue exactly defines the retention function of a graph. The same holds for the special case community C in $G2$.

We denote the largest eigenvalue of the Q matrix for C of $G1$ and $G2$ as λ_1 and λ_2, respectively. The retention distance between the community in both these graphs is then given as $d_{ret}(C, G1, G2, t) = \lambda_1^t - \lambda_2^t$, and the maximum retention distance as $\max\{\lambda_1^t - \lambda_2^t\}$.

We can now derive the time at which the maximum retention occurs as follows:

$$\begin{aligned}\frac{\partial}{\partial t}(\lambda_1^{t_{max}} - \lambda_2^{t_{max}}) &= 0 \\ \frac{\lambda_1^{t_{max}}}{\ln(\lambda_1)} &= \frac{\lambda_2^{t_{max}}}{\ln(\lambda_2)} \\ (\frac{\lambda_1}{\lambda_2})^{t_{max}} &= \frac{\ln(\lambda_1)}{\ln(\lambda_2)} \\ t_{max}\ln(\frac{\lambda_1}{\lambda_2}) &= \ln(\frac{\ln(\lambda_1)}{\ln(\lambda_2)}) \\ t_{max} &= \frac{\ln(\frac{\ln(\lambda_1)}{\ln(\lambda_2)})}{\ln(\frac{\lambda_1}{\lambda_2})}\end{aligned} \tag{4.9}$$

Hence, λ_1 and λ_2, which represent the edge density of the two communities, are the key factors influencing the maximum retention time.

We now use this expression to investigate how the magnitude of λ_1 and λ_2 influences the size of maximum retention time. We highlight that $\lambda_1 < 1$ and $\lambda_2 < 1$. To obtain a large maximum retention time, we want a large numerator ($\ln(\frac{\ln(\lambda_1)}{\ln(\lambda_2)})$) and small denominator ($\ln(\frac{\lambda_1}{\lambda_2})$).

First, we consider the case where $\lambda_1 > \lambda_2$. A small denominator can be obtained if $\frac{\lambda_1}{\lambda_2}$ is close to 1 - both communities need to have similar edge densities. A large numerator can be obtained if $\frac{\ln(\lambda_1)}{\ln(\lambda_2)}$ is close to zero. This can occur if λ_1 is close to 1 - a community with a high edge density. Hence, large times are found if both communities have high and similar densities. Note that this corresponds with a small maximum retention distance.

On the other hand, we assume that $\lambda_1 < \lambda_2$. Now, to obtain a small denominator $\frac{\lambda_1}{\lambda_2}$ again needs to be close to 1 - thereby both communities need to have similar densities. To obtain a large numerator, $\frac{\ln(\lambda_1)}{\ln(\lambda_2)}$ needs to be large. Therefore λ_2 needs to be close to zero - a community with sparse edge density. For this case, large times are obtained if both communities are sparse and have similar densities. Note that this also results in a small maximum retention. From both cases, we see that large maximum times are obtained with corresponding small maximum retention distances.

4.5.3.2 Maximum retention time for general community structure

Next, we investigate how the maximum retention time is impacted by changes in a community's modular structure more broadly (not just in our special case). Using a set of toy networks, we investigate the influence of changes in community edge density and connectivity on the maximum retention time (as well as its relationship to the maximum retention value).

First, we probe the impact of changing the community's edge density (while keeping its connectivity constant) on the maximum retention time. To do this we construct a clique-like community where all nodes are connected outwards to RON. We then successively reduce the inner-community edge weights thereby reducing the inner-community edge density but keeping the same community edge connectivity. This also reduces the modular structure of the community. We compare the original community to each successive community and investigate how this change impacts the maximum retention distance and the maximum retention time.

In Figure 4.8 A, we illustrate four toy networks (A-D) with this clique-like inner-community structure and where each node is connected to RON. In each successive community, we see a decrease in the inner-community edge weights and thereby also a reduction in community modular structure. We compare the modular structure of the blue community in A to its structure in B, C and D. In sub-figure E, we plot the difference in the community's edge density and the difference in the community depth versus the maximum retention time. We measure the edge density by the conductance - the ratio of the number of outer-community edges over the volume of the community. We measure the community depth as the maximum FMPT of a walker starting on any node in the community reaching RON - the largest number of steps it takes for a walker to be absorbed by RON. We observe that for each successive graph comparison the magnitude of the difference in community edge density and depth increases.

In sub-figure F, we plot the resulting retention distance curve for each comparison. Note that as we reduce the inner-community density, the maximum retention

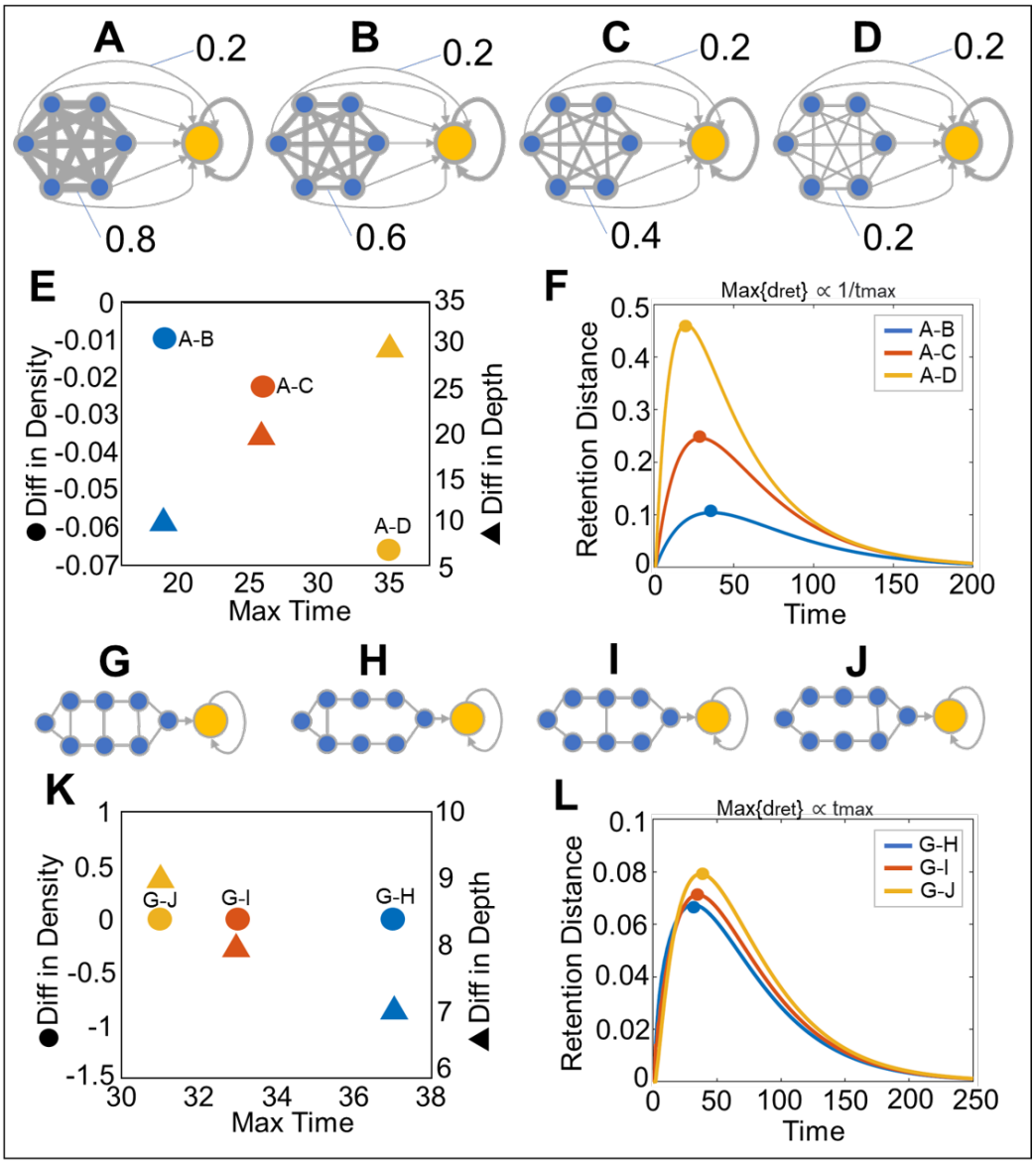


Figure 4.8: Toy network comparison illustrating how edge density and connectivity influence the maximum retention time. (A-D) Four clique-like toy networks with where we reduce the inner-community edge weights in each successive network. (E) The relationship between the difference in the density (measured as conductance) and depth (measured as the maximum FMPT) of the blue community in A to each of the other graphs and the maximum retention time obtained. (F) A graph showing the retention distance for each pairwise community comparison. We observe an inversely proportional relationship between the maximum retention and the maximum time. (G-J) An illustration of four toy networks where in each network we remove edges deeper in the community. (K) We show the difference in the community edge density and depth of the blue community in G and its structure in each of the other graphs. (L) We illustrate the retention distance for each consecutive graph pair comparison. Here, we observe a positive relationship between the maximum retention and the maximum time.

distance (the magnitude of the curve's peak) increases. Furthermore, the time at which the curve peak decreases. We, therefore, observe, that a larger change in the

inner-community density increases the maximum retention distance but reduces the maximum retention time. Here, we observe that the maximum retention is inversely proportional to the maximum time.

Next, we probe the impact of edge connectivity on the maximum retention distance and the maximum time. More specifically, we investigate the impact of changing the depth of the community while keeping the number of edges constant. To do this, we remove two edges in the community that are successively deeper (further away from RON) in the community. In sub-figure G-J we illustrate 4 toy networks where we compare the modular structure of the blue community in G to its structure in each of the other networks. Network H-J, are constructed by removing two edges in G - where we remove two edges that are successively deeper within the community. In K, we again plot the corresponding difference in density and depth for each network comparison versus the maximum retention time. We see that the magnitude of the difference in depth increases, but the difference in edge density remains zero and constant.

In sub-figure L, we show the resulting retention distance curves for each comparison. In contrast to the influence of edge density, we see an increase in both the maximum retention distance and the maximum time. Intuitively, this relationship makes sense, as the deeper within the community the difference occurs; the longer the walker needs to explore the community to pick up this difference. Here, the maximum retention is positively related to the maximum retention time.

When comparing a set of networks, we can harness the relationship between the maximum retention and the maximum time to unveil potential cases where the community differs across the graphs either through a difference in edge density or connectivity. We also highlight, that there may be other factors that influence the maximum retention time - such as the size of the community. As we compare communities of the same size, this variable's influence is negligible in our comparison.

4.5.4 Application to inter-industry labour flow networks

In this section, we illustrate the use of our maximum retention distance on two sets of real-world networks.

4.5.4.1 Comparing inter-industry labour flows across European countries

We start by investigating how the modular structure of individual communities within the Irish SRN compare to their structure in other European countries' SRNs. We

therefore find the modular structure of the Irish SRN and use its' structure as our pre-defined communities. We then compare the modular structure of each community across the different countries' SRNs using our maximum retention distance.

First, we illustrate the adjacency matrices of the different countries' SRNs in Figure 4.9 A. Recall, that when we compared the global modular structure of Ireland's SRN to the other countries' SRNs using the BiDir distance, we found that Ireland's modular structure is closest (and a nested form) of the German SRN's structure. Now, we further investigate to which degree this is true for each community individually. In other words, are certain communities closer to the German structure than other?

Next, in sub-figure B, we visualise the Irish SRN as a network. Each node in the network represents an industry, and each edge the skill-relatedness between its two corresponding industries. The node layout is based on a spring algorithm, namely 'Force Atlas' in Gephi, where more skill-related industries are positioned closer together. The modular structure of the network is shown via node colouring. The modular structure is found using the Markov Stability community detection algorithm. We choose a robust (defined as a partition with a low variation of information) partition which divides the nodes into 18 communities. We also label the communities alphabetically and indicate the industrial sector they best represent.

We are particularly interested in comparing the modular structure of the three sectors - circled in red in the figure. First, we consider the Irish Financial sector (community 12). We have observed that this sector is highly isolated from the broader labour market in the Irish case. In C and F, we show the resulting retention distance and the maximum retention distance, respectively. For the retention distance, we observe a single-hump shape with a maximum for all pairwise comparisons. We see that the financial sector is particularly isolated in the Irish case, and most similar to the UK. We also show the comparison of this community to its structure in the Irish SRNs configuration null model. This provides a reference for the distance - the distance obtained when comparing the community to a graph which has no modular structure. As all comparisons are smaller than the comparison with the configuration model, we known that this community has some form of modular structure in all the networks.

Next, we focus on comparing the Software and Telecommunication sector (community 16). Similarly, we show our resulting retention distance and its maximum in sub-figure D and G. In D, we observe that the retention distance curve is multi-hump with an initial local minimum and a later local maximum for all pairwise comparisons. This shows us that initially, all other countries are more modular (have a

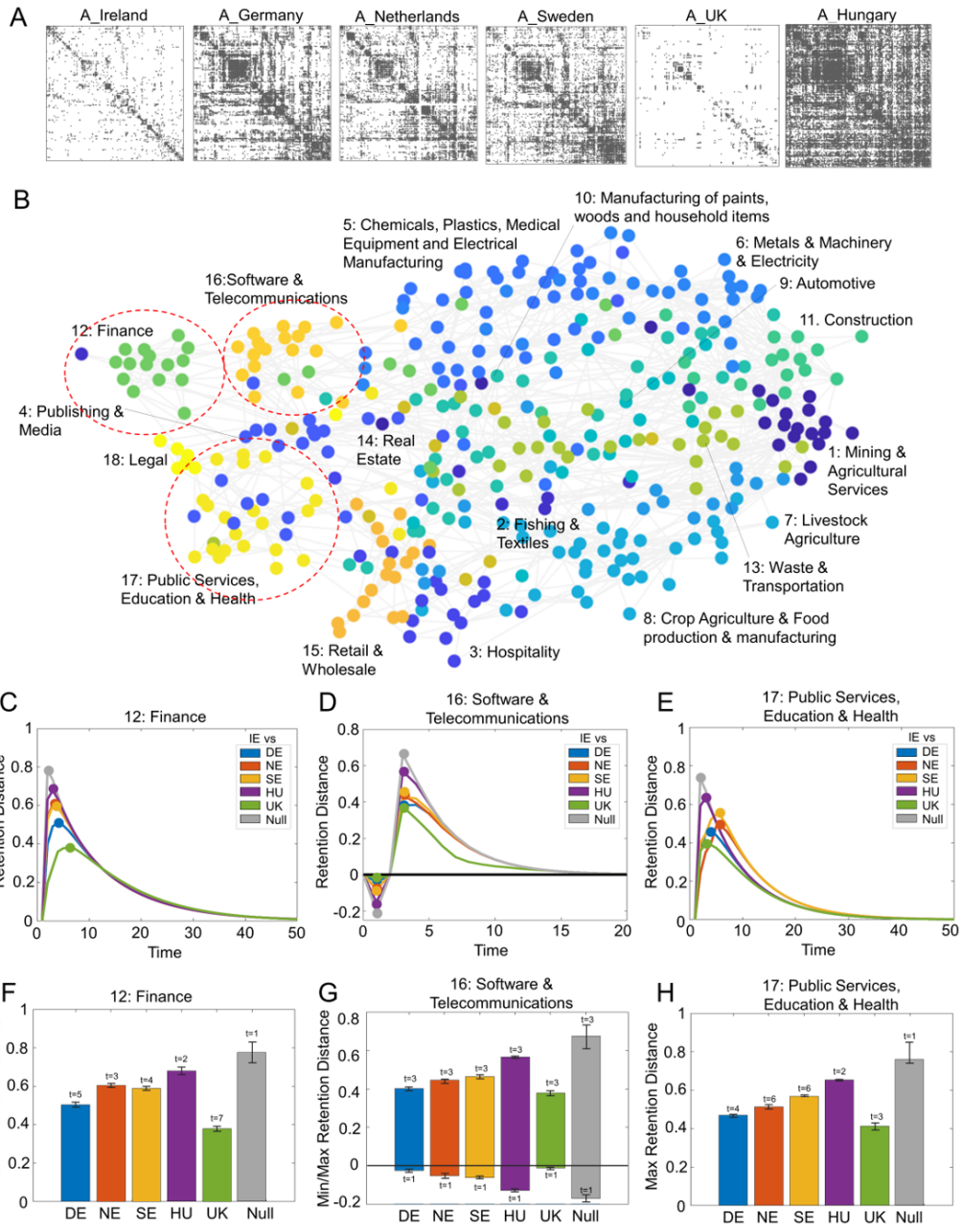


Figure 4.9: The comparison of the modular structure of three communities in the Irish SRN to their structure in different European countries' SRNs. (A) The adjacency matrices of 6 different European countries' SRNs. (B) A visualisation of the Irish SRN where nodes represent industries and edges the degree of skill-overlap between the two related industries. The node colouring indicates the different communities found using the Markov stability community detection algorithm. We encircle three communities which we intend to compare across the different countries. (C-E) We show the retention distance curve obtained for the comparison of each of these communities across the different countries. (F-H) We show how we summarise each of these curves using the maximum retention distance, respectively.

higher relative inner-strength boundary). However, as the walker explores more of the connectivity, the Irish community becomes more modular. Hence, in the Irish case, the boundary nodes are not as well connected inside the community as in other countries. But, the core nodes are more internally well connected and more isolated from the rest of the labour market.

Finally, we compare the Public Services, Education and Health sector. We again show the retention distance and its maximum in sub-figures E and H, respectively. Here we see that the UK sector is again the closest in structure. What is interesting about this community comparison is the relationship between the maximum retention distance and the maximum time. We observe that when comparing Ireland to the UK, Germany and the Netherlands: as the retention distance increases, so does the time at which it occurs. This shows that the difference in modular structure is occurring deeper in the community. When more closely investigating the community in these three networks, we observe that they have similar densities, and that it is in fact the connectivity of edges that is causing the difference - in each successive comparison (UK, Germany and the Netherlands), the inner-community edge connectivity is deeper in the community.

In each of the above sectors, we see a similar ranking of countries in their modular distance to Ireland. However, this is not the case for all communities. In Figure 4.10, we show the maximum retention distance obtained for each of the 18 communities. Note that apart from community 16[2], all communities display a single-hump retention distance curve. Furthermore, we do not include the UK in the comparison when the community is not connected in its' SRN.

In sub-figure A, we show a bar-graph of the maximum retention distances obtained for the various community comparisons. We order the communities on the x-axis according to their size (number of industries within the community) - where 5 is the largest community and 2 the smallest. Note that the number represents the community label. Below each bar, we colour the label of the community by the country in which it is closest in structure.

In sub-figure B, we again illustrate the Irish SRN, but now colour each node by the country its community is closest to in structure. In general, we find that service-related sectors (positioned on the left of the Irish SRN) are most similar in their structure to the UK SRN. In contrast, manufacturing-related sectors (positioned on the right of the network) are overall closest in structure to the German SRN. Only very few sectors are closest in structure to either the Netherlands or Sweden. This shows

[2]For community 16 we only display the maximum retention in the figure.

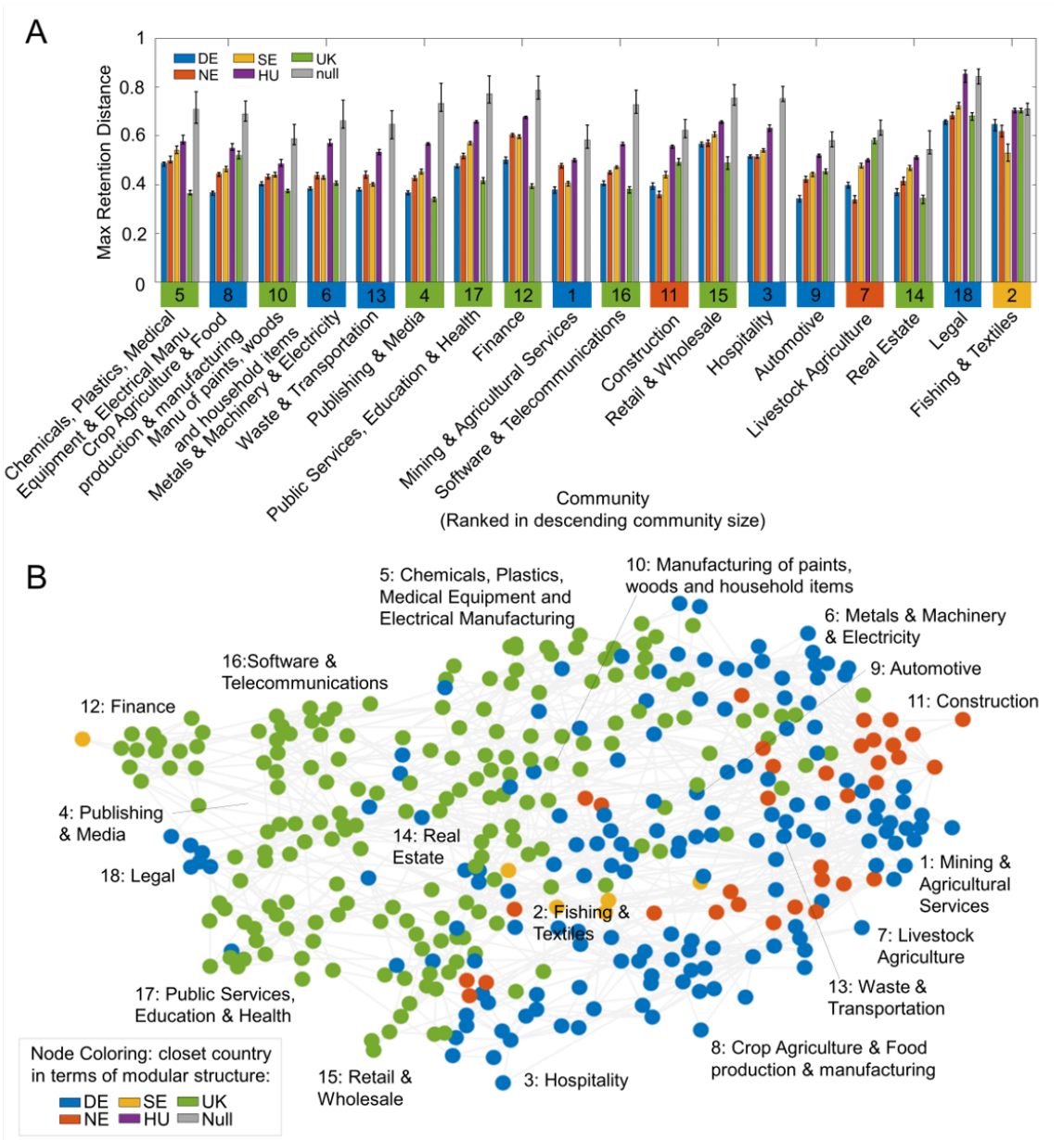


Figure 4.10: (A) bar graph showing the maximum retention distance between each of the communities in the Irish SRN and the communities in 5 other European countries' SRNs. (B) A visualisation of the Irish SRN, where nodes are coloured by the country to which their community is most similar in modular structure.

that there is heterogeneity in the distance between different communities' modular structure across the network set.

4.5.4.2 Comparing gendered inter-industry labour flows

As a second real world application of our method, we compare the individual communities (skill basins) within the male and female SRN for South Africa. Comparing

the communities in these networks enables us to investigate in which communities female (or male) workers are more isolated from the broader labour market.

In Figure 4.11 A we show a visualisation of the South African SRN (constructing using both male and female labour flows). Each node in the network represents an industry and each edge the skill-relatedness between its two corresponding industries. The industries are classified according to an internal industry classification. The node layout is again based on a spring algorithm 'Force Atlas' in Gephi, where more skill-related industries are positioned closer together. The modular structure of the network is shown via the node colouring. Furthermore, each of the 14 communities are labelled alphabetically from A to N along with the industrial sector they best represents. The communities were detected using the Markov Stability Algorithm. These communities represent South Africa's skill basins or industrial clusters.

Here, we are interested in comparing the modular structure of these communities across the male and female SRN to investigate the degree to which they differ. We therefore use these pre-defined communities and superimpose them onto the male and female SRN. First, we show the male and female SRN adjacency matrices in sub-figure B and C, respectively. Each of these networks were constructed using only male or female labour flows, respectively. The two networks differ only in their edges. We also highlight, that when calculating the skill-relatedness (methodology reviewed in §2.5) we normalized by the total size of male and female labour flows. Therefore, the measure inherently normalizes for the size of male and female workers within each industry, respectively.

We show the resulting maximum retention distance obtained when comparing each of the sectors (labelled in A) between the male and female SRN via the bar graph in sub-figure D. We obtained a single hump retention distance curve for all community comparisons. We therefore only display the maximum retention distance. Furthermore, as we are only comparing two communities (and not a set) we are not able to harness the maximum time. We order the communities on the x-axis of the bar-graph by descending size (number of industries within the sector). Observe, that we obtain both positive and negative distances. A positive distance indicates that the community is more isolated in the male case, while a negative distance indicates that the sector is more isolated in the female case. We colour the sector labels (on the x-axis of the bar graph) according to whether the sector is more isolated in the male (blue) or female (red) case. If the label is not coloured, we observe that the sector does not (significantly) differ in modular structure.

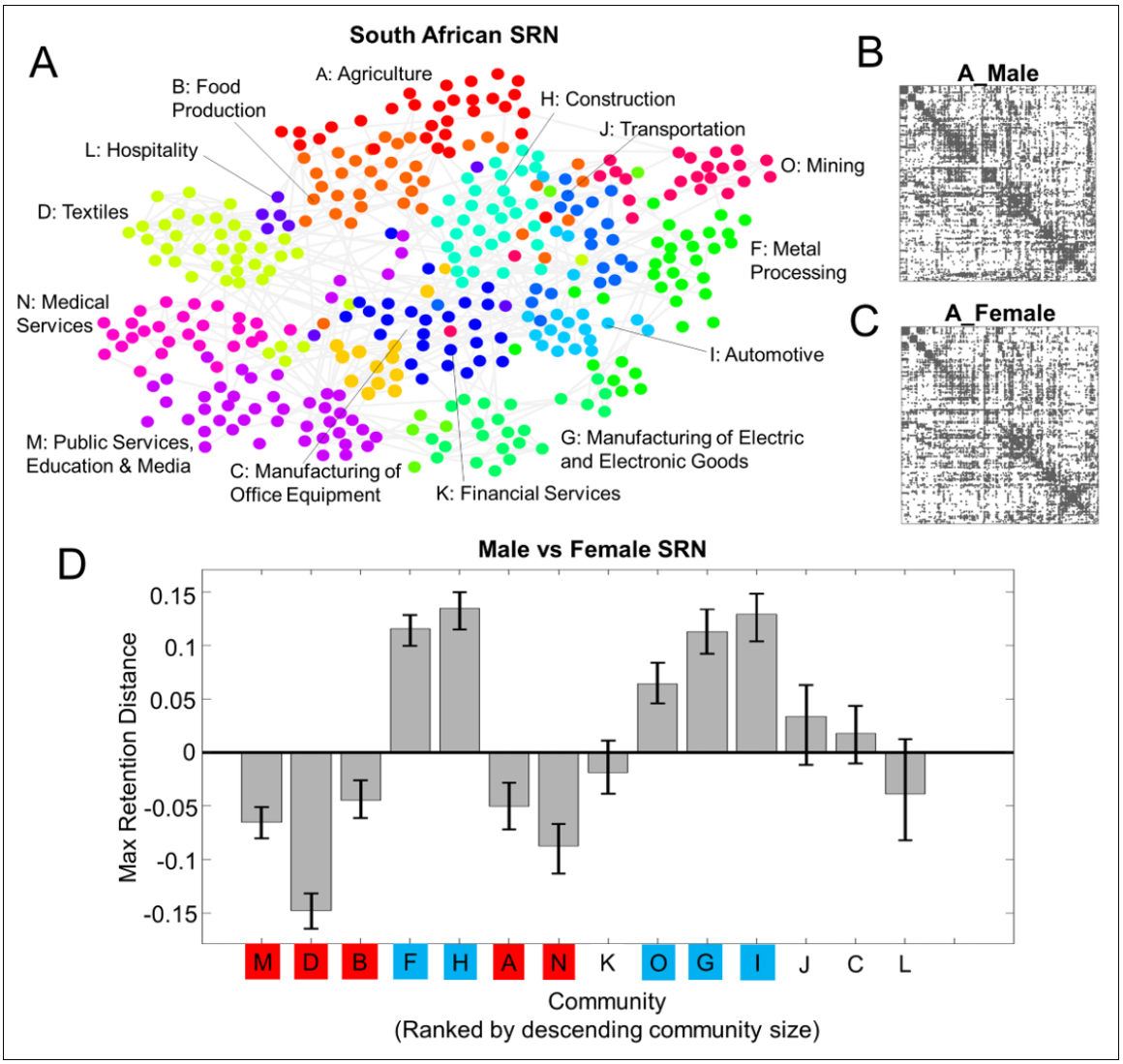


Figure 4.11: *The maximum retention distance between the modular structure of different sectors in the South African male and female SRN.* (A) A network visualisation of the South African SRN with communities (skill basins) shown via node colouring. (B) The adjacency matrices for the male and female SRNs. (D) The maximum retention distance between the male and female SRN for the different skill basins where the colouring of the community labels indicates whether the male (blue) or female (red) community is more isolated.

Overall, we observe that our distances are small and four of the communities show to have no significant difference in their modular structure. However, we do still find some interesting differences in some of the communities modular structure. Our results indicate that women in public services, medical services, agriculture, food production and textiles are more isolated from the wider labour market (or just display a stronger community structure) compared to their male co-workers. These communities are all positioned on the upper left of the network. On the other hand, male workers within the mining, metal processing, electric and electronic goods

manufacturing, automotive and construction sector are much more skill-isolated (or again display a stronger community structure) than their women colleagues. These communities are all positioned on the right of the network. It is interesting that we again observe that certain regions of the network display similar results - either more or less isolated than the opposite gender. Although being skill-isolated can allude to skill-specialisation, workers may also be at a higher risk of becoming unemployed in the case of a sector-shock. It is therefore crucial for policy-makers to understand these differences when designing industrial policy aimed at a particular skill basin or industrial sector.

4.6 Conclusion

Comparing the modular structure of a single community across networks provides insight into how a part or functional unit of a complex system compares, or in the case of temporal networks, how their structure has developed over time. Despite its utility, no current modular network comparison techniques in the literature take (or can be adapted to take) a single-community approach. Furthermore, popular local quality functions focus primarily on inner- and outer-community edge density separately or ignore the influence of edge connectivity (*i.e.* how edges are distributed) when measuring the quality of a single community in a network.

To address this problem, we propose the maximum retention distance, which compares the degree to which a community can conserve dynamics across two different node-aligned networks. This approach is advantageous as it captures both the influence of edge-density and its' connectivity simultaneously when comparing the structure of the community across networks. Secondly, as the measure does not quantify the community quality through a null-model comparison, it can compare the modular structure of a community across different networks. In addition, as the measure is local it allows for the sole comparison of a single community's modular structure without the topology of the rest of the network being considered. The local nature of the distance measure also allows for it to be used more widely. For example, in the case when only partial information of a network is available or when the network is so large that it is computationally intractable to consider the entire network topology (which is often the case for real-world networks). Finally, as the distance measure adheres to normal network distance behaviour it can easily be used and interpreted.

4.6.1 Limitations of the maximum retention distance

Although we have shown that the maximum retention distance is an important new tool to compare the modular structure of a single community across a set of node-aligned graphs, it also has some limitations which reduce its usability.

First, both communities need to be connected (the community cannot contain an isolated node or set of nodes) and at least one community node needs to be connected outwards to the rest of the network (the community is not isolated in the graph). Although this assumption reduces the usability of the metric, these assumptions do hold for most sets of nodes that display modular structure.

Second, there are certain cases where the maximum retention distance cannot capture differences in a community's modular structure across a set of networks. In these cases, the maximum retention of both graphs will be equivalent even though the community topology is different. This can occur due to symmetries between the community in the two graphs. For example, if the communities are isomorphic (a relabelling of nodes in the second graph results in the same adjacency matrix A), the resulting retention curve will be equivalent. It can also occur if the community has dynamically equivalent node classes. Here, we can identify a permutation of the node labels across the community in the two graphs which leaves the system matrix Q invariant. This can occur, for example, if all nodes have the same k_{in}/k_{out} ratios even though their exact connectivity may differ. Or if both communities can exactly be described by the same quotient graph (such as in an EEP). In these cases, our measure will not be able to differentiate between these communities. However, by applying the parametric bootstrapping to obtain a confidence interval, we mitigate this problem as the changes in the edge weightings can break some of these symmetries and show how they differ.

We have also numerically searched for cases where we compare a community across two graphs and obtain the same maximum retention but different maximum retention times. Although theoretically such cases can potentially exist, we found them to be particularly rare (and were unable to find such a case). Note, that the same maximum retention can more easily be obtained when independently comparing two different communities (of different sizes) across two different sets of networks. However, the maximum retention distance should only be directly compared when comparing the same community across a node-aligned network set.

Thrid, as our measure can obtain a distance of zero even when the underlying community structure is different across two graphs, the measure does not comply with the identity property required by mathematical distance functions. Hence, our

measure is not a formal mathematical metric. However, as we have shown, the measure does comply with normal network distance behaviour (defined by [195, 118]) despite this limitation.

Fourth, the maximum retention distance measures modular structure as a group of nodes in which flow is isolated from the rest of the network. Note, that this may differ from the way other quality functions define and measure community structure. For example, the measure does not define modular structure as a group of nodes where the dynamics acting on the community display homogeneous properties. When using this measure, this should be carefully considered. The distance is therefore best suited to comparing communities where this definition of modular structure suits the application at hand.

Furthermore, as our distance measure evaluates and compares pre-defined communities - it does not find the communities in the graphs. However, often before the measure is implemented a community detection algorithm will be used to find the community structure of one of the graphs (which we wish to compare across the network set). Care should be taken to ensure that the community detection algorithm also takes a dynamic or/and cut-based perspective and thereby similarly defines modular structure as our distance measure. This will allow communities with a strong modular structure to be compared.

Recall that the retention function (that our distance adopts) does not use a null model. This allows the measure to be able to compare a community's modular structure across different networks. However, this is also a limitation. A null model is often adopted to ensure that the signal one captures is greater than what would be expected at random. In the case of community structure, this would ensure that the community displays a modular structure (a degree of dynamical flow isolation) that is greater than what would occur randomly. As our measure does not account for this, care needs to be taken when interpreting our distance measure. To do this, we should compare the distance to the maximum retention distance that would be obtained when comparing the community to its structure in a graph with no modular structure (the graph's configuration model). Note, that this would represent the degree of modular structure that would occur randomly. This value can offer a reference to the distance - if the distance is greater than this reference point the community is just being compared to a network with no modular structure.

Similarly, the local nature of our distance measure is advantageous as it allows for the sole comparison of the community's modular structure. However, by ignoring the topology of nodes outside of the community we also lose information on how

the community influences the network more broadly. For example, the measure will not differentiate if the community lies in the core or the periphery of the network. This can be important, for example, when the community is a key bridge connecting various other communities or parts of the network. Furthermore, the measure cannot measure or compare the distances between different communities. This is because the rest of the network is collapsed into a single absorbing state. To compare these features of the community and broader network a different network distance metric will need to be adopted.

Finally, to implement the measure, the user needs to consider the full dynamics of the walker on the community in both graphs. This is because we need to first investigate the retention distance's shape and then find its maximum. Hence, this measure can become computationally expensive if large communities are considered.

4.6.2 Future Work

To address some of the current distance limitations, we propose a few avenues for future work.

To address that the measure is computationally expensive (especially as the communities increase in size), one could investigate if we can algebraically predict, or at least provide bounds, for the maximum retention value. This would allow the dynamics only to be calculated up to these bounds. As a starting point, one could consider approximating the retention function using the largest eigenvalue of each graph's system matrix (λ^t_{max}) and finding their difference. This function would approximate the retention distance. It also represents the retention distance of a randomly shuffled version of the community in each graph. Recall that this value exactly describes the expected community retention curve if the community displays ER graph structure with a constant inner- and outer-community edge probability. We could then use the maximum value of this retention distance curve as an approximation, and try to characterise its distance from the actual maximum retention distance. This could also be a measure of the influence of community connectivity (how edges are distributed within the community).

Another interesting avenue of future work could investigate if we can algebraically predict the shape of the retention distance curve (whether it has a single- or multiple-hump shape). What makes this a challenging task is that the curve is the difference between the dynamics on two different graphs. A multiple-hump-shape occurs when the speed of the dynamics on one graph changes in relation to the other graph (*i.e.*, where the speed is initially faster and then becomes slower than the other graph, or

vice versa). Here, we could start by investigating if we can predict when the retention curve is zero. If it occurs between the retention curve's starting and ending point - it could indicate a multi-hump shape.

Furthermore, our measure only considers weighted, undirected graphs. Future research could extend this framework to incorporate directed graphs. Here, one would need to investigate how to deal with edges pointing into the community. Currently, in our measure, we apply a graph transformation to a community that makes all nodes in the rest of the network an absorbing state, and all outer-community edges point to this state. It is therefore not apparent whether we would ignore edges pointing into the community or make these edges incident to a separate collapsed state that forms part of the initial random walker probability distribution. The implications of these choices on measuring the modular structure of a community would need to be investigated.

Finally, the measure assumes that the community is connected in both graphs. For the directed case, we would need to assume that the community is strongly connected. This is a much stronger assumption and may further restrict the measure's use. Here, we could extend our measure's applicability by allowing for random walker teleportation (similar to that used in the Page-Rank centrality measure [36]). The implications of this on the metric's ability to quantify and compare the community's modular structure would need to be carefully investigated.

Chapter 5

An Application: The Irish Labour Flow Network

In this chapter, we focus on the modelling and analysis of the Irish SRN to answer a key economic question: How does the presence of related multinational enterprises impact the dynamics of domestic industries through knowledge spillovers in Irish regions? We contribute methodologically by constructing a novel dynamical-based centrality measure for the SRN. This measure captures the potential for higher-order linkages and knowledge spillovers between a focal industry and other industries in the region.

5.1 Introduction

A central position in the global research agenda is to investigate the avenues through which regions can generate economic prosperity and growth. Recall that foundation theories emerging from evolutionary economic geography suggest that regions grow by combining existing capabilities to create new economic activity [150]. This is because it is costly to develop new activities that require locally unavailable capabilities. However, regions may also look outwards to access new capabilities through external actors such as suppliers in neighbouring regions, migrants or foreign direct investment (FDI). In particular, attracting multinational enterprises (MNEs) is seen as a key channel to *import* new capabilities and generate knowledge spillovers via technology [132, 11] and skill transfer [95, 15]. These spillovers are thought to enrich a region's capability base and thereby enhance domestic diversification opportunities.

MNEs are generally viewed as beneficial to a host economy as they transfer financial resources [107], create new market opportunities [52] and influence the productivity and innovation of co-located domestic firms through spillover effects [107].

Spillover effects can emerge through various channels, including demonstration effects, competition effects, and labour mobility [26]. These are most often captured empirically through supply-chain linkages, but we focus on inter-sectoral labour mobility here to better proxy for knowledge spillovers [15, 95]. However, these spillover effects may not always materialise, as MNEs actively protect their know-how and skills to prevent competition [4], or the capability gap may be too large between domestic and MNE firms and workers, limiting absorptive capacity [117, 26].

In this study, we are interested in whether cohesion to MNEs in related sectors leads to knowledge spillovers that drive new domestic industry entries at a regional level. While a huge number of studies have investigated the effect of MNEs *within* an industry [91, 100, 52], there have been fewer studies focusing on *inter-industry* impacts. These include the impact of supply-chain linkages to MNEs on domestic entry [12], and the cohesion of MNE entries to the local knowledge base [71]. Here we focus on the role of inter-industry knowledge spillovers from local MNEs for domestic industry entry.

Further, we are interested in whether MNEs have a protective effect on regional domestic industry survival. Within the evolutionary economic geography literature, resilience is studied from an evolutionary perspective. It defines the resilience as a region's ability to successfully diversify into new growth paths when faced with an economic shock [189]. The current consensus is that the more variety and the more closely an industry is related to a region's industrial basket, the more likely it is to survive [150, 147, 14]. In terms of inter-industry spillover effects from MNE to domestic industries at a regional level, Szakálné Kanó *et al.*[194] found that the greater the variety of MNEs within a region's industrial portfolio the higher the chance of firm survival and the greater the region's resilience. Here we extend this literature, focusing on the impact of cohesion to MNEs in terms of knowledge or skill linkages on domestic industry exits.

In order to quantify the potential for knowledge spillovers, encompassing a range of potential mechanisms, we deploy the skill-relatedness metric [146]. Recall that skill-relatedness is a pair-wise measure of industry skill-similarity based on inter-industry labour mobility. Typically, this relatedness measure is then used with a cohesion measure to quantify the degree of relatedness between an industry and the wider existing industrial basket of a region. There is a range of cohesion metrics in the literature, such as the closeness measure [147] or the density measure [104]. However, most of these metrics only consider local or 'nearest neighbour' links. Therefore, these measures implicitly ignore the connectivity of their related industries (or neighbours)

- both between themselves and other industries in the region. In other words, they do not consider the impact of high-order linkages that form a densely connected group of related industries in a region. We need a metric that captures these inter-related sets of industries that share skills and know-how in the greater neighbourhood of a node or industry. To do this, we propose a new dynamic-based cohesion measure, the *strategic closeness*, which captures cohesion to industries which are themselves well-connected in the greater neighbourhood, thus quantifying relatedness *between* industries in a region.

We carry out our analysis on a subset of government-supported Irish firms that covers the vast majority of manufacturing and exporting firms in Ireland. Knowledge spillovers from MNEs are most likely to occur amongst these firms as they represent the most productive and complex part of an economy [116]. Unlike previous studies, we separately investigate the impact of so-called 'overlapping industries' - those that have both MNE and domestic employment in a region - and MNE-only industries ('exclusive MNE industries') on domestic industry entry and exit. We focus on three distinct periods, before the Financial Crisis (2006-9), the recession (2010-14) and the recovery period (2015-19), which coincided with the Brexit referendum of 2016.

We find that cohesion to overlapping industries is positively associated with both entry and survival of government-supported domestic export and manufacturing industries. In contrast, we find that if a domestic industry is proximate to MNE-only industries it reduces the industry's chance of entry and survival. Our results suggest that domestic industries are unable to benefit from spillovers in this case due to a large technological and know-how gap. Hence, while it is difficult for domestic firms to 'leap' into these more complex and cognitively distant MNE-dominated industries, once they have successfully entered and coexisted with MNEs, overlapping industries appear to induce further new entries via knowledge spillovers successfully. In the most recent period studied, 2015-19, we observe domestic export firms entering MNE-exclusive industries. Although we show no causal link, this coincides with significant financial support injected in 2017-18 to enhance firm diversification and generate new domestic-MNE links in response to Brexit fears. During this period, domestic industries seemd to make larger cognitive 'leaps' and break into MNE-dominated industries. Finally, we find that the type of cohesion matters. The presence of densely connected overlapping industries is associated with more entries and better survival of domestic export and manufacturing activities in the recovery period.

We briefly provide an overview of the structure of the rest of this chapter. Following a comprehensive literature review, we introduce the data and definitions of

industry entries and exits and present some preliminary statistics and trends in the data. We then focus on the methodological development of the closeness measure of Neffke *et al.* [147] and the introduction of a new cohesion measure, the strategic closeness, before adapting both measures to account for domestic and MNE industries separately. We then present our econometric model and our results. Finally, the paper concludes by discussing some potential policy implications of our work.

5.2 Literature

5.2.1 MNEs as agents of structural change

Regional diversification is often depicted as a branching process in which a region develops new economic activities by drawing on and recombining capabilities, particularly know-how and skills embedded in workers, that are present within the region [104, 83]. This is because search costs rapidly rise as the gap between the regionally available skills and know-how and those that are required for the new economic activity widens. Furthermore, new activities unrelated to the existing knowledge base of a region tend to have a lower probability of survival [150, 147]. Hence, related diversification (diversification into industries that are cognitively similar within a region) is the dominant channel for industrial diversification, while unrelated diversification (diversification into industries that are cognitively dissimilar) is rare [84, 175]. What is less clear within this literature is the role that external actors (e.g. suppliers and customers in neighbouring regions or foreign direct investment (FDI)) play in the development paths of regions.

As the number and importance of MNEs have risen globally, many governments have developed industrial policies aimed at attracting FDI and other kinds of MNE engagement. This is because MNEs are seen as key generators of income, innovation and growth for the host economy. Examples of potential mechanisms through which MNEs bestow a beneficial effect on host countries' economies include directly via financial resources (spending on local suppliers, capital investment, employment, tax revenue), technology (R&D), know-how in terms of management and training of the workforce, as well as through linkages to value chains [107]. Furthermore, market access spillovers from MNEs to domestic firms are also important as they connect regions to global markets (and thereby induce domestic exporting activity) [54]. Our focus here, knowledge spillovers from MNEs to domestic firms, have been suggested to primarily occur along three channels: demonstration effects where domestic firms gain knowledge by imitating MNE firms, competition effects, and knowledge transfer

through labour mobility [26]. This dimension is particularly important with respect to regional industrial dynamics and the import of know-how into a local labour force.

Spillover effects to the domestic host economy may not always materialise [90, 55]. This can be attributed to MNE characteristics, which include actively protecting their know-how to reduce knowledge leakages to domestic competitors [4], or out-competing domestic firms in the labour market by providing better employment conditions to workers [3, 22]. A related branch of literature has specifically investigated how the 'absorptive capacity' of domestic firms influences spillover effects [117, 26]. The absorptive capacity of a firm is defined as a firm's ability to recognise valuable new knowledge, integrate it into the firm and use it productively [207]. Various authors have argued that the lack of spillovers from MNEs to domestic firms is due to a wide skill or technology gap between the two groups [116, 86]. Empirical studies have shown that the strength of MNE-domestic spillover effects rises as the size and productivity of domestic firms increases [25].

There have been a variety of empirical studies investigating how MNEs influence their host economy. These studies vary in the country they study, which aspect of economic development they consider (e.g., employment growth, industrial diversification, or firm productivity and innovation), and how they define MNE presence (FDI, value added to GDP, or measures related to R&D expenditure, sectoral output, foreign equity, sales, employment etc.). As a result of diverging research findings, perhaps owing to the heterogeneous research designs employed, there is currently no consensus within the literature on the impact of MNEs on a host region.

We focus here specifically on inter-industry spillovers. Within this burgeoning literature, we highlight a few studies of particular relevance which focus on domestic industry entry and cohesion to MNEs as proxied by inter-industry linkages. Görg and Strobl [92] and Ayyagari and Kosová [12] find that the presence of related MNEs (supply-chain linkages) is positively associated with domestic entries in the manufacturing and service sectors respectively. Lo Turco and Maggioni [129] looked at product entries, finding that cohesion to MNEs enhances entry, particularly for more productive, established and local selling firms. Finally, some studies [37, 25] focus on firm productivity, and similarly found that cohesion to MNEs via supply chain linkages promotes productivity particularly for larger firms and those focused on the domestic market.

We choose Ireland due to the country's profile as a highly developed open economy [164] with substantial MNE presence and an FDI-oriented industrial policy. In fact, Ireland is one of the world's most active countries in terms of industrial policy. Key

objectives include promoting export-led growth [34], generating industry-university R&D partnerships [20], and developing better linkages between industries [19, 164]. Some of the strategies developed to foster linkages and knowledge spillovers between domestic and MNE activities include the development of collaborative R&D infrastructure and clusters [62], which have been linked to the emergence of industrial clusters [162].

Various authors have investigated the role of MNEs on the Irish economy; however, the impact of MNEs on domestic activities remains unclear. Studies have found that the presence of MNEs within the same sector influences the entry rate, productivity, and employment growth of domestic firms [94]. Specifically, for productivity and employment growth, benefits have only been observed in high-tech domestic sectors [93]. In contrast, authors have found a negative effect of MNEs on domestic exporting firms' wages and productivity [21]. Most related to our work, Görg *et al.* [91] suggested that related MNEs support domestic industry entrance through supply chain linkages, but a negative (or non-existent for R&D-active firms) impact of MNE supply chain linkages on domestic firm productivity has also been found [65].

Our study differs in five key aspects from previous work. First, we focus on the cohesion between new domestic export sectors and existing MNE activities as captured by inter-industry labour mobility patterns, a proxy for a broad range of potential knowledge spillovers. Secondly, we decompose MNE industries into two sets - overlapping industries (those with also domestic presence) and exclusive MNE industries. This enables us to distinguish the role of these distinct sets, particularly with respect to their capability gap to domestic firms. Thirdly, we investigate which type of cohesion to MNEs matters. Specifically, as we outline below, we develop a new cohesion to capture the complex structure of linkages present between industries in a region. Fourthly, we conduct our analysis at a granular industry-region level, unlike the majority of studies which focus on industries at the national level or regions neglecting the industry dimension. Finally, we investigate the role of MNEs in three different economic eras: before and after the 2008 financial crisis and during a recent period (2015-19) characterised by rapid domestic growth and Brexit.

5.2.2 Regional resilience

Regional resilience has become a prominent focus area in the research and policy agenda, featuring in the target indicators of the Sustainable Development Goals. Indeed, Target 1.5 aims to *'By 2030 build the resilience of the poor and those in vulnerable situations, and reduce their exposure and vulnerability to climate-related*

extreme events and other economic, social and environmental shocks and disasters' [13]. A growing literature investigates the determinants of a region's ability to adapt to an external shock such as the Financial Crisis [53, 82, 203]. However, despite the growing popularity of resilience in the research and policy agenda, there are concerns over the usefulness of the concept stemming from a lack of clarity on its definition [134]; the appropriate theoretical and empirical frameworks to measure and analyse it [68]; its determinants [134]; and tools to design or implement appropriate policies.

While the concept of resilience is multidimensional and has been interpreted in various ways, there are three main conceptual approaches: engineering, ecological, and adaptive [189, 136]. Scholars advocating an *engineering-based approach* emphasise the ability of an economic system to return to its stable or pre-crisis equilibrium state after a crisis. The *ecological-based view* concerns the magnitude of a shock that a system can weather without shifting to a new equilibrium state. The *evolutionary approach* departs from these equilibrium-based frameworks and defines resilience as the ability of an economy to diversify and branch out into new growth paths successfully, thereby countering economic decline [135, 29]. In the latter approach, which we adopt in this study, authors typically study regional resilience by investigating industry exits or survival after an economic shock [68].

A range of studies have investigated the relationship between the industrial portfolio of a region and industry exits [147, 76, 194]. Studies have found that a region with a wide variety of industries can better adapt to sector-specific shocks [29, 194]. Furthermore, a study showed that regions with a high degree of relatedness to existing technologies in which the region does have a comparative advantage (competitive presence) had a greater capacity to weather technological crisis [14]. Overall, there is evidence in the literature to support both industry variety and relatedness as key factors in industrial resilience.

Authors have also argued that MNE-domestic linkages can act as a protective effect on domestic industries in times of crisis and enhance the resilience of the host economy. This can occur via several mechanisms. Primarily, knowledge and experience built up by MNEs on external shocks may transfer to domestic firms [78] via demonstration effects and labour mobility. Labour mobility may also assist via reallocation of workers between sectors in a region [68]. Furthermore, productivity gained from technological and knowledge spillovers is expected to reduce the average cost of domestic production, which may help firms survive in the face of economic shocks [93].

Very few empirical studies have investigated the role of MNEs in regional resilience. Closest to our work, Szakálné Kanó *et al.* [194] showed that a greater variety of MNE industries reduces the likelihood of domestic firm exit, using data on Hungarian regions. This effect was particularly strong for regions undergoing economic transition. Focusing on Ireland, the authors found that the presence of MNEs in the same industry increases the domestic firm's chance of survival through technological spillovers [93]. This was only significant for high-tech sectors, while no effect of MNE presence was found for low-tech sectors. We add to this literature by investigating whether cohesion to MNE industries, measured at an industry-region level, provides a protective effect.

5.2.3 Industrial cohesion

In this study, we aim to investigate how the presence of MNE and domestic industries within a region's industrial basket impacts an industry's entry or exit. We are therefore interested in measuring the capability or knowledge-distance between an industry and a region's current industry basket. Within the literature, cohesion is defined as the degree of relatedness amongst industries within a region; a measure of the opportunity for knowledge spillovers [147, 84]. Cohesion measures are typically used to quantify the degree of structural change induced by an industry as it enters, or leaves, a region. When an industry enters (or leaves) a region, it brings new (or removes current) capabilities. Hence, according to which industries enter or exit, and how they are related to the current industrial portfolio, they differently impact the region through changes in the combined total of the region's capabilities.

Cohesion measures are typically derived from the structure of an industry network [149, 102]. This is a network where nodes represent industries, and edge-weights correspond to the degree of capability-overlap between industry pairs. The advantage of using an industry network is that it allows for the topological structure resulting from the relatedness amongst all industry pairs to be analysed via a complex network approach. A measure of relatedness amongst industries is required to construct an industry network. Since the actual level of capability overlap cannot be directly measured, an outcomes-based approach is taken to infer the degree of relatedness. This type of approach varies according to the data source considered and capability type. For example, the co-location of industry pairs on patents has been used to measure the degree of technological-relatedness [110] and supply chain (IO) linkages have been used to estimate the degree of supplier-buyer sharing or similarity between industries [1]. In this study, we adopt the skill-relatedness index, based on labour

mobility between industries [146], as we primarily focus on knowledge spillovers and labour pooling between MNEs and domestic firms.

Neffke *et al.* [147] introduced one of the first cohesion measures. The authors defined the *closeness* of an industry as the count of the number of related industries (neighbours in the industry network) present within a region's industrial portfolio. This effectively captures how connected or embedded an industry is to other local industries. More formally, given that industry j's presence in region r at time t is defined by $P(j, r, t)$, where $P(j, r, t) = 1$ if the industry is present and $P(j, r, t) = 0$ if not. The skill-relatedness between industry i and j is also given as $A_{SR}(i, j)$. The closeness of industry i to the region's industrial basket is then defined as:

$$\text{Closeness}(i, r, t) = \sum_j P(j, r, t) \times (A_{SR}(i, j) > 0.25)$$

Note that the measure considers an industry cohesive if it has many neighbours (regardless of their skill-relatedness). This measure is, therefore, particularly useful in a study where the presence of related industries rather than their degree of relatedness is most important. The authors found that for manufacturing industries in Sweden, industries that enter a region have a higher closeness, while those that exit have a lower closeness.

Another well-known cohesion measure is the *density* [102] (or related employment[1]) measure. This metric measures the strength of the relatedness (edge weight) between an industry and its neighbours relative to the strength of relatedness to all industries. More formally, density is defined as:

$$\text{Density}(i, r, t) = \sum_j \frac{A_{SR}(i, j)}{\sum_j A_{SR}(i, j)} \times P(j, r, t)$$

Note that this measure considers an industry cohesive if a high fraction of its neighbours (weighed by their relatedness) are present within the region. This measure has been used in various applications to predict regional industry diversification and employment growth [147, 33, 102]. In all of these studies, industries with a higher density are more likely to enter a region.

Both measures mentioned above are one-step measures as they only consider direct neighbours in the network. Consequently, all neighbours are treated homogeneously. These measures fail to capture the importance of the connectivity and embeddedness of their neighbours (the greater industry network structure). Various authors have

[1]Related employment is very similar to the density measure but also considers the employment size of each industry.

argued that it is not only the presence of a related industry but also the assemblage of these industries, as well as other industries present in the region, that generate collective efficiency and further knowledge spillovers [133]. Hence, spillovers are generated from the presence of a cluster of densely connected economic activity around an industry [180]. Therefore, being more deeply embedded into the industry network enhances the chance of spillovers. Several influential studies have also argued that economic activity more distant from an industry can enhance innovation. However, the economic activities cannot be too cognitively distant that learning cannot occur [157, 84]. In this study, we develop a new cohesion measure that captures the presence of related industries and their connectivity to other industries within a region. The measure, therefore, captures the impact of higher-order linkages that may occur through the broader concentration and inter-connectivity of economic activity but is not too skill-distant from an industry.

Another cohesiveness measure, *related variety*, has also been widely adopted within the literature [84, 194]. A region with a wide related variety has employment spread over various industries within only a few sectors. In contrast, one with a low related variety has employment spread over industries within different sectors. The authors hypothesised that regions benefit from employment distributed in various industries as more variety implies more potential for spillovers. However, the variety should primarily occur amongst industries in the same sector as limited spillovers occur amongst industries in different sectors. The measure is directly computed at a regional level and based on an entropy calculation for employment distributed within industries spread across sectors.

A central disadvantage of the related variety measure is that it relies on the hierarchical structure of the standard industrial classification system as a measure of relatedness. Thereby, when considering the related variety of a region, all industries within the same sector (2-digit industry class) are assumed to have the same relatedness. Secondly and most importantly for our study, the measure is calculated at a regional level making it less suited to our application. Our new metric, however, does capture some of the ideas behind related variety in that it identifies the presence of groups of related industries - although, in our case, we quantify relatedness via the skill-relatedness measure rather than the official industrial classification - in a region. The key difference, however, is that it enables us to look at the cohesion of a particular industry to these groups, resulting in an industry-region level variable.

5.3 Data and definitions

5.3.1 Industry data

For this study, we use data covering the majority of exporting and manufacturing firms within the Irish economy. The data derives from the Irish Department of Business, Enterprise, and Innovation Annual Business Survey of Economic Impact, and includes firms assisted by the three Irish enterprise development agencies: Industrial Development Authority (IDA) Ireland, Enterprise Ireland, and Údarás na Gaeltachta.

The dataset covers the period 2006-2019 and includes total employment (in assisted firms) at an industry-region-year level of aggregation. This is further broken down by firm ownership-type level (Irish or foreign). Industries correspond to 4-digit NACE 2 industry level, and regions are at NUTS level 3. Further data descriptors across region, time and ownership type are presented in Table C.1 in Appendix C.1.

The dataset includes approximately 80% of all manufacturing employment and 7% of services employment within Ireland. Furthermore, it accounts for 90% of total merchandise exports as well as 70% of services exports (which comprise of approximately half of total Irish exports) [34]. It also includes approximately 63% of the employment in all foreign-owned firms in Ireland. The dataset, therefore, covers both the majority of domestic and foreign manufacturing and exporting firms. As these firms are typically highly productive, complex and export-focused, there is a higher likelihood of MNE-domestic knowledge spillovers occurring amongst them [116, 25]. These firms also act as leading drivers of economic development, thus also offering an important indicator of regional economic prosperity.

5.3.2 Industry presences, entries and exits

We start by dividing our 14-year time-span into three time periods, namely 2006-2009, 2010-2014 and 2015-2019. Each of these periods can be associated with a distinct economic era in Ireland. The 2006-2009 period falls largely before the 2008 financial crisis started to take effect. A recession then characterised the 2010-2014 time period [49]. Finally, the 2015-2019 period was characterised by fast growth of the domestic economy and Irish industrial policy support for both domestic and MNE firms in response to Brexit [182].

In this study, we investigate the entry and exit of domestic export and manufacturing industries with respect to their cohesion to three mutually exclusive sets of existing industries within a region. These sets are 'exclusive MNE' industries in

which only MNE firms are active, 'exclusive domestic' industries in which only domestic firms are active and 'overlapping' industries in which both MNE and domestic industries are active.

We define the presence (denoted as $P(j,r,t)$), entry and exit of an industry j in region r at time t as follows:

- An MNE industry is *present* if there are more than 5 employees in foreign-owned firms ($P_M(j,r,t) = 1$ and otherwise 0), while a domestic industry is present if there are more than 5 employees in Irish-owned firms ($P_D(j,r,t) = 1$ and otherwise 0).

- A domestic industry *entrant* (Entry$(i,j,t+1)$) is an industry that had less than 5 employees in the beginning of the time period, and then becomes present in the industrial portfolio of a region at the end of the time period ($P_D(j,r,t) = 0 \cap P_D(j,r,t+1) = 1$).

- A domestic industry *exit* (Exit$(i,j,t+1)$) is an industry that had more than 5 employees in the beginning of the time period and then was no longer present in the industrial portfolio of the region at the end of the time period ($P_D(j,r,t) = 1 \cap P_D(j,r,t+1) = 0$).

Similarly,

- An exclusive MNE industry is an industry in which only MNE firms are active, $P_{exclM}(j,r,t) = 1$ if $P_M(j,r,t) = 1 \cap P_D(j,r,t+1) = 0$.

- An exclusive domestic industry is an industry in which only domestic firms are active, $P_{exclD}(j,r,t) = 1$ if $P_D(j,r,t) = 1 \cap P_M(j,r,t+1) = 0$.

- An overlapping industry is an industry in which both MNE and domestic firms are active, $P_{overlap}(j,r,t) = 1$ if $(P_M(j,r,t) = 1 \cap P_D(j,r,t+1) = 1)$.

We choose 5 employees as our threshold measure to indicate the presence of an industry within a region [145], and hence less well established and potentially dormant industries are removed [2].

As a preliminary step, we first investigate the magnitude of domestic regional structural change following the approach of Neffke *et al.* [147]. Figure 5.1 shows the dynamics of domestic industries over the entire period in our study. Regarding all industry-region combinations within our dataset, only 87% present within 2006

[2]We do check our results with other thresholds in the range of $(2,8)$ and still find robust results.

still exists in 2019. Taking the reverse perspective, 85% of domestic industries in 2019 already existed in 2006. For comparison, Neffke *et al.* [147] found that 78% of industries in 1998 in Sweden were still present in 2002 and 68% of industries in 2002 were still present in 1998. Slightly lower churn levels are not unexpected in our case as we consider just a subset of Irish firms (*i.e.*, those supported by government agencies).

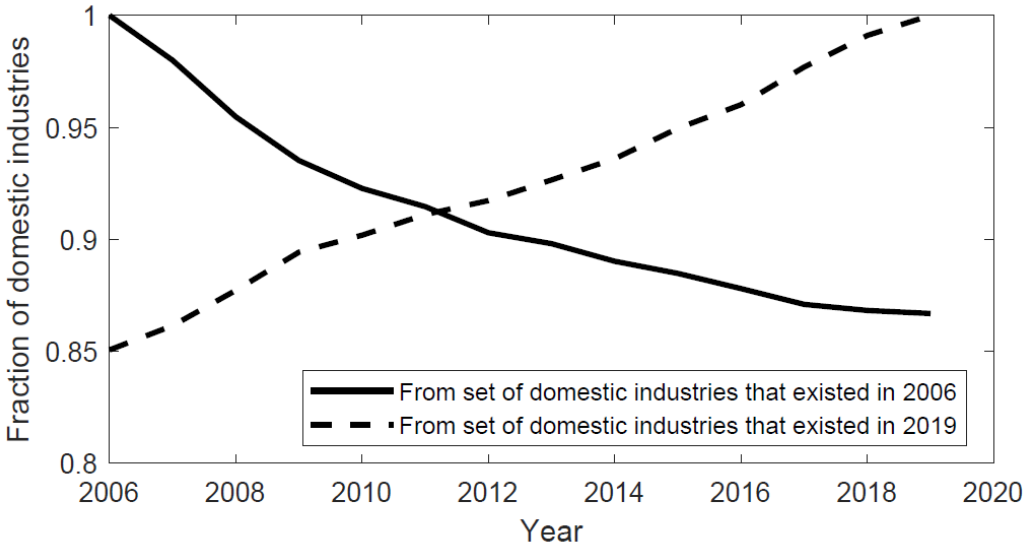


Figure 5.1: Graph showing the domestic structural change in Irish regions between 2006–2019. The solid line shows, for all regions in Ireland, the share of domestic industries that belong to the original set of domestic industries in 2006 as a percentage of the total amount of domestic industries in each consecutive year. The dotted line shows the share of domestic industries in each preceding year that still existed in 2019.

In Figure 5.2, we illustrate the number of new domestic industries entering into exclusive MNE industries each year. We observe a sharp increase in entries within the 2017 and 2018 periods. Note that, in 2017, due to high levels of uncertainty and fear of loss of UK markets by Irish exporting firms, the Irish government made a large amount of capital available to Irish firms, particularly small and medium enterprises, supported by government agencies [63]. This investment and a range of corresponding policies aimed to both provide adequate support to Irish exporting firms [35], and enhance the diversification of export markets and promote domestic entry into existing markets [72, 98]. This is a potential hypothesis for the sharp increase we observe.

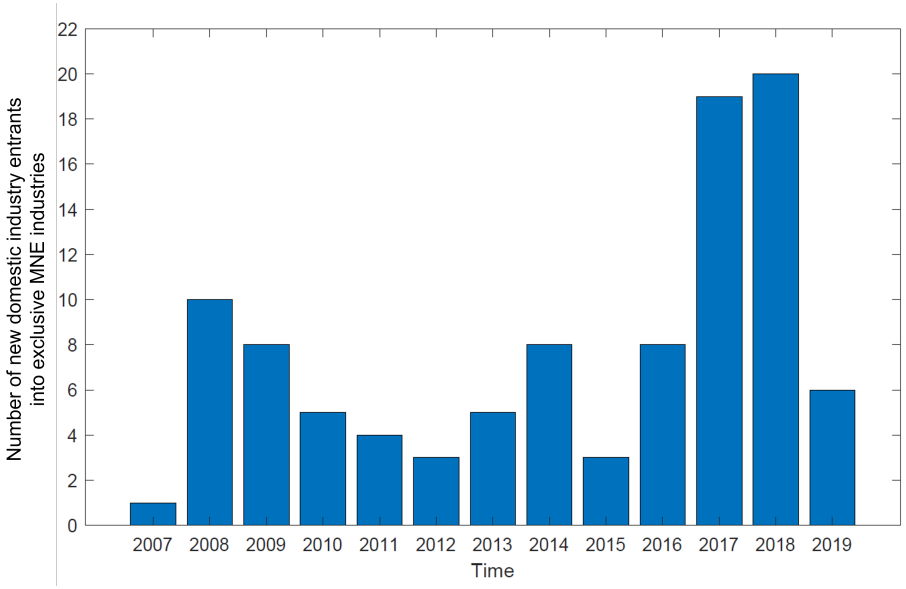


Figure 5.2: The number of new domestic industry entries into industries in which only MNEs are active in all regions in Ireland within the 2006-2019 period.

5.3.3 Skill-relatedness matrix

We use a second dataset to measure the skill-relatedness (A_{SR}) between industry pairs, following [149]. Recall that the skill-relatedness value is calculated by considering the degree of labour mobility between industries. The reader is referred to §2.5 for a more in depth discussion of the construction of skill-relatedness measure.

O'Clery *et al.* [161] previously constructed the skill-relatedness measure for Irish industries using an anonymised administrative dataset from Ireland's Central Statistics Office. The dataset contains the employment records[3] of each registered employee within the Irish formal economy. The dataset covers the 2005-2016 period[4]. Hence, entry $A_{SR}(i, j)$ corresponds to the average number of workers that transitioned between industry i and j between 2005-2016 normalised by the number that would have been expected to switch at random given the total worker flows of the corresponding industries. Recall, that the matrix is normalised and all values lie in range of $[0, 1]$.

Within this dataset industries are classified using the 4-digit NACE 1.1 industry classification. Note, that this is a slightly different industry classification compared to our export firm employment dataset, which define industires according to the 4-digit NACE 2 industrial classification. We therefore need to match the two datasets so that

[3]The employment records are constructed from SPP35 annual tax returns filed by employers on their employees to the Irish Revenue Commissioners.

[4]Note that there is an overlap in the time period to the above mentioned dataset. As the datasets differ and the skill-relatedness value has been shown to remain relatively constant across smaller time periods [149], we do not consider this to be a problem.

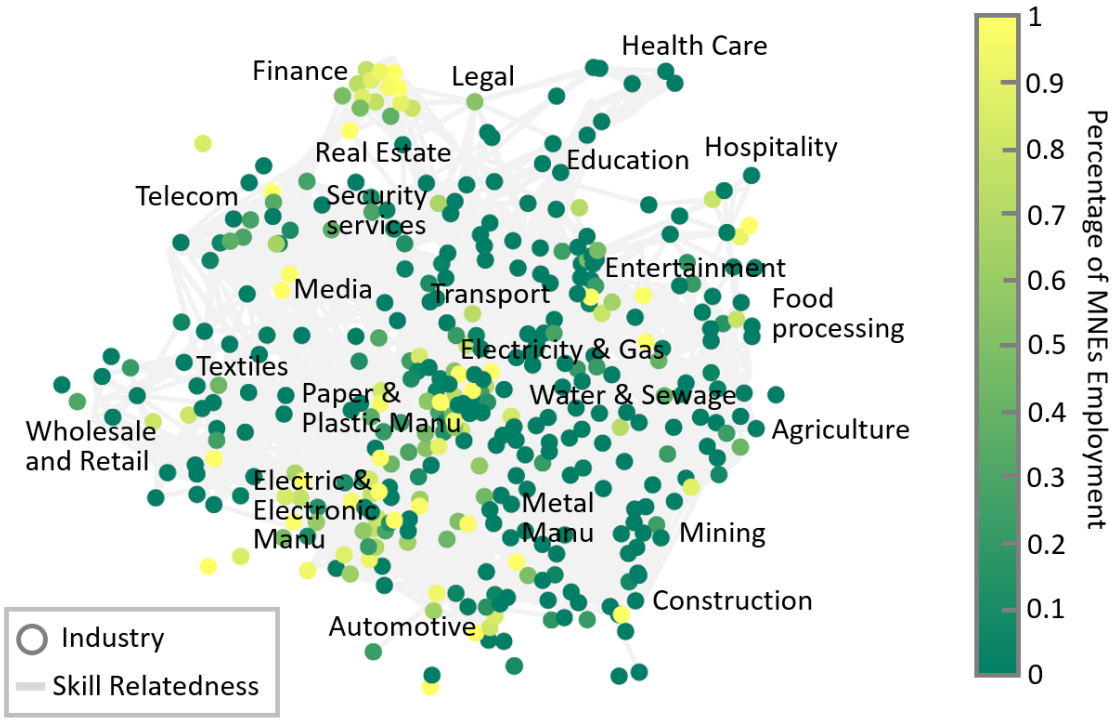


Figure 5.3: A visualization of percentage of MNE employment in each industry on the skill-relatedness network for Ireland. Each node represents an industry and each edge a skill relatedness linkage. Nodes are coloured according to the percentage of MNE employment within each industry. The network layout was generated using 'Force Atlas' in Gephi - a spring algorithm in which related industries are positioned closer together.

industries are classified by the same industry codes. To ensure industries are similarly defined we convert the second dataset into the NACE 2 industry classification. We do this by following the methodology of Diodato [66]. We use the correspondence tables between NACE 2 and NACE 1.1 released by the European Commission, as well as there detailed documentation of how these classifications differ [77].

We visualize the skill-relatedness matrix as a network shown in Figure 5.3. Each node represents an industry and each edge the skill-relatedness between the corresponding industries. A spring algorithm called 'Force Atlas' in Gephi is used to generate the spatial layout of nodes with more related industries positioned closer together. We have added a general labelling of groups of industries on the network for orientation, as well as coloured the nodes by the percentage of MNE employment. We observe that MNEs (yellow nodes) are concentrated mainly within the finance and high-tech manufacturing sectors. We highlight that the only difference between this network and the one presented previously in the thesis is that nodes now represent industries classified by the NACE 2 industrial classification.

5.4 Measuring the cohesiveness of an industry

Here we introduce the *weighted closeness* and *strategic closeness* of an industry to the existing industrial basket of a region.

5.4.1 Weighted closeness

The *weighted closeness* (WC) of an industry is similar to the closeness measure of Neffke *et al.* [145] in that it captures the number of related industries (to that industry) that are present in a region. This measure quantifies the cohesion of an industry to the industrial basket of a region as the number of related industries present in the region weighed by their relatedness.

Let the relatedness between industries be encoded in matrix A_{SR}[5]. The WC of industry i in region r at time t is then given by,

$$WC(i,r,t) = \sum_{j \neq i} A_{SR}(i,j) \times P(j,r,t), \tag{5.1}$$

where $P(j,r,t) = 1$ if industry j is present in region r at time t (and otherwise 0).

Similar to closeness [145], weighted closeness only considers the presence of directly related industries, and does not take into account the wider 'global' structure of the industry network. Furthermore, although we weigh the presence of each directly related neighbour by its relatedness, the presence of each neighbour is treated homogeneously. Hence, the measure does not consider the connectivity of these neighbouring industries to other related industries within the network. Next, we introduce a new measure that is able to capture these higher-order connections.

5.4.2 Strategic closeness

We propose a new measure, *strategic closeness* (SC), which does not only consider directly related industries (as in the case of the WC) but also their connectivity to both each other and other industries present within the region. In other words, the measure picks up the presence of higher order connections (of two steps away) in the local industrial basket. An industry with high SC is not only related to industries in the region but these industries are themselves highly connected to both each other and other industries in the region. These are in a sense 'strategic' or highly embedded

[5]While we use the skill-relatedness measure within our analysis, the cohesion measures can be used with any type of relatedness measure.

neighbours. These more distant industries increase the variety of skills and know-how an industry has access to, and are thought to promote innovation [84, 157].

The *SC* adopts a dynamic perspective by modelling a 2-step diffusive process on the network. This perspective coincides with the common approach of modelling regional diversification as a diffusive process [83]. Intuitively, it can best be understood by considering a random walker on the industry network. The random walker is initially positioned on the network with a uniform probability distribution across all industries present within the region's portfolio and is then allowed to move on the network. The walker jumps from one industry to another with probability proportional to the edge weights connecting them. After the first jump, only if the industry that the walker is now positioned at is present within the industrial basket of the region is the walker able to make a second jump. If the industry is not present the walker is removed. After the second jump, the probability distribution of the walker (across all nodes) corresponds to the cohesion measure that takes into account both the presence and inter-connectivity of 'neighbours of neighbours' in the network.

More formally, the industry network is defined via adjacency matrix A_{SR} with entries corresponding to the skill-relatedness between industry pairs. Recall, that the strength vector is denoted as $\mathbf{k}$, where $k(i)$ calculates the sum of all edge-weights that are connected to node i. This is given as $k(i) = \sum_j A_{SR}(i,j)$. Furthermore, the diagonal matrix of strengths is defined as $D = \text{diag}(\mathbf{k})$. We now define a unbiased random walker process on the industry network, similar to as described in §2.2. Here, the probability of a walker leaving a node is split amongst the edges of a node according to their relative weight. The transition probability for an edge connecting industry i and j is given by $A_{SR}(i,j)/k(i)$. The random walker dynamics on the skill-relatedness networks is modelled by:

$$\mathbf{x}_{\tau+1} = \mathbf{x}_\tau D^{-1} A_{SR}, \tag{5.2}$$

where $\mathbf{x}_\tau \in \mathbb{R}^N$ a probability vector representing the probability of finding a random walker at node i at time step τ. Recall that given an initial probability vector $\mathbf{x}_0$, the process can also be described as

$$\mathbf{x}_\tau = \mathbf{x}_0 (D^{-1} A_{SR})^\tau, \tag{5.3}$$

Now, we define the starting probability of a random walker for region r and at the base period (t_{base}) as the uniform distribution across all industries that are present in the region, this is given as:

$$\mathbf{x}_0(i,r) = \frac{P(i,r,t_{base})}{\sum_j P(j,r,t_{base})}. \tag{5.4}$$

Using Eq (5.3) and Eq (5.4), we now model the first step of the our random walker process as:

$$\mathbf{x}_1(:, r) = \mathbf{x}_0(:, r)\,(D^{-1}A_{SR}), \tag{5.5}$$

where : denotes all elements of the vector.

We then remove all the walkers that are positioned on industries that are not present within the industrial basket of the region. We do this by multiplying the probability vector of the random walker after one step with the binary vector, denoted $\mathbf{P}(:, r, t_{base})$, indicating which industries are present in the region as defined in §5.3.2. We then allow the walker to take a second step. This is given as:

$$\tilde{\mathbf{x}}_2(:, r) = (\mathbf{x}_1(:, r) \times \mathbf{P}(:, r, t_{base})) \times (D^{-1}A_{SR}). \tag{5.6}$$

Finally, we define the strategic closeness of industry i as

$$SC(i, r, t_{base}) = \tilde{x}_2(i, r). \tag{5.7}$$

Note that when calculating the $SC(i, r, t)$ we need to ensure that $P(i, r, t) = 0$. This is to ensure that we do not consider the impact of 2-step walks starting and ending on the focal node. Although this is always true when considering the SC of industries entering an industry, it is not when measuring the SC of an industry exit. Therefore, in this case, we redefine $P(i, r, t) = 0$ before calculating the SC.

Our measure only consider a 2-step dynamic process. This is to ensure that we pick up higher-order linkages that occur from the connectivity of neighbouring industries, as well as other industries that are more distant. However, as argued in the literature, these industries can't be too cognitively far away (more than 2 steps away), as otherwise learning cannot take place [157].

We illustrate the additional information gained from using the SC measure alongside the WC measure in Figure 5.4. Here we show the industrial portfolio of a mock region displayed on an industry network. For this example, an unweighted industry network is used. Nodes in blue represent industries that are present within the region, while those in grey are absent. We observe that both node A and node B are directly related to four other industries that are present within the region and hence both of these industries have the same WC. However, we can easily see that industry B is directly connected to industries that are themselves both inter-connected and connected to other industries present in the region. Industry B is therefore connected to a larger agglomeration of related industries which could provide access to a larger range of capabilities and opportunities to develop a variety of linkages. Hence, industry B has a higher SC cohesion value than industry A (which has an SC of zero as

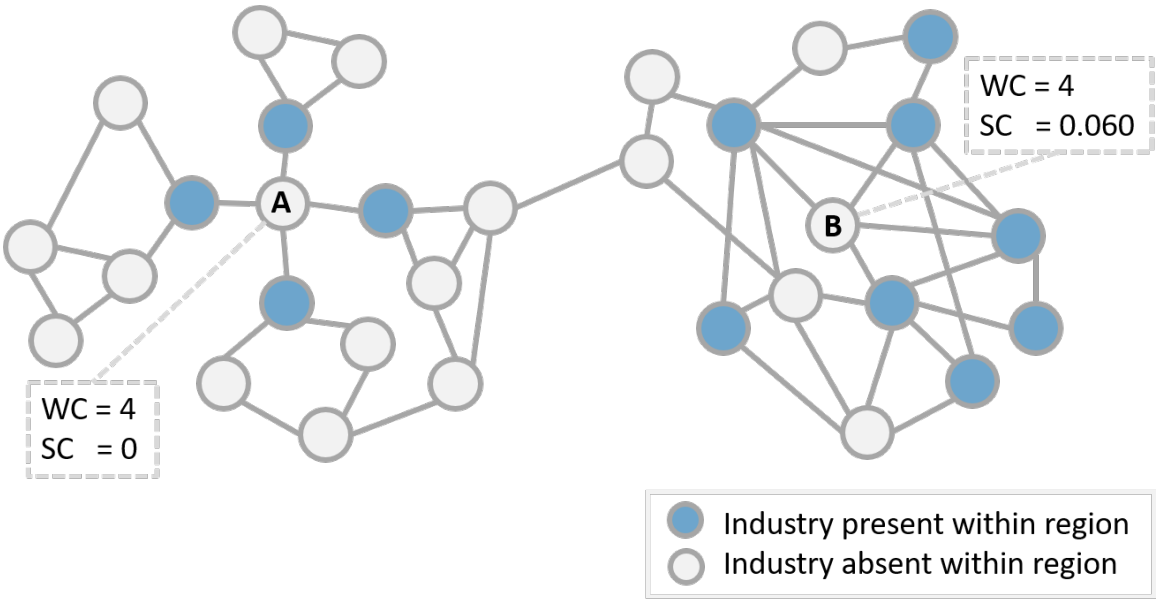


Figure 5.4: The comparison of two industries' (A and B) cohesion to the industrial portfolio of their region using the weighted closeness and strategic closeness metrics. The network represents a toy industrial network on which the mock region's industrial portfolio is shown. Each node represents an industry and each edge the level of relatedness between the corresponding two industries. Blue nodes are industries that are present within the region, while grey ones are absent.

its related industries are not connected to each other or other industries present in the region).

5.4.3 Cohesion to domestic and MNE industries

Here we adapt WC and SC to account for the presence of exclusive domestic industries, exclusive MNE industries and overlapping industries separately.

First, we adapt the weighted closeness measure to include only exclusive domestic industries within the industry portfolio of a region. This is given by,

$$WC_{exclD}(i,r,t) = \sum_{j \in N_i, j \neq i} A_{SR}(i,j) \times P_{exclD}(j,r,t). \tag{5.8}$$

where $P_{exclD}(j,r,t) = 1$ if industry j contains only domestic employment in region r at time t. This measure captures the cohesiveness of industry i to exclusive domestic industries in region r. The measure is analogously defined with respect to the presence of exclusive MNE industries (denoted as $WC_{exclM}(i,r,t)$) and the presence of overlapping industries (denoted as $WC_{overlap}(i,r,t)$) within a region.

We similarly adapt the strategic closeness measure to capture the presence of different types of industries in a region. In the case of exclusive domestic industries,

the starting probability of the random walker on industry i within region r at base time t is defined as

$$x0_{exclD}(i, r, t_{base}) = \frac{P_{exclD}(i, r, t_{base})}{\sum_j P_{exclD}(j, r, t_{base})},$$

again ensuring (otherwise redefining) $P_{exclD}(i, r, t) = 0$. Then

$$\mathbf{SC}_{exclD}(:, r, t) = \mathbf{x0}_{exclD}(:, r, t)\,(D^{-1}A_{SR}) \times \mathbf{P}_{exclD}(i, r, t) \times (D^{-1}A_{SR}). \tag{5.9}$$

where $\mathbf{x0_{exclD}}$, D, A_{SR} and $\mathbf{P}_{exclD}$ are defined as before. The strategic closeness of industry i is then given as $SC_{exclD}(i, r, t)$. The measure is similarly defined for the strategic closeness to exclusive MNE industries (and denoted $SC_{exclM}(i, r, t)$) and for the strategic closeness to overlapping industries (denoted as $SC_{overlap}(i, r, t)$) within a region.

5.4.4 Correlation analysis

In Table 5.1 we show the descriptors of our dependent and explanatory variables for all time periods. We also show the pairwise correlation between the various cohesion measures and the domestic industry entry and exit variables in Table 5.2. We see a positive but small correlation between the entry of domestic industries and the various cohesion measures. In accordance with the literature, this suggests that the more cohesive an industry the higher the likelihood of its entrance. On the other hand we see a negative relationship between the exit of domestic industries and the various cohesion measures (except for the cohesion to exclusive MNEs). Once again, this agrees with the dominant view in the literature and suggests that the less cohesive an industry the higher its chance of exit. In contrast, we find a positive relationship between exits and the WC and SC to exclusive MNEs showing that the less cohesive an industry to these types of industries the higher its chance of survival. We now further investigate these relationships controlling for various effects using econometric models.

5.5 Econometric framework

We aim to investigate the relationship between domestic industry entry (and exit) and the cohesiveness of the industry to exclusive MNE, exclusive domestic or overlapping industries in the region across three distinct periods. To detect these relationships we set up a panel probit regression model in a similar frame to Neffke *et al.* [147] and

Szakálné Kanó *et al.* [194]. We adopt a probit model as our dependent variable is binary. We highlight that although the coefficients of these models are not as easily interpreted as in an OLS model, we can use the signs and significance levels of the parameters to investigate the influence of the various variables.

For our first model, we investigate the relationship between domestic industry entrants and their cohesion to the different types of industries within the industry portfolio of the region. We run a fixed effects panel probit model for each of the three time periods separately. The model is given by:

$$\begin{aligned}\Pr(\text{Entry}(i,r,t)=1|Z(i,r,t-1),\omega P_M(i,r,t-1),\gamma(i),\tau(r)) \\ = \phi(\alpha+\beta Z(i,r,t-1)+\omega P_M(i,r,t-1)+\gamma(i)+\tau(r)+\epsilon(i,r,t)),\end{aligned} \quad (5.10)$$

where $\phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and $Z(i,r,(t-1))$ is the value of the explanatory variable (cohesiveness measure) included in the model. This can be the cohesion to exclusive MNEs (WC_{exclM}, SC_{exclM}), exclusive domestic industries (WC_{exclD}, SC_{exclD}) or overlapping industries ($WC_{overlap}$, $SC_{overlapping}$). β is then coefficient of the cohesion explanatory variable. P_M indicates whether an MNE industry is already present, with a corresponding coefficient vector ω. $\gamma(i)$ and $\tau(r)$ are industry and region fixed effects, respectively. Through these fixed effects, we control for within-region and within-industry variance. Regional fixed effects account for the number of MNE or domestic industries in a region. Similarly, industry fixed effects account for the size of each industry. This ensures that our results are driven by the absence or presence of MNEs rather than region or industry-specific dependencies. We also include a coefficient term (α) to absorb other dependencies that we have not controlled for. Furthermore, we include a robust standard error term (ϵ), account for heteroskedasticity in the model's unexplained variation and ensure we have unbiased standard errors. In this model, we only consider domestic industries not yet present within a region as an observation.

Table 5.1: Panel descriptors of dependent and explanatory variables for all time periods

Variable	N	Mean	SD	Min	Max
$Entry_D$	8312	0.0348	0.1616	0	1
$Exit_D$	2752	0.0296	0.1694	0	1
P_M	11064	0.0273	0.1630	0	1
WC_{exclD}	11064	3.4312	3.9117	0	29.5297
WC_{exclM}	11064	0.5004	0.9435	0	9.9874
$WC_{overlap}$	11064	1.3173	2.2900	0	21.2523
SC_{exclD}	11064	0.00065	0.00078	0	0.0062
SC_{excl_M}	11064	0.00014	0.00042	0	0.0045
$SC_{overlap}$	11064	0.00039	0.00079	0	0.0061

Table 5.2: Pairwise correlation of independent variables and explanatory variables for all time periods.

	Entry_D	Exit_D	P_M	WC_{exclD}	WC_{exclM}	$WC_{overlap}$	SC_{exclD}	SC_{exclM}	$SC_{overlap}$
P_M	0.0898	-0.0728	1						
WC_{exclD}	0.0392	-0.0763	-0.0091	1					
WC_{exclM}	0.0330	0.4342	0.0688	0.4340	1				
$WC_{overlap}$	0.0319	-0.0584	0.0684	0.4477	0.4287	1			
SC_{exclD}	0.0375	-0.0730	-0.0265	0.6847	0.3016	0.3733	1		
SC_{exclM}	0.0123	0.0067	0.1242	0.2555	0.6657	0.3617	0.1788	1	
$SC_{overlap}$	0.0315	-0.0588	0.0759	0.4434	0.5162	0.6829	0.2970	0.3813	1

Note: Correlation values displayed in the Entry_D column only include observations within Model 1 (as in Table 5.3 and Table 5.4). Correlation values displayed in the Exit_D column only include observations within Model 2 (as in Table 5.5, Table 5.6). Correlation values displayed in all other columns include all values with different industry-region-time combinations.

In our second model, we investigate the relationship between the exit of domestic industries and their cohesion to exclusive MNEs, exclusive domestic or overlapping industries within a region. Using a very similar model as previously, we run a fixed effect panel probit model for the various time periods, given by:

$$\begin{aligned}\Pr(\text{Exit}(i,r,t) = 1 | Z(i,r,t-1), \omega P_M(i,r,t-1), \gamma(i), \tau(r)) \\ = \phi(\alpha + \beta Z(i,r,t-1) + \omega P_M(i,r,t-1) + \gamma(i) + \tau(r) + \epsilon(i,r,t)), \quad (5.11)\end{aligned}$$

where variables are similarly defined as in the first model. In this model, we only consider domestic industries that are already present within a region as an observation.

5.6 Results

5.6.1 Domestic industry entrance

The results of our econometric model in Eq. (5.10) are reported in Table 5.3 and Table 5.4. Each table is also sub-divided into three sections horizontally representing the three time periods we investigate independently.

Recall that the consensus within the regional branching literature is that the presence of related industries enhances industry entry. This is as regions grow by building on existing expertise and fostering new economic activities in related industries [84, 147, 31].

Here, we focus on the relationship between entry of domestic exporting and manufacturing industries and the existing local MNE presence in the form of overlapping and MNE-exclusive sectors. First we consider the pre and post recession periods, and further down we consider the recession period itself.

We start with the impact of overlapping industries on entries. For the first (2006-2009) and third (2015-19) non-recession periods we observe a positive and significant relationship between new domestic entries and cohesion to overlapping industries. This holds for both weighted and strategic closeness in the first period, and strategic closeness in the third. These industries are the more complex industries within a region's industrial basket, mostly consisting of medium-high tech manufacturing, information and telecommunication as well as professional service activities. Well-known examples include Ireland's world-renowned baby food sector, which includes both domestic and foreign-owned global industry leaders. In particular, powdered milk has grown rapidly due to Ireland's large dairy sector. It appears that domestic firms tend to enter new industries proximate to existing dynamic sectors, already home to a mix of domestic and foreign-owned enterprises.

We note that in the third post-recession period, domestic industry entries are associated with strategic closeness to overlapping industries (and not weighted closeness). Hence, during economic recovery, cohesion to a cluster of overlapping industries (which are strongly connected to each other in a region's industrial basket) is associated with a higher probability of domestic industry entry. This result highlights the important role of dense linkages *between* these dynamic sectors in enhancing the domestic export diversification potential of a region.

Turing to MNE-exclusive industries, we observe the opposite effect. Specifically, for the first 2006-2009 period, we find that a domestic firm is unlikely to enter an industry that is close or strategically close to related MNE-only industries. These industries are highly complex and less related to the region's domestic skill-base. For domestic industries to enter MNE-only industries (or those proximate to them) they need to make large cognitive 'leaps' to bridge the capability gap.

Although we cannot disentangle the exact reasons why domestic firms are failing to enter industries linked to complex MNE-dominated industries, our findings very much relate to those on firm absorptive capacity [117, 26], also thought to be factor in Irish domestic firm productivity [18, 65] (absorptive capacity is proxied in both cases by the presence of R&D activity). Other possible reasons include: a potential lack of appropriate training and skill development of Irish workers to be able to work and learn from more complex industries dominated by MNEs, a lack of incentives for MNEs to engage in R&D collaboration with domestic firms [138], as well as potentially less strategic investment decisions by government agencies to encourage domestic firms to enter into MNE markets [49, 50].

What is particularly striking about the recovery period is the strong significant and positive relationship between domestic industry entrance into MNE-only industries. Although we show no causal link, our analysis might be picking up the Irish response to Brexit, and particularly the effect of two schemes that aimed to support domestic firms. These are the 'Global Sourcing Initiative' which aimed to create business opportunities for Irish-owned companies with MNEs and the 'Market Discovery Fund' which supported Irish-owned exporting firms to move into new products.

We now consider the crisis period (2010-14). During this period there is no significant relationship between industry entrance and cohesion to overlapping or exclusive MNE industries. It appears that entries into industries characterised by relatedness to complex sectors ceased during this difficult period. However, we observe a positive and significant relationship for both *weighted* and *strategic* closeness to exclusive domestic industries which are dominated by low tech manufacturing and agriculture-related sectors. Our results suggest that in a time of recession, new domestic export and manufacturing activities enter into regions where there is densely connected existing (exclusive) domestic activity. This contrasts with more entries into regions with related overlapping industries in non-recession periods as seen above.

We show results combining WC and SC in a single model in Table C.3 in Appendix C.2. We generally find that our results hold. Our cohesion measures remain significant when controlling for the other cohesion measure, demonstrating empirically that these measures pick up different dimensions of cohesion.

Overall, we observe that both before and after the recession period cohesion to overlapping industries is associated with a higher probability of entry. In contrast, cohesion to MNE-only industries is associated with a lower probability of entry pre-crisis. Post crisis, however, domestic entries are associated with MNE-exclusive industries, perhaps due to various schemes aimed at generating strong domestic-MNE links in response to Brexit.

5.6.2 Domestic industry exit

The results of our second econometric model in Equation 5.11 are shown in Table 5.5, and Table 5.6 for the cohesion measures WC and SC, respectively.

Within both the regional branching and resilience literature it is generally accepted that industries which are less related to a region's industrial basket are more likely to exit [147]. Hence, relatedness amongst industries provides a protective shield against industry exit in response to economic shocks [76, 14].

Here we focus on the presence of MNE activity and the survival of domestic government supported export and manufacturing industries, particularly during the crisis period. Again, we look at the distinct impact of overlapping and MNE-exclusive industries.

As before, we first examine the role of overlapping industries. In the 2006-2009 period, we see a significant and negative effect of the presence of MNEs within an industry on exit of the domestic counterpart. Hence, co-existing with MNEs within an industry (and being an overlapping industry) enhances domestic resilience - but only in the pre-recession period. Similarly, Görg and Strobl [93] found a positive impact of MNEs on domestic manufacturing plants in high-tech sectors between 1973-96.

Similar to entries and co-existence with MNEs, *cohesion* to overlapping industries also plays a key role in domestic industry survival. We find a negative and significant relationship between weighted closeness (within the first two periods) and strategic closeness (within the latter two periods) and domestic industry exit. Hence, the more cohesive an industry is to these dynamic industries the more likely it is to survive. The protective role of these industries during the recession period, a time when industry exits peaked, accentuates their importance.

We again observe, similar to entries, that in the latter recovery period it is only the relationship between domestic entries and strategic closeness to overlapping industries that remains significant (and no longer that of weighted closeness). This again highlights the importance of the dense linkages between related sectors and thereby the formation of a cluster or concentration of related and dynamic activities for the survival of domestic industries.

Next, we turn to MNE exclusive sectors. In contrast to the protective effect of overlapping MNEs, in the recession period we observe a significant and positive effect of both weighted and strategic closeness to MNE-exclusive industries. Hence, government supported domestic exporting and manufacturing activities that were related to MNE-only industries were more likely to exit, highlighting the fragility of domestic firms operating in sectors most distant from the broader domestic capability base and perhaps dependent on their role in international supply chains.

We also show the results for the SC variable when controlling for the corresponding WC measure in Table C.4 in Appendix C.2. As before, our cohesion measures remain significant when controlling for the other cohesion measure.

Overall, it appears that cohesion to overlapping industries has a protective effect in a crisis. Cohesion to exclusive MNE industries, however, was associated with an increased probability of exit during this period.

Table 5.3: Panel probit regression results for domestic industry entrance between 2006-2019 and their weighed closeness cohesive measure as independent variable

Baseline period	2006-2009					2010-2014					2015-2019				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
P_M	-0.439	-0.433	-0.477	-0.568	-0.578	0.021	-0.024	0.048	0.020	0.011	1.673***	1.642***	1.690***	1.715***	1.673***
	(0.315)	(0.315)	(0.319)	(0.332)	(0.333)	(0.306)	(0.306)	(0.308)	(0.306)	(0.309)	(0.451)	(0.453)	(0.457)	(0.465)	(0.470)
WC_{excD}		0.047			0.047		0.132*			0.134*		0.065			0.106
		(0.054)			(0.056)		(0.076)			(0.078)		(0.100)			(0.107)
WC_{excM}			-0.259**		-0.185*			0.209		0.246			0.169		0.11
			(0.138)		(-0.146)			(0.145)		(0.165)			(0.218)		(0.229)
$WC_{overlap}$				0.231***	0.222***				0.020	0.066				0.168	0.186
				(0.072)	(0.074)				(0.075)	(0.089)				(0.107)	(0.115)
Region FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Constant	-8.841	-8.405	-8.302	-8.323	-8.590	-42.689	-9.867	-19.189	-25.688	-19.177	-16.145	-17.634	-15.652	-18.6160	-14.952
	(8.14e+06)	(4.97e+06)	(8.14e+06)	(2.85e+06)	(4.21e+06)	(5.76e+06)	(5.76e+06)	(5.76e+06)	(5.75e+06)	(5.75e+06)	(5.76e+06)	(5.75e+06)	(5.75e+06)	(5.75e+06)	(5.75e+06)
N	2522	2522	2522	2522	2522	2494	2494	2494	2494	2494	2507	2507	2507	2507	2507
AUC	0.9480	0.9481	0.9480	0.9508	0.9509	0.9700	0.9703	0.9706	0.9700	0.9710	0.9814	0.9811	0.9819	0.9819	0.9821

Notes: Robust standard error in parenthesis; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5.4: Panel probit regression results for domestic industry entrance between 2006-2019 and their strategic closeness cohesive measure as independent variable

Baseline period	2006-2009					2010-2014					2015-2019				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
P_M	-0.439	-0.440	-0.322	-0.529	-0.404	0.021	0.014	0.019	0.020	0.021	1.673***	1.656***	1.681***	1.668***	1.631***
	(0.315)	(0.315)	(0.320)	(0.327)	(0.332)	(0.3055)	(0.307)	(0.306)	(0.305)	(0.307)	(0.451)	(0.453)	(0.453)	(0.495)	(0.502)
SC_{excD}		45.200			105.200		640.199**			686.141**		178.106			349.103
		(183.202)			(185.686)		(262.956)			(271.562)		(360.212)			(372.565)
SC_{excM}			-529.264**		-507.548**			-142.417		-153.707			-80.054		-29.998
			(286.142)		(295.121)			(238.175)		(253.688)			(398.120)		(464.021)
$SC_{overlap}$				209.728*	191.619*				150.549	210.868				964.004***	1039.400***
				(140.517)	(149.109)				(197.860)	(216.355)				(437.841)	(466.363)
Region FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Constant	-8.841	-8.177	-8.175	-8.109	-8.281	-42.689	-30.333	-26.834	-31.729	-22.358	-16.145	-19.776	-24.293	-14.870	-18.664
	(8.139e+06)	(3.767e+06)	(4.055e+06)	(3.593e+06)	(2.987e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)
N	2522	2522	2522	2522	2522	2494	2494	2494	2494	2494	2507	2507	2507	2507	2507
AUC	0.9480	0.9481	0.9490	0.9485	0.9494	0.9700	0.9725	0.9699	0.9701	0.9726	0.9814	0.9814	0.9814	0.9826	0.9829

Notes: Robust standard error in parenthesis; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5.5: Panel probit regression results for domestic industry exits between 2006-2019 and their weighted closeness cohesive measure as independent variable

Baseline period	2006-2009					2010-2014					2015-2019				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
P_M	-0.607**	-0.570**	-0.593**	-0.586**	-0.569**	-0.140	-0.140	-0.118	-0.036	-0.048	0.025	0.028	0.028	0.040	0.114
	(0.271)	(0.281)	(0.272)	(0.274)	(0.286)	(0.255)	(0.255)	(0.257)	(0.264)	(0.265)	(0.334)	(0.337)	(0.335)	(0.334)	(0.339)
WC_{exclD}		-0.304***			-0.337***		-0.068			-0.103		-0.146			-0.128
		(0.097)			(0.102)		(0.069)			(0.077)		(0.086)			(0.090)
WC_{exclM}			0.094		-0.049			0.340***		0.235**			-0.726		-0.269
			(0.150)		(0.163)			(0.166)		(0.191)			(0.169)		(0.173)
$WC_{overlap}$				-0.148**	-0.204**				-0.372***	-0.372***				-0.093	-0.105
				(0.084)	(0.102)				(0.098)	(0.111)				(0.335)	(0.078)
Region FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Constant	-1.193*	1.165	-1.207*	-1.030	1.623	-21.305	-14.167	-10.715	-16.713	-16.943	-15.163	-12.568	-14.402	-14.734	-15.568
	(0.622)	(0.979)	(0.623)	(0.641)	(1.036)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)
N	1166	1166	1166	1166	1166	1194	1194	1194	1194	1194	1181	1181	1181	1181	1181
AUC	0.9466	0.9512	0.9460	0.9455	0.9514	0.9479	0.9478	0.9501	0.9549	0.9561	0.9650	0.9657	0.9656	0.9645	0.9662

Notes: Robust standard error in parenthesis; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 5.6: Panel probit regression results for domestic industry exit between 2006-2019 and their strategic closeness cohesive measure as independent variable

Baseline period	2006-2009					2010-2014					2015-2019				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
P_M	-0.607**	-0.627**	-0.600**	-0.569**	-0.589**	-0.141	-0.144	-0.175	-0.001	-0.031	0.025	-0.001	0.024	0.068	0.036
	(0.271)	(0.275)	(0.272)	(0.273)	(0.277)	(0.256)	(0.255)	(0.258)	(0.263)	(0.265)	(0.334)	(0.335)	(0.334)	(0.337)	(0.341)
SC_{exclD}		-553.214**			-552.652**		-143.621			-295.173		-334.402			-441.893
		(284.890)			(286.079)		(262.738)			(275.891)		(296.985)			(303.295)
SC_{exclM}			214.396		152.654			546.481**		574.382**			31.224		54.843
			(296.377)		(303.343)			(290.018)		(294.286)			(427.907)		(450.899)
$SC_{overlap}$				-222.324	-204.248				-790.051***	-754.955***				-513.383**	-563.335**
				(217.059)	(217.172)				(260.896	(276.587)				(240.303)	(244.238)
Region FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Constant	-1.1930*	-0.077	-1.200	-1.103	-0.011	-21.305	-16.306	-12.153	-13.642	-17.084	-15.163	-16.384	-13.290	-14.663	-17.446
	(0.622)	(0.859)	(0.822)	(0.835)	(0.866)	(3.323e+06)	(3.323e+06)	(3.323e+06)	(3.323e+06)	(3.323e+06)	(3.323e+06)	(3.323e+06)	(3.323e+06)	(3.323e+06)	(3.323e+06)
N	1166	1166	1166	1166	1166	1194	1194	1194	1194	1194	1181	1181	1181	1181	1181
AUC	0.9466	0.9480	0.9461	0.9462	0.9473	0.9479	0.9475	0.9491	0.9516	0.9515	0.9650	0.9655	0.9649	0.9654	0.9659

Notes: Robust standard error in parenthesis; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

5.7 Conclusions and policy implications

It is well-established that regions grow by learning how to recombine complementary capabilities to move into more complex and sophisticated economic activities. The growth trajectory of a region is therefore conditioned on it's current capability base. With the rise of multinational enterprises globally, an issue of key concern is whether and how MNEs can act successfully as a channel to 'import' new capabilities to a region and promote local domestic diversification.

Focusing on a set of government-supported Irish firms active in manufacturing and exports, we find a strong role for so-called overlapping industries - those that have both domestic and MNE employment in a region. Specifically, domestic industries are both more likely to enter and less likely to leave a region if they are closely related to these industries. These are some of the most dynamic domestic sectors, home to global brands such as Glanbia, the Irish baby food manufacturer. While we cannot separate the role of MNEs from domestic firms within this subset of industries in terms of new entries, we can deduce that there exists a set of related sectors in which domestic firms already successfully thrive alongside MNEs, and it is this set that is driving new entries.

In contrast to overlapping industries, we find a negative impact from 'exclusive MNE' industries. In particular, we find that cohesion of a domestic industry to MNE-only industries both reduces its chances of entry into a region as well as reduces its chance of survival in a crisis. These results suggest that domestic firms are unable to 'leap' into MNE-proximate industries, likely due to a technology or know-how gap that is too large to bridge. Furthermore, those that do are less likely to survive, suggesting weak ties.

Finally, the type of cohesion matters. We differentiate between simple relatedness and strategic relatedness, or the presence of 'higher order' connections *between* industries in a region. Our results show that, specifically in the recovery period, strategic closeness to overlapping industries is highly significant for both the entry and survival of domestic industries, while weighted closeness is not, in a multivariate setting. Hence, entries tend to occur in industries proximate to regional 'clusters', or groups of existing interconnected overlapping industries. Similarly, exits are negatively correlated with strategic closeness to overlapping industries, further suggesting that deep embeddedness in a regional network of related industries is key to success.

O'Clery *et al.* [158] identified two distinct export clusters in the Irish economy which they visualised in the Product Space network [104]. One of these clusters,

located centrally in the network, contained industries with both Irish and MNE activities such as food and agriculture, while the other cluster was more peripheral, containing very complex sectors such as chemicals, pharmaceuticals and electronics and dominated by MNE. The authors' hypothesised that the 'distance' between these sectors would likely prevent domestic firms emerging in MNE sectors due to the huge capability gap. Our findings here very much accord with this hypothesis. Specifically, we find that the 'rich get richer' in the sense that overlapping industries in a region tend to attract further related industries. On the other hand, domestic entries into MNE dominated industries are not only rare but decrease in likelihood the closer the target industry is to MNE activities. The only exception to this dynamic is a recent cluster of entries into MNE-exclusive sectors, which might be explained by government-led efforts to stimulate and boost domestic-MNE links in response to Brexit.

5.7.1 Policy implications

There are a number of clear policy implications from our study. Firstly, our study suggests that industrial policy should distinguish types of MNEs, looking closely at sectoral concentrations and the potential for linkages to the domestic economy. Crucially, it suggests that policy should prioritise MNEs in overlapping industries, or those proximate to overlapping industries, as these have the greatest likelihood of inducing domestic transformation. This idea is somewhat at odds with a general strategy aimed at increasing the industrial or export complexity of a nation, or focusing on taxation income alone, irrespective of the domestic capability base. These findings are particularly salient within a context of finite investment resources and a potential global re-organisation of MNE activities resulting from international agreements on MNE taxation.

Secondly, in line with a burgeoning literature, methods from network science and data science can provide invaluable novel insights into the structure and dynamics of economic processes. In this case, we harness data on inter-sectoral labour mobility and network science to quantify the cohesion between industries. Establishing a mathematical model for cohesion, and differentiating types of cohesion, is key to predicting which industries are well-placed to enter a region, providing an evidence base for industrial policies such as grant funding, training and R&D programmes, and infrastructure investment. Such modelling is rarely conclusive on its own, but forms part of a package of analysis which can be used to develop informed and strategic investments.

Thirdly, our data suggests a possible connection between domestic industry entry in the final period and various schemes designed to generate stronger domestic-MNE links in response to Brexit. One scheme in particular, the 'Global Sourcing Initiative', aimed to stimulate new domestic-MNE partnerships via one-to-one meetings. Over 450 such meetings were held in 2017 during a single two-day event. Another scheme, the 'Market Discovery Fund', supported Irish-owned exporting firms to move into new products and geographies, and effectively doubled the size of the IDA (Investment and Development Agency) budget between 2016 and 2017. While we do not establish any statistical or causal relation between domestic entry and these schemes here, this would undoubtedly be an interesting and worthwhile avenue to pursue for future work. Furthermore, it would also be interesting to investigate the survival of these Brexit era entries to see whether their survival is comparable, better or worse than entries during other periods. These questions are important for a wider literature which studies leapfrogging, or unrelated diversification, which is known to be rare [175]. A better understanding of the conditions under which firms can 'leap' into sectors that are distant from the underlying capabilities of the local economy is a key priority for evidence-based industrial policy.

5.7.2 Limitations and future work

We note that there are some limitations to this study. Most obviously, we study the dynamics of a particular subset of Irish firms, namely those that are supported by government agencies. It is possible that our results are influenced by some characteristics of this set relative to other firms. For example, its entirely plausible that domestic firms that are somehow related to MNEs may be more likely to receive government assistance by virtue of either exporting or important links to MNE firms. Yet, in this case, our results distinguishing the effect of overlapping industries from MNE exclusive industries remain potent and suggest that any effort to support domestic firms in or proximate to MNE exclusive sectors has had a limited effect. Another possible issue is that the industries we observe as new entries are in fact just entries of existing domestic firms into this dataset, perhaps motivated by the various Brexit schemes mentioned above. Examination of a subset of firm level data for domestic firms entering MNE exclusive sectors in 2017-18 (not shown due to privacy concerns) suggests that less than 5% of firms entered with more than 10 employees during this time, and so this is not likely a major driver of entries. A fuller analysis of all Irish firms is warranted to generalise the observed patterns, should such data be made available to researchers in the future.

Additionally, our data does not include services outside the exporting sector which are increasingly a large part of economic activity and employment both in Ireland and globally. Due to the knowledge intensive nature of these activities, we would expect that the relationships found here would both generalise and become stronger for service industries [67] but this remains to be tested. Finally, as discussed above, we are limited by the aggregate nature of our data, particularly in disentangling the influence of overlapping industries. Firm level data would enable a deeper analysis in future work.

Finally, our results show that directly related industries and higher-order linkages influence industrial diversification processes. Therefore, this work provides a rationale for developing modelling tools that consider more of the underlying SRN structure. Most techniques (*e.g.*, density) only consider the presence of neighbouring industries and thereby take a local approach. However, as argued in the previous two chapters and by O'Clery *et al.* [161] the modular structure of these networks indicates skill basins. These skill basins represent labour pools where workers can move freely between industries. The authors also showed how employment within these labour pools is a good predictor of industry employment growth. An extension of our work could consider the impact of the presence and number of MNEs in these industrial clusters on domestic industry dynamics. Furthermore, how the structure of these labour pools (*i.e.*, having high modular structure and being particularly isolated or being well connected to the rest of the network) influences the degree to which MNE-domestic knowledge spillovers occur.

Chapter 6

An Application: The South African Labour Flow Network

In this chapter we model and analyse the South African labour flow network to investigate how an affirmative action policy has influenced labour mobility patterns. Using quasi-experimental techniques (the regression discontinuity design) with network analysis tools we show the casual impact of the policy on local network structure changes.

6.1 Introduction

There is a vast body of interdisciplinary literature investigating the impact of group-based affirmative action (AA) policies on labour market outcomes. These include the impact of these policies on employment representation [140, 127], wages [121] and occupational mobility [120], to name a few. However, few studies have investigated how these policies impact labour mobility patterns. More specifically, their indirect effects on the sectoral diversity of newly hired employees.

The skills and knowledge present within a firm's workforce is one of its key competitive advantages and heavily influences its performance. One of the key mechanisms by which firms grow, is by gaining new skills and knowledge through hiring new employees. These skills can then be recombined with existing skills within the firm to enhance firm's performance, produce new products or services, or allow the firm to enter into new markets [10, 23, 32]. Thereby, inter-firm labour mobility is key in allowing firm learning and firm growth to occur [131, 75].

In the innovation management literature, a significant focus has been placed on the skill diversity of employees. It has been shown that firm's with a highly skill-diverse workforce displays higher degrees of firm creativity and innovation [150]. Similarly,

new employees which enhance the skill-diversity of the firm enhance the firm's innovative ability. However, this literature also stresses that not all new employees bring valuable skills that trigger innovation and positively benefit a firm's performance [23, 157].

For newly hired workers to benefit their firm positively, they need to bring in skills and knowledge closely related to the knowledge already present within the firm. This literature stresses two key factors. First, the firm's absorptive capacity which refers to the firm's ability to communicate, understand and integrate external knowledge [47]. Secondly, the type of external knowledge needs to be cognitively close but not too similar to the existing knowledge base of the firm [157]. This is often considered to be employees, not in the same industry as the firm, but rather from a prior background in a skill-related industry [30]. This is because the type of knowledge needs to be similar enough to the firm's existing skill-set to ensure low co-ordination costs, but not too similar to avoid lock-in and competition occurring between employees.

As previously discussed within this thesis, evolutionary geographers also stress the key role of skills and knowledge in the economic development path of regions [101, 104]. Here, emphasis is placed on labour mobility as the primary channel through which skills and knowledge diffuse within a regional economy [146]. This channel allows for new combinations of existing skills and know-how present within a region that allows for the creation of new economic activities. Similar to the firm-level, this literature emphasises the importance of skill-diversity. An industry that receives workers from a wide diversity of different sectors display higher employment growth and higher resilience in the face of economic shocks [84].

To ensure regional economic prosperity and growth, it is crucial for governments to understand the patterns of labour mobility within a region. Similarly when implementing AA policies, one needs to understand how these policies influence these labour mobility patterns, specifically their influence on the sectoral diversity of newly hired workers both of firms and industries.

In this study, we focus on the impact of the South Africa Employment Equity (EE) Act No. 55 of 1988 on gendered labour mobility patterns. The EE act was introduced by the South African government to address labour market inequalities cause by the Apartheid regime. The act's primary aim is to ensure fair representation of black people[1], women and people with disabilities in all sectors, occupations and levels of the workforce through implementing affirmative action [39]. The act requires

[1]Black people consists of all African, Coloured and Indian people, as well as people of Chinese descent.

designated employers to increase employment opportunities and the hiring of blacks and women. This includes meeting set racial and gender profile numerical targets, reduce differences in remuneration between employees of different races and genders, and reduce any discriminatory labour market practices.

The act's impact is controversial amongst both the public and policy-makers [143]. It has been criticized for creating a brain drain [106], causing a greater skill mismatch [69] and reducing the productivity of the workforce [38, 119]. Advocates of the policy, however, argue that the act is vital to reverse the self-reinforcing inequality caused by the structure of the labour market [200]. Various studies have offered evidence that the act has enhanced the employment and occupational status of blacks and women [166].

In this study, we contribute to this literature by investigating the impact of the act on a *firm*'s sectoral diversity of newly hired male and female workers. We hypothesise that the act induced firms to diversify their recruitment of female workers to a larger number of sectors to meet their gender profile numerical targets. Furthermore, as not all labour inflows *per se* positively influences firms, we also consider how the act has influenced the *related-* and *unrelated variety* [84] of newly hired skilled male and female workers in firms. We also investigate whether male-dominant *industries* have been more heavily impacted by the act as they require the largest workforce restructuring and therefore display a larger sectoral diversity of newly hired female workers. According to the author's knowledge, we are the first study to investigate an AA policy's indirect impact on newly hired employees' sectoral diversity.

To investigate the impact of the act, we first construct an inter-firm labor flow network. This is a network of firms with edge weights corresponding to the count of worker transitions between firms. To quantify the sectoral-diversity (the diversity of sectors from which new workers flow into a firm) we construct different entropy-centrality measures on the network (total inflow diversity, the related-variety and unrelated-variety). We then use a regression discontinuity design with these centrality measures to evaluate the act's impact. Specifically, our analysis exploits a clear cutoff created by the act's adoption legislation that requires all firms with 50 or more employees to comply with the act. Therefore, we compare firms with slightly less than 50 employees who are exempt from the act to those with slightly more than 50 who comply.

The RD results reveal that the act increases a firm's total inflow diversity (the total sectoral diversity) of newly hired female workers, while the corresponding impact for men was found to be negligible. Furthermore, the act increased the related variety

(the diversity of new worker's skills that are related but not the same as the firm) of newly hired skilled female workers. However, unlike in the literature, where these types of labour flows have been shown to enhance firm performance, we do not find any significant impact of the act on firm performance.

We further investigate the act's impact on industries relative to their male employment share. We do this by evaluating the relationship between the percentage of male employment within an industry, and the industry's female labour inflow diversity[2] through a regression model. We find that male-dominant industries have a higher sectoral diversity of newly hired female workers. This relationship is significantly stronger amongst the group of firms who comply with the act compared to those exempt from the act.

In the next section, we explore related literature on the EE Act, labour mobility and labour market regulation evaluation methods. This follows, in Section 6.3, with a discussion of the data used in this study. Next, network and metric construction are introduced in Section 6.4, this includes the construction of a new industry classification based on the skill-relatedness network. The regression discontinuity design and regression models are then presented in Section 6.5. In Section 6.6 our results are shown. Finally, in Section 6.7 the results and implications of our study are discussed along with potential avenues for future work.

6.2 Literature

6.2.1 Employment Equity Act

The apartheid regime led to a high level of inequality within the South African labour market by systematically and purposefully restricting the majority of black people, women and people with disabilities (collectively referred to as previously disadvantaged individuals or PDIs) from economic and social opportunities. This has greatly affected income distribution and gender and racial representation within sectors, occupations and workforce levels in the South African labour market [45, 64].

Although much has been done to rectify these effects, the current, legally nondiscriminatory South African labour market is still socially inequitable. A Green Paper on Employment Equity [64] uncovered high levels of labour market discrimination, showing that race and gender (even after controlling for education, age, occupation and sector) are strong factors in determining an individual's probability of obtaining

[2]Computed analogously to the firm-level inflow diversity using an inter-industry labour flow network.

work and predicting their corresponding remuneration. Whites earn 104% more than blacks, and men receive 43% higher wages than women who are similarly qualified, in the same sector, and occupation [64].

In response to these issues, the Employment Equity Act was introduced in 1998 to enhance the re-entry of PDIs into the broader economy and accelerate their upward career trajectory into higher-paying and higher-skilled occupations and sectors. The act aims to achieve this by requiring firms to implement an AA plan [196]. This consists of implementing measures, such as preferential treatment in recruitment to enhance the representation of PDIs in all occupational categories and levels in the workforce. Firms must also retain and develop PDIs by implementing appropriate training and employment growth opportunities. Furthermore, in consultation with the EE committee, firms set targets for their racial and gender quotas. They are also obliged to report annually on these targets, as well as differences in the remuneration and benefits received by their employees across various categories, including gender and ethnicity. As of 2022, the act is currently still in effect.

The act is compulsory for all *designated employers*. A designated employer is any South African firm with more than 50 employees. However, suppose a firm has a workforce smaller than 50 employees but generates an annual revenue above a certain threshold (dependent on the industry in which the firm operates). In that case, the firm is also classified as a designated employer [64]. Various industries are exempt from the act, namely: the South African defence force, the secret service and the national intelligence service. For further detail regarding which firms need to comply with the act, the reader is referred to the act's documentation (see [196]).

There are financial penalties for a failure to comply with the act by designated employers. The size of these penalties vary on a case by case basis which depends on the degree of non-compliance and number of previous offenses. For a first time offense, where a designated employer has failed to prepare or implement an EE plan, an employer is subject to a fine the greater of R1.5million (approximately £74 000) or 2% of the employer's turnover. For repeated offenders the fine increases to the greater of R2.7 million (approximately £134 000) or 10% of their annual turnover [196].

The act is highly controversial among policy-makers, the private and public sectors. Advocates of the act contend that preferential policies break down negative views about previously disadvantaged individuals by enabling them to demonstrate their capabilities [48]. Many economists also argue that market forces alone are unable to solve the problem of discrimination, and therefore, regulations that impose change on structural labour market characteristics are vital [200]. Critics, however,

argue that the EE Act has led to a brain drain [106], as it incentives immigration of the skilled minority population. The act has also been blamed for reducing firms' productivity by dropping general standards of labour and thereby increasing the cost of doing business [38, 119]. It has also been criticized for reducing foreign investment in South Africa [69]. Furthermore, it has led to many high-skilled vacant jobs, and under-skilled employees as there is a limited labour supply which meets both the requirements of the act and high-skilled job specifications [106].

Various studies have investigated the act's impact on high-level equality and economic outcomes. For example, research suggests that the act has increased the representation of PDIs within managerial positions and both highly-payed and higher-skilled occupations [28, 197]. Furthermore, studies have indicated that the act has impacted employee engagement, levels of discrimination in the workforce [39] and levels of foreign direct investment [69]. However, most of these studies are case-study based (focusing on a specific firm, industry or region [197]). We add to this literature by investigating the act's indirect impact on the sectoral diversity of newly hired PDIs.

6.2.2 Inter-industry labour mobility and skill diversity

A vast literature shows that human capital (the knowledge and skills embodied by workers) within a firm is a key competitive advantage and positively impacts a firm's performance and growth [150]. For firm's to expand their human capital they can hire new employees who bring new sources of knowledge and trigger new ideas. This can positively influence a firm by enhancing it's productivity, enabling the development of new products and services, and also provide the right skills for the firm to move into new growing markets [131, 75]. Hence, labour mobility is the key channel through which firm learning and growth occurs.

Many empirical studies have confirmed this. For example, Almeida *et al.* [6] showed how labour mobility was responsible for knowledge spillovers which allowed for the rise of many successful firms in Silicon Valley. Similarly, Pinach *et al.* [174] found that labour mobility facilitated the knowledge creation and innovation that characterised the British motorsport industry. Another study also claims that the higher rates of labour mobility found in urban centres explain their 'urban-productivity' premium [8].

Although labour mobility can positively enhance firm performance, it may also hinder human capital developments due to labour poaching. High levels of labour mobility may be detrimental to firms as they lose key personnel to their competitors.

This may reduce their incentives to train and up-skill their workforce [114]. Argote *et al.* [2] found that high levels of personnel inflow reduced organisational learning and productivity. Various studies investigating the overall impact of high labour inflows on firm performance have also found no positive effect [30, 74].

More recently, the literature has started to consider the type of skills and knowledge that new employees bring to a firm. Authors have found that a high diversity of skills enhances a team's creativity and innovation [23]. Consider a firm as a team, where employees work together to produce a service or product. The greater the variety of skills and know-how amongst employees, the higher degree of specialization possible - each employee is able to specialize in a subset of tasks to produce the final product or service. This high degree of specialization gives a firm competitive advantage and enhances its productivity. However, a too broad skill diversity can also lead to high coordination costs [24]. This is the diversity of skills leads to more challenging communication and skill integration, which reduces a firm performance. Therefore, the type of skills that a new employee brings to a firm should be carefully considered in relation to the firm's existing skill base.

When focusing on the hiring of new employees, authors have found that firms benefit the most from new skills and knowledge that are close to the firm's existing knowledge base. This is as these new employees are more easily understood and integrated into the firm [47]. However, other authors have also argued that new skills that are too close to the firm's existing skills and know-how will only produce competition for existing employees and negatively influence firm performance. Noteboom *et al.* [157] argues that inter-firm learning requires a certain degree of cognitive proximity between firms to enable effective communication, but not too much cognitive proximity to avoid lock-in. The authors showed a U-shaped function between the cognitive distance between large firms' technology-based alliances and the firm's innovation performance.

Various empirical studies support Noteboom's [157] claim. Song *et al.* [191] found that the inflow of engineers that were technologically related but different from the existing expertise of the firm had a greater positive effect on a firm's growth. Furthermore, Boschma *et al.* [30] found that plants performed better when new employees were recruited from industries technologically related to the industry of the plant. In contrast, they found that employees recruited from non-related industries or the same industry reduced or had no effect on plant performance.

Inter-industry labour mobility is also vital for regional economic development. Leaning on the evolutionary economic geography and economic complexity literature,

regions grow by combining existing knowledge and skills in new ways to create new economic activities [150, 104]. Inter-industry labour mobility is the primary mechanism through which these different skills and knowledge are recombined. Hence, the patterns of labour mobility can either foster or hamper regional industrial diversification opportunities [147, 83]. In this literature, authors have found that regions with high inter-industry labour mobility amongst a diverse set of industries are associated with greater employment growth, and higher levels of industrial diversification [84].

Apart from the knowledge transfer argument, inter-industry labour mobility also enables structural change in an economy which is key to enable long-term economic growth and the survival of an economy. This is because an economy is always under a form of economic decline in some sectors. It therefore needs to reallocate redundant workers from these sectors to growing sectors [170]. To allow this process of creative destruction to occur smoothly and with a limited cost of adjustment, inter-industry labour mobility is required. Hence, high inter-industry labour mobility allows for a resilient labour market, where workers can be absorbed into related industries in the face of a industry-specific shock [68]. However, high inter-industry labour mobility may also be problematic for an economy. It may prevent the formation of specialized knowledge [68], lead to job insecurity and workers who struggle to cope with the impact of change [176].

It is therefore crucial to understand labour mobility patterns within an economy. In this study, we investigate how the EE Act has influenced the sectoral-diversity of newly hired male and female workers. Furthermore, taking into account that the type of skills matter, we also consider how the act effected the related- and unrelated variety [84] of these newly hired male and female workers. Related variety refers to the skill-diversity of new employees that have related skills (come from a prior industry related to the new industry in which they are employed), while unrelated variety is the skill-diversity of new employees that have unrelated skills (come from a prior industry unrelated to the new industry in which they are employed) according to the firm where they have been hired.

In South Africa, Kerr [113] was the first to quantify the degree of labour mobility, specifically inter-firm labour mobility. The authors found that approximately 53% of all workers switched firm per year during the 2011–2014 period. Worker flows were also found to be highly heterogeneous across various factors, including: firm size, firm earning rates and industry. With respect to industries, Kerr investigated worker

flows *within* 34 industry sectors[3]. The largest worker flows were found within the household services and hospitality sectors. On the other hand, the smallest worker flows were found within the public administration and mining sectors. In our study, we investigate the sectoral-diversity of firms' new hires, which corresponds to *cross* sectoral flows.

As we investigate sectoral-diversity of new hires along gender lines, we also draw from a related literature regarding the socio-economic factors which influence the female labour participation rate[4] and female labour mobility. South Africa's female participation rate was 48.77% in 2019, which is lower than the male participation rate of 62.59% [39]. The main factors in the literature influencing participation include the level of economic development, the level of female educational attainment, social dimensions (such as social norms around marriage, fertility and the woman's role outside the household), access to credit and access to childcare [111, 85]. In general, policy reforms focused on increasing female participation and mobility aim at influencing one of these aforementioned factors. The EE Act is a long-term policy reform that is focused mainly on increasing recruitment opportunities for women and removing discriminatory recruitment practices in the workplace. The act therefore aims to influence social dimensions (specifically social norms on how employers see women), as well as long term female educational attainment and training (by providing more opportunities for women to enter into higher-skilled employment) [192, 166].

6.2.3 The evaluation of labour market regulations

To evaluate the impact of the EE Act on the sectoral diversity of worker inflows, as well as the more specific related- and unrelated-variety of skilled inflows, we require a quantitative approach that can robustly estimate the causal impact of the act on these outcome variables. We, therefore, turn to quasi-experimental designs. These are statistical techniques that construct a counterfactual without the need for randomised controlled trials. A counterfactual refers to a comparison group established to assess the difference in outcomes for individuals who are subjected to intervention and those who are not. In this study, we specifically adopt the Regression Discontinuity Design (RDD) as our quasi-experimental design approach.

[3]The industry sectors were classified according to a high-level SARS industry classification measure.

[4]The labour force participation rate is the proportion of the country's working-age population that actively engages (either by working or seeking work) in the labour market [111].

An RDD is a well-known method that has emerged as one of the most credible non-experimental strategies for the analysis of causal effects [42]. It does this by exploiting a discontinuity in the treatment assignment [42, 108]. For example, in our case, whether a firm must comply with the act is based on the firm's size. Note: this method applies specifically to cases where an administrative decision is based on a continuous measure. The method calls for a control and treatment group to be constructed and their outcome variables compared. Observations that are very close to the "cut-off" (the value which divides the two groups) can be assumed to be the same. Comparing these observations, therefore, simulates a random assignment.

However, in order to quantify the sectoral diversity of new hires, we need to combine the RDD approach with a network perspective. Although the popularity of networks has grown dramatically over the last decade, very few studies have combined network analysis with quasi-experimental design policy evaluation tools. This is one of the first studies to integrate these techniques to evaluate the impact of a labour market regulation on changes in the diversity of labour inflows.

In labour economics, labour flow networks (LFN) have primarily been used to understand the structure and topology of a labour market. Omar *et al.* [96] were the first to construct an inter-firm LFN for an entire economy. This is a network in which the nodes represent firms and the edges the number of workers who transition between the corresponding firms. The authors then used various network analysis tools to investigate the degree of labour mobility within the labour market as well as its community structure. Inter-industry LFN is analogous, with industries as nodes and the number of worker transitions between the corresponding industries as edges. In this study, we construct both an inter-firm and inter-industry LFN for the formal labour market in South Africa. These networks enable us to quantify the sectoral diversity, as well as the related- and unrelated-variety, of new hires to both firms and industries.

We also construct a skill-relatedness network (SRN) [149], which is a normalised inter-industry LFN that 'corrects' the total flows (edge weights) to take into account the size of flows to and from the origin and destination industry. A detailed review of this network is presented in §2.5. This network is used to redefine sectors, in a similar manner to [161], within our study. We redefine sectors by grouping industries according to their degree of skill similarity using a community detection algorithm applied to the SRN. The redefined sectors enable us to more accurately quantify the diversity of newly hired workers by reducing errors due to, *e.g.*, differences in the granularity of industries in the official classification. Our study further adds to this

literature on labour flow networks, not only by constructing the first SRN for South Africa but also by using this network in order to construct a new skill-based industry classification.

6.3 Data

To evaluate the impact of the EE Act on the labour inflow diversity, we use an administrative dataset constructed from anonymised tax records for the period $2011 - 2014$ [173]. The National Treasury and the South African Revenue Services (SARS) recently made this dataset available to researchers. This longitudinal dataset comprises all employer-employee spells within the formal economy. For each, details regarding the employee, the employer and the duration and terms of employment are included. To evaluate the labour inflow diversity, we use this dataset to count inter-firm and inter-industry worker transitions. Note that we only include worker transitions of formal employees in full-time employment.

As not all labour flows positively impact a firm, we also investigate the act's impact on the related- and unrelated-variety of newly hired *skilled* workers. Here we do not include all labour mobility but only count worker transitions if the worker is a high-income earner and thereby earns more than R400 000 annually (this is the average starting salary of an engineer). We, therefore, roughly only include the top 25% of formal workers according to their annual wage. This subset of workers is a proxy for the skilled workers within the South African economy.

Secondly, to investigate the inflow-diversity of inter-industry labour flows, we assign each firm to an industry (given by an industry code) according to its primary economic activity. The industry classification system was internally constructed by SARS and has 388 different 4-digit industry codes.

In this study, we divide all firms into two groups according to whether a firm is exempt or is expected to comply with the EE Act. Recall that the act only applies to firms with a workforce larger than 50, except if the firm is within an exempt industry. However, firms with annual revenue above a certain threshold must comply with the act even if their workforce is below 50. All firms and their workforce exempt from the act are our control (exempt) group, and all other firms and their workforce are our treatment (compliant) group. We illustrate the division of our dataset into these two groups in Figure 6.1. We have 95156 firms within our exempt group and 17138 firms within our compliant group. Although the number of firms is smaller within the compliant group, they contain many more employees.

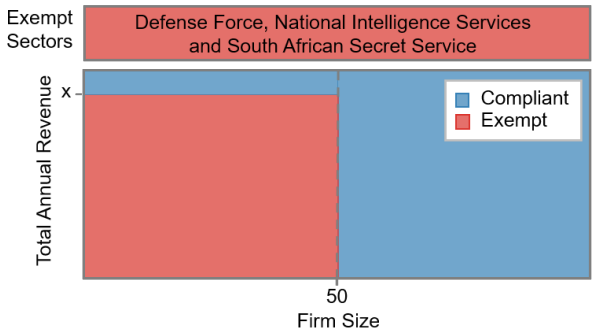


Figure 6.1: An illustration of the division of our firm dataset into a treatment (compliant) and control (exempt) group. All firms with more than 50 employees or earning an annual revenue above a certain threshold of x are obliged to comply with the act. The threshold x is industry specific.

6.4 Network and metric construction

6.4.1 Network construction

To analyse the impact of the EE Act on labour inflow diversity, we construct three networks, namely: the inter-firm labour flow network (LFN), the inter-industry LFN and the skill-relatedness network (introduced in the next section). The differences between these networks are summarised in Table 6.1.

Table 6.1: Properties of the three different networks constructed in this study.

Network	Nodes represent	Edges represent	Size of network	Use of network in study
Inter-firm LFN	firm	Average number of worker transitions between firms per year	122 294	Quantifying the diversity of labour inflows for firms
Inter-industry LFN	industry	Average number of worker transitions between industries per year	388	Quantifying the diversity of labour inflows for industries
SRN	industry	Skill overlap between industries	388	Constructing a new industry classification

The first two networks, the inter-firm and inter-industry LFN, are used in this study to quantify the labour inflow diversity (e.g., the diversity of sectors from which workers are hired) for a firm and industry, respectively. The inter-firm LFN is a network in which each node represents a firm and edges the number of workers transitioning between the two corresponding firms. For the inter-industry LFN, each node represents an industry and edges the number of workers transitioning between the two corresponding industries.

Formally, let $L_S(i, j, t)$ denote the observed labour flows from firm i to firm j between years t and $t + 1$. We define the positive and non-symmetric adjacency matrix $A_S(t)$ for the inter-firm LFN as $A_S(i, j, t) = L_S(i, j, t)$. Similarly, let $L_I(i, j, t)$ denote the observed labour flows from industry i to industry j between years t and $t + 1$. We define the positive and non-symmetric adjacency matrix $A_I(t)$ for the inter-industry LFN as $A_I(i, j, t) = L_I(i, j, t)$. Both networks are weighted, directed graphs.

We use various subsets of these flows to build inter-firm and inter-industry LFN variations. For example, we include flows depending on whether the workers are male or female, whether they transition to firms that either comply or are exempt from the EE Act and whether they are skilled employees. The adjacency matrices of these networks are denoted as $A_\alpha^{\beta,\gamma,\theta}$, where

- $\alpha \in \{S, I\}$ indicates whether we are constructing an inter-firm or inter-industry LFN,
- $\beta \in \{M, F\}$ indicates whether we are considering male or female workers,
- $\gamma \in \{C, E\}$ indicates whether workers who transition to either compliant or exempt firms are used, and
- $\theta \in AW, SW$ indicates whether we consider all workers or only skilled workers.

For example, $A_S^{F,C,SW}$ indicates an inter-firm LFN which includes the transitions of skilled female workers who move to firms that comply with the Act.

6.4.2 Constructing a new industry classification

To evaluate whether the EE Act has prompted firms to hire new workers from a more diverse set of prior sectoral backgrounds, we need to construct an inflow diversity measure. Recall that the inflow diversity measures the array of different sectors, defined as a group of industries that share a high degree of skill and knowledge, from which a firm receives new workers. A firm with a high inflow diversity will have hired workers with a wide variety of skills. Constructing this measure consists of two steps. First, we need to construct a new industry classification due to the limitations of the current SA internal industry classification scheme. Specifically, we group industries into clusters or sectors with similar skills. We can then measure the diversity of inflow sectors, defined according to our new industry classification, as discussed in the following section.

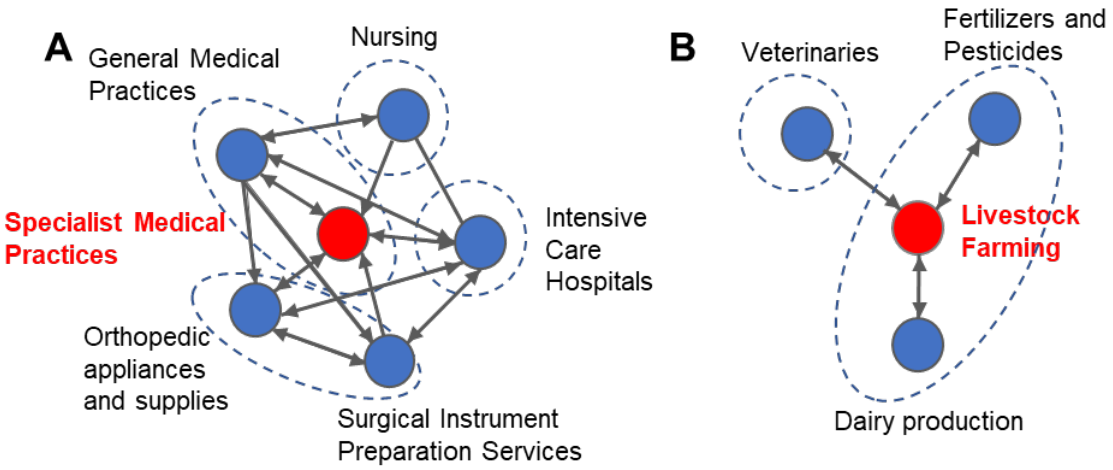


Figure 6.2: (A&B) Shows two toy networks, where nodes represent industries and edges represent the presence of a high level of labour flow between the two corresponding industries. In (A) we show the Specialist Medical Practices industry and in (B) the Livestock Farming industries, as well as the industries from which they hire new workers, respectively. We observe that the industries are defined in different levels of granularity, with industries in B more granular than in A. This makes the count of industries from which they receive workers uninformative as a measure of skill-diversity. Furthermore, the industries are grouped into sectors as shown by the blue dotted lines. Here we observe, that these groupings do not coincide with the degree of labour flows between them. Industries grouped together do not always have corresponding high labour flows. This makes the current classification not a good indicator of skill-diversity.

Various limitations associated with the current industry classification prevent us from using its predefined 4-digit, 2-digit or 1-digit classification as different sectors within our analysis. First, we cannot purely use the 4-digit industries as these industries are defined at different levels of granularity. Grouping industries into sectors that represent similar levels of granularity can solve this. However, the current 2-digit (or 1-digit) classification assigns industries according to their primary economic activity. This results in industries being grouped together even when they share no skills or knowledge. Therefore, using these sectors to calculate the sectoral diversity would not be a good indication of the skill diversity of workers.

We illustrate these limitations in Figure 6.2. First, to see that industries are defined at different levels of granularity, consider the inflow of workers into the *Specialist Medical Practices* industry and the *Livestock Farming* industry shown in red in (A) and (B), respectively. The Specialist Medical Practices industry receives workers from 5 different industries, compared to Livestock Farming which only receives workers from 3. However, when examining the industries more carefully, we can see that the first industry's neighbouring nodes are defined in much more detail than the second. This makes considering the diversity of 4-digit industries from which workers

flow uninformative.

Next, we observe how the neighbouring industries have been grouped into different sectors (by the SARS administrative industry classification). Dotted lines in A and B illustrate the sectoral groupings for both these industries. We see that the neighbouring industries of Specialist Medical Practices (shown in A) can be classified into four different sectors, while the neighbouring industries of Livestock Farming (shown in B) are classified into two different sectors. However, in A, these sectors have a high overlap of knowledge and skills shown by the edges connecting them and representing high degrees of labour flows between these industries. In contrast, in B, we find no skill overlap amongst any neighbouring industries, not even those in the same sector. Therefore, by looking at the degree of skill overlap between the neighbouring industries, we can conclude that Specialist Medical Practices are, in fact, less diverse than Livestock Farming. Therefore, we need to consider both the degree of skill overlap between industries and their granularity level when grouping them into the same industry cluster or sector.

In order to overcome these problems we construct a new industry classification. To do this, we construct a third network: the skill-relatedness network (SRN) [149]. In this case, each node represents an industry and each edge is a measure of the skill similarity or overlap between the two corresponding industries. This is calculated by comparing the number of worker transitions between two industries to what we would expect at random (under a standard configuration model [142]). The reader is referred to §2.5 for a detailed description of the SRN construction. We do however highlight, that we construct the skill-relatedness matrix (A_{SR}) for each year (2011-2014) and then find its average, as this allows for the most accurate skill-relatedness value to be obtained.

We illustrate the skill-relatedness network in Figure 6.3 (A). In this figure, each node represents an industry, each node is sized by the average employment over the full time period, and the weight of each edge its skill-relatedness. The node layout used is generated by a spring algorithm called 'Force Atlas' in Gephi, which positions industries connected by larger edge weights closer together. We have manually added labels indicating the general position of sectors to the figure.

In order to extract meaningful groupings of 4-digit industries based on skill overlap, which we will use as a higher level sectoral classification, we apply a community detection algorithm to the SRN. This approach was previously deployed by O'Clery *et al.* [161] to unveil communities (which can be interpreted as skill-basins or labour pools) within the Irish SRN. More specifically, we use the Markov Stability algorithm

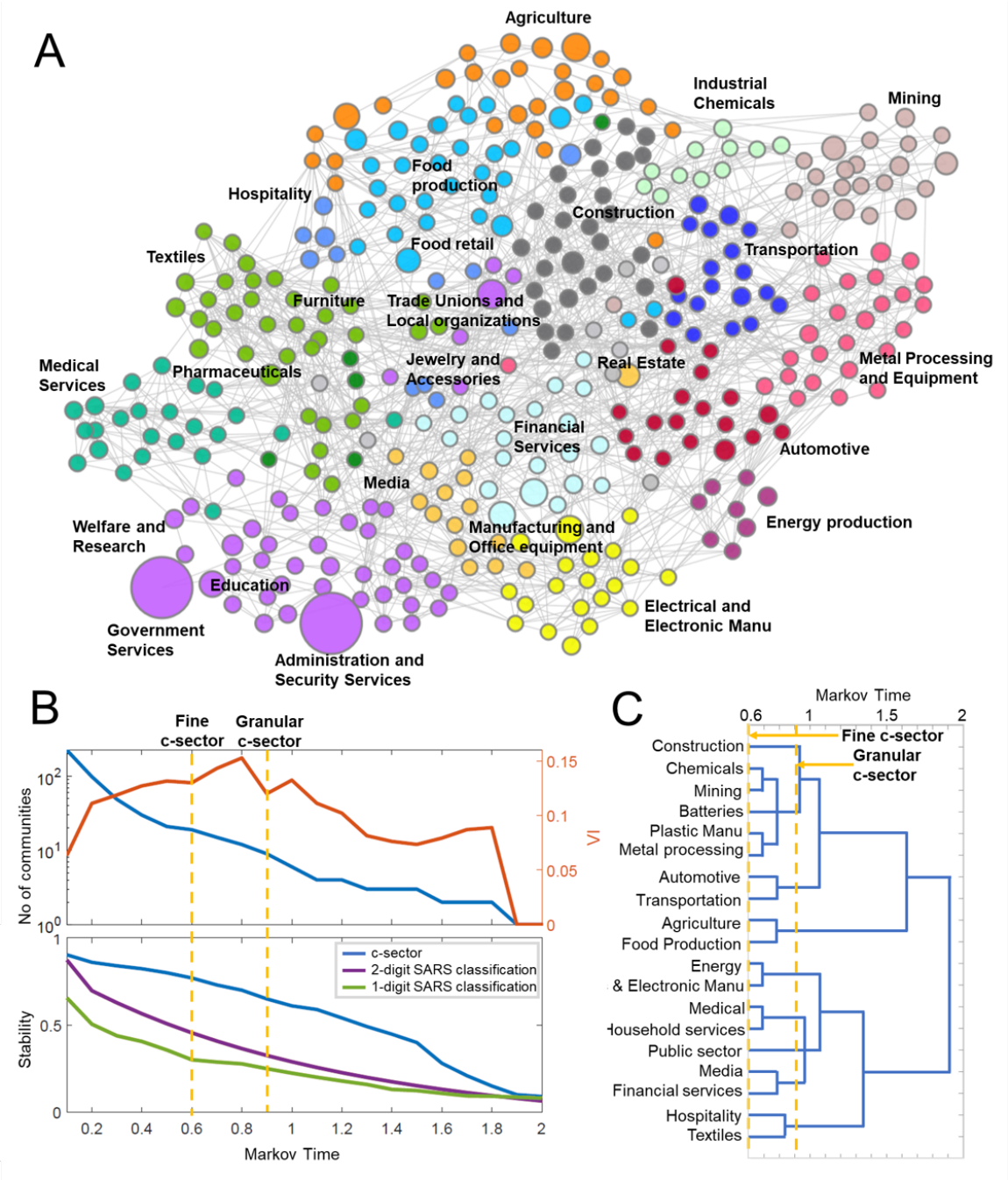


Figure 6.3: Visualisation of the skill-relatedness network for South Africa. Each node represents an industry and each edge the skill-relatedness between the corresponding industry pair. Only edges above a skill-related index of 0.6 are shown. Nodes are sized by the average employment over the $2011 - 2014$ time period. The node layout is based on the 'Force Atlas' spring algorithm, whereby more skill-related industries are positioned closer together. Nodes are coloured according to their c-sector which is chosen as our new industry classification.

[61] as our community detection algorithm. This algorithm is a dynamical community detection algorithm based on random walker diffusion dynamics. The algorithm aims to maximise the probability that a random walker remains in the community in which it started during a time interval. Intuitively, if a random walker is allowed to jump from node to node, then if the walker gets trapped in a group of nodes for an extended period it indicates a region of high connectivity corresponding to a node community. For a detailed explanation of the workings of the Markov Stability algorithm, the reader is referred to § 2.4.2.

Markov stability is a multi-resolution community detection algorithm. We, therefore, obtain a range of partitions that group industries into clusters at different scales (from many small clusters to only a few large clusters). We show the number of communities, the variation of information and the stability of each resulting partition at different Markov times in Figure 6.3 (B). We choose two robust partitions (which display a low variation of information), which we mark in yellow, and which well describe the network's modular structure. The first partition contains 19 communities and is also illustrated in Figure 6.3 (A) by node colouring. The second contains nine communities. We refer to the former partition as the fine c-sectors and the latter as the granular c-sectors. We use these two partitions to group industries at different scales. We also show the hierarchical structure of our new c-sectors through a dendrogram in Figure 6.3 (C)[5].

Our grouping of industries (both in our fine and granular c-sectors) differs from that of the official sectors in the 2-digit (and 1-digit) SARS industry classification. This is because the c-sectors classify industries according to skill and knowledge overlap, while the administrative sectors group industries according to their primary economic activity. We compare the stability obtained using our new measures with those of the SARS industry classification in subplot (B). We observe that we obtain higher stability using our c-sectors. Hence, our sectors are better representations of the network's modular structure. Furthermore, our c-sector classification also rectifies the heterogeneity of the SARS industry classification scheme by allowing industry clusters to be of varying sizes.

6.4.3 Constructing the various inflow diversity measures

To evaluate the impact of the EE Act on labour mobility we need to construct various inflow diversity measures (*e.g.*, the array of different sectors, as defined by our c-

[5]Note that the network partitions found using the Markov Stability algorithm are not naturally hierarchical. We use a majority rule to assign smaller communities to larger ones.

sectors above, from which a firm or industry hires workers). We evaluate three aspects of the inflow diversity, namely: the total inflow variety of all workers ($Inflow\_V$), the related inflow variety ($Inflow\_RV$) and the unrelated inflow variety ($Inflow\_UV$) of skilled workers.

6.4.3.1 Total inflow variety

The *total inflow variety* ($Inflow\_V$) measures the array of different fine c-sectors from which a firm hires new workers. A firm with a high inflow variety will therefore have hired workers from many different sectors with a wide variety of skills.

To quantify the total inflow variety for either a firm or an industry we need to take into account both the range of sectors a firm or industry's workforce flows from, as well as the number of workers that flow from each, we therefore construct an entropy measure.

We let $Inflow\_V(i, A)$ be the inflow variety of node i in a network represented by adjacency matrix A. Let the fine c-sector to which node i belongs be defined as $\text{fsec}(i)$. Furthermore, we let the 4-digit industry to which node i belongs be denoted as $\text{ind}(i)$[6]. Now, let $e^{fsec}(j, i)$ represent the number of workers who flow from fine c-sector j to node i, so that $e^{fsec}(j, i) = \sum_k^N A(k, i)\delta(\text{fsec}(k), j)(1 - \delta(\text{ind}(k), \text{ind}(i)))$ where δ is the Kronecker delta. Note that we only consider inter-industry labour flows (even in our firm-level network), we therefore do not count for labour flows within each 4-digit industry accounted for by the term $(1 - \delta(\text{ind}(k), \text{ind}(i)))$. We then define,

$$p^{fsec}(j, i, A) = \frac{e^{fsec}(j, i)}{\sum_{j=1}^{19} e^{fsec}(j, i)}.$$

This represents the fraction of node i's inter-industry worker inflows which come from fine c-sector j. The inflow diversity is then given as

$$Inflow\_V(i, A) = -\sum_{j=1}^{19} p^{fsec}(j, i, A) \log(p^{fsec}(j, i, A)). \tag{6.1}$$

Therefore, the larger the array of different sectors, as well as the larger the number of workers who transition from each fine c-sector to node i, the higher the inflow variety of node i. To obtain the male and female, exempt and compliant inflow varieties we use the corresponding adjacency matrices in our measure.

[6]Note that in the industry network $\text{ind}(i) = i$

6.4.3.2 Related inflow variety

The related inflow variety ($Inflow\_RV$) measures the array of different sectors, which are related but not similar to a firms own industry, from which a firm hires new skilled workers. As noted in Section 6.2.2, we assume that the more related, but not identical, the set of skills are that are brought in by newly hired workers, the greater the opportunities for learning and innovation within a firm, and thereby the more the firm will benefit from the array of skills and knowledge.

To measure the related variety a weighted sum entropy of the industries (defined by their 4-digit industry code) within each fine c-sectors is calculated. This measure was first introduced by Frenken *et al.* [84]. We assume that new employees from the same fine c-sector as the firm have related skills and are best able to understand each other, but they require variety at the industry level to induce real learning and knowledge spillovers.

Formally, the related variety is calculated as follows. Recall, that the 4-digit industry of node i is given by $\text{ind}(i)$. Similarly, the number of workers who flow from industry j to node i is given as $e^{ind}(j,i) = \sum_k^N A(k,i)\delta(\text{ind}(k),j)(1-\delta(\text{ind}(k),\text{ind}(i)))$. In accordance with the literature, we again only account for inter-industry labour flows and not labour flows within each industry. The fraction of node i's inter-industry worker inflows which come from industry j is then also defined as $p^{ind}(j,i,A) = \frac{e^{ind}(j,i)}{\sum_{j=1}^{388} e^{ind}(j,i)}$.

Now, each industry belongs to a fine c-sector. Therefore, we can also derive the fine c-sector's share of $p^{fsec}(k,i,A)$ by summing all the 4-digit industry shares $p^{ind}(j,i,A)$ belonging to $\text{fsec}(k)$:

$$p^{fsec}(k,i,A) = \sum_{j \in \text{fsec}(k)} p^{ind}(j,i,A). \tag{6.2}$$

Thereby, the related variety is given as:

$$Inflow\_RV(i,A) = \sum_{k=1}^{19} p^{fsec}(k,i,A)H(k) \tag{6.3}$$

where:

$$H(k) = -\sum_{j \in \text{ind}(k)} \frac{p^{ind}(j,i,A)}{p^{fsec}(k,i,A)} \log_2\left(\frac{p^{ind}(j,i,A)}{p^{fsec}(k,i,A)}\right). \tag{6.4}$$

Note that the RV indicates how evenly the degree of inflows are spread across the industries within the node's fine c-sector. The value of the $Inflow\_RV$ can differ from 0 (when all inflows from the industry's fine c-sector comes only from one single industry) to a theoretical upper bound of $\log_2(388) - \log_2(19)$, when an equal number

of workers arrive from all the industries in the c-sector. Note that the higher the $Inflow\_RV$ value, the more evenly employment is spread across the industries in the fine c-sector and therefore the higher the diversity of related sectoral backgrounds.

6.4.3.3 Unrelated inflow variety

The unrelated inflow variety ($Inflow\_UV$) measure the array of different sectors which are completely unrelated to the sector of the firm from which a firm hires new workers. This measure was also first introduced by Frenken *et al.* [84]. This variable captures the degree of unrelated competencies that newly hired workers bring to a firm. As noted in the literature, very different competencies hinder interactive learning processes due to communication problems between workers. Therefore these newly hired inflows will have negative effect on firm performance.

The degree of unrelated inflows is measures as the entropy at the granular c-sector level. More formally, let the number of workers who flow from granular sector j to node i be given as $e^{gsec}(j,i) = \sum_k^N A(k,i)\delta(\text{gsec}(k),j)(1-\delta(\text{ind}(k),\text{ind}(i)))$. Now, the fraction of node i's inter-industry worker inflows which come from granular c-sector j is given as $p^{gsec}(j,i,A) = \frac{e^{gsec}(j,i)}{\sum_{j=1}^{9} e^{gsec}(j,i)}$. Then,

$$Inflow\_UV(i,A) = -\sum_{j=1}^{9} p^{gsec}(j,i,A)\log_2(p^{gsec}(j,i,A)). \tag{6.5}$$

Again, this measure can vary from 0 where all unrelated employees are hired from a single granular c-sector to $\log_2(9)$ when an even number of new unrelated employees are hired from each of the differ granular c-sectors. Again, the higher the $Inflow\_UV$, the more diverse the backgrounds that are unrelated to that of the firm or industry.

6.5 Methodology

6.5.1 Firm-level regression model

Before investigating the act's impact on labour mobility, we first investigate if there is a relationship between the sectoral diversity of newly hired workers and firm performance. To do this, we follow the same methodology of Boschma *et al.* [30] who investigated this relationship for plants in Sweden.

Following the literature, we focus only on skilled labour flows and the firms that have hired skilled workers. This is because skilled employees can contribute new

knowledge and skills to a firm where they can be recombined to create new economic value.

We use an ordinary least square model to investigate the relationship between the related and unrelated variety of skilled labour flows and firms' labour productivity growth. We measure firm performance by using firms' labour productivity growth over a two-year period. We measure labour productivity as the turnover per employee [32]. This is because it is the most straightforward measure of labour productivity and is available for all firms in our dataset. We expect that the effects of the sectoral diversity of newly hired employees will only materialize at the firm level after some years. This is why we use the productivity growth only after two years as the dependent variable. All our independent variables (the inflow related variety, inflow unrelated variety and firm size) are measured at the beginning of the observed period.

Formally, our OLS model is given as:

$$\begin{aligned}\log(\text{Change in Labour Productivity}(i, t+2)) =& a + b_1 \log(Inflow\_RV(i, A_S^\gamma)) \\ &+ b_2 \log(Inflow\_UV(i, A_S^\gamma)) \\ &+ b_3 \log(Emp(i,t)) + \rho(i) + \tau(t) + \epsilon(i,t),\end{aligned} \tag{6.6}$$

where $\rho(i)$ and $\tau(t)$ are industry and time fixed effects, respectively. Through these fixed effects, we control for within-industry and within-year variance. Furthermore, $Emp(i,t)$ indicates the size of firm i in the base year t. These control variables which are often co-explaining determinants of productivity are therefore accounted for. Note that we run this OLS regression for firms who comply with the act and those that are exempt ($\gamma \in C, E$). Furthermore, we pool together all observations in the first two time periods (2011 and 2012).

6.5.2 Firm-level regression discontinuity design

Next, we adopt a regression discontinuity (RD) design to determine the impact of the EE Act on the sectoral diversity of newly-hired male and female workers. This enables us to exploit the discontinuity between the requirements prescribed by the act according to firm size.

Our approach relies on the hypothesis that firms slightly below and slightly above the cut-off have similar characteristics. This provides a unique opportunity to compare compliant and exempt firms within the cut-off neighbourhood while controlling for confounding factors. Hence, differences in the outcome variable will show the act's impact.

Specifically, within our RD design, our independent variable is the size of employment within a firm. Following the legislation of the EE Act, firms with more than 50 employees were compelled to comply with the act, while those with less than 50 employees are exempt from the act. Note that firms who either have less than 50 employees but still abide by the act as their annual revenue is above the threshold value or belong to a non-compliant industry are removed from our sample. The key feature of the design is that the probability of complying with the act changes abruptly, from 0 to 1, at the 50-employee cut-off value. The discontinuous change in this probability is used to learn about the local effect of the act on the outcome variables: a firm i's male and female total inflow diversity, $Inflow\_V(i, A^{\beta})$, (where $\beta \in \{M, F\}$). Firms that are just below the cut-off (have a firm size of 49) are compared to firms just above the cut-off (have a firm size of 51). As the size of these firms is very similar, it is expected that the outcome variable should be very similar without the act. Hence, a significant discontinuity between the groups' outcome variable indicates the act's impact.

To estimate the discontinuity, we fit a polynomial function to the data on each side of the cut-off. We adopt a local linear polynomial approach and choose a polynomial of order 1. We do not choose a higher-order polynomial, as these polynomials provide poor approximations at boundary points [7]. However, the issue with choosing a lower-order polynomial is that the size of the neighbourhood (the interval) considered within the analysis heavily influences the result. We adopt a bandwidth of 10. We choose ten, as firms within this interval (specifically, those with 50-60 employees) fall under the same category under the act. To ensure that our fit is less sensitive the size of the neighbourhood, as well as over-fitting boundary problems (variance of the local linear polynomial estimator), we adopt a triangular kernel function which assigns non-negative weights to each observation based on its distance to the cut-off. The largest weight is assigned to the values at the cut-off. The weight then symmetrically and linearly decreases as we move away from the cut-off [42].

More formally, the independent variable x_i is equal to the firm i's size ($Emp(i, t)$). The dependent variable is denoted as y_i, which is equal to the male or female inflow diversity ($Inflow_V(i, A^{\beta}(t))$). We again pool all observations across our time periods together (t= 2011-2014). The cutoff value is chosen at $c = 50$ and the bandwidth denoted as $h = 10$. The implementation of our discontinuity regression approach is

[7] For a detailed explanation of how higher-order polynomials provide poor approximation at boundary points, the reader is referred to the Runge phenomenon in approximation theory [199])

based on estimating the following linear regression:

$$y_i = \begin{cases} a_L + b_L x_i + \varepsilon_L, \text{ if } (c-h) \leq x_i < c, \\ a_R + b_R x_i + \varepsilon_R, \text{ if } c \leq x_i \leq (c+h). \end{cases} \tag{6.7}$$

Therefore a different slope and intercept are found when fitting data on each side of the cutoff. We adopt the triangular kernel function, $K(\boldsymbol{u}) = (1 - |\boldsymbol{u}|)\mathbf{1}(|\boldsymbol{u}| \leq 1)$, which assigns non-negative weights to each transformed observation $u_i = (x_i - c)/h$. The weight is maximized at $x_i = c$, and declines symmetrically and linearly as the value of x gets farther from the cutoff. Finally, the RD treatment effect is calculated by determining the vertical distance at the cutoff point:

$$\tau_{RD} = \lim_{x \xrightarrow{R} c} \mathbf{E}[y_i | x_i = c] - \lim_{x \xrightarrow{L} c} \mathbf{E}[y_i | x_i = c]. \tag{6.8}$$

One of the disadvantages of this method is that it is valid only with respect to the mean impact of the act around the cutoff. In other words, an RD design is not a valid measure for capturing treatment effects that occur for units away from the cutoff value. It is, therefore, unable to predict the overall relationship between the two variables [42].

We apply this methodology using the male and female inflow variety ($Inflow\_V$) of all workers and the male and female related- ($Inflow\_RV$) and unrelated inflow variety ($Inflow\_UV$) of skilled workers within this study.

6.5.3 Industry-level regression model

Next, we are interested in investigating how the act impacts different industries. We hypothesise that male-dominant industries are associated with a higher female inflow diversity. This is because these industries require the most significant workforce restructuring and have the lowest availability of female workers within their sectors. They, therefore, need to recruit female workers from other related industries to meet the act's obligations.

We, therefore, investigate the relationship between the fraction of male employment and the male and female inflow diversity of an industry. We first evaluate this relationship for workers within firms that are compliant with the act. Then we evaluate it for workers within firms that are exempt from the act. These two relationships are then compared. We now elaborate on this setup.

We show our multivariate linear regression model for the case in which the impact of the fraction of male employment FME is compared to its female inflow variety

$Inflow\_V(A_I^F)$ of an industry. First, when considering firms that comply with the act, the model is given by

$$\begin{aligned} Inflow\_V(A_I^{F,C}(t)) =& \alpha + \beta_1 FME_I^C(t) + \beta_2(Emp_I^{F,C}(t)) + \\ & \beta_3(Emp_I^{F,C}(t+2) - Emp_I^{F,C}(t)) + \beta_4 Wage_I^{F,C}(t) + \epsilon. \end{aligned} \tag{6.9}$$

We control for the average employment size, the employment growth over 2 years and the average wage of each industry within the model. We run this regression for all industries in $t = 2011$ and $t = 2012$ together. Secondly, when considering firms that are exempt from the act, a similar model is given by

$$\begin{aligned} Inflow\_V(A_I^{F,E}(t)) =& a + b_1 FME_I^E(t) + b_2(Emp_I^{F,E}(t)) + \\ & b_3(Emp_I^{F,E}(t+2) - Emp_I^{F,E}(t)) + b_4 Wage_I^{F,E}(t) + \epsilon. \end{aligned} \tag{6.10}$$

We again, run this regression for all observations in $t = 2011$ and $t = 2012$ together. These two regression models are then compared. Finally, we repeat this regression analysis to investigate the impact of the fraction of male employees on male inflow diversity.

6.6 Results

6.6.1 The impact of labour inflow diversity on firm performance

Before investigating how the act influences labour mobility, we first investigate whether inter-industry labour mobility affects the performance of firms. Recall that according to the literature, the inflow of skills that are related to the existing knowledge base of the firm has a positive effect on firm performance, while the inflow of new employees with skills that are unrelated to those already present in the firm has either no effect or a negative effect [30]. We investigate whether this is true in the South African case.

Following the same methodology as Boschma *et al.* [30] we investigate how the sectoral diversity of newly hired employees influences firms' labour productivity. First, we extract the inter-industry labour flows of *skilled* employees and the corresponding firms where they were hired. Our data-set now consists of 11 880 firms. Next, we determine the effects of the total number of workers who are hired from other industries, as well as the related- and unrelated variety of these in-flowing workers on the change in labour productivity of firms (measured over a two-year period).

	Compliant Firms		Exempt Firms	
Log(Change in Labour Productivity)	(1)	(2)	(3)	(4)
Log(Inter-industry inflows)	-0.0517*		-0.0963*	
	(0.0699)		(0.7876)	
Log(RV)		0.1921***		0.2084**
		(0.1107)		(0.1168)
Log(URV)		-0.0479		-0.0592
		(0.1289)		(0.1289)
Log(Firm Size)	-0.6130***	-0.6023***	-0.7853***	-0.8173***
	(0.1009)	(0.1058)	(0.1082)	(0.1129)
Constant	11.2864**	11.5163**	10.7912**	10.6486**
	(1.8216)	(1.8962)	(1.9372)	(1.9014)
Industry FE	Y	Y	Y	Y
Time FE	Y	Y	Y	Y
N	8 586	8 586	3 294	3 294
R^2	0.4022	0.4127	0.3794	0.3840

Table 6.2: Table showing an OLS model in which we investigate the association between skilled labour mobility and a change in firms' labour productivity both for firm who exempt and those who comply with the Employment Equity Act.

In Table 6.2, using the OLS model presented in §6.5.1, we assess whether the inflow of workers from different industries and their sectoral diversity is associated with firm labour productivity growth. In models (1) and (3), we observe that for both compliant and exempt firms, the total inter-industry inflows of newly hired employees have a negative and weakly-significant (only to a 90% confidence interval) relationship with the change in labour productivity growth. This confirms the stance of the literature that not all labour mobility *per se* positively influences firm performance. It is, therefore, crucial to differentiate between different types of skills when assessing the effect of labour mobility on firm performance - this is as the type of skill that flows into firm matters. We also observe that firm size has the greatest effect on productivity growth with a negative and significant relationship. Hence, small firms show higher levels of productivity growth than larger ones. This outcome is in line with the literature [30].

We now consider the type of skills that flow into the firms separately. In models (2) and (4), we find that the inflow of related workers is significantly and positively associated with increased labour productivity. Hence the inflow of new skills should be related to but not similar to the firm's existing knowledge base to have a positive economic effect. In contrast, we do not find a significant relationship between the change in a firm's labour productivity for the inflow of unrelated skills. Our model suggests that the cognitive distance between the knowledge and skills brought by the new employees (previously employed in different sectors to that of the firm) is too cognitively distant. Therefore the firm cannot effectively absorb the incoming

knowledge and benefit from it.

Using this as a motivation on why it is crucial to understand changes in the sectoral diversity of newly hired workers, we next investigate the act's impact on labour inflows.

6.6.2 Impact of the EE Act on firm labour inflow diversity

First we investigate the act's impact on the total variety of newly hired male and female employees within firms. Here, we again consider all inter-industry labour flows and not only those associated with skilled employees. We apply an RD design taking both the male and female inflow variety as the dependent variable in Eq.(6.7). The results are illustrated in Figure 6.4 (A) and (B), respectively. We observe that for both male and female workers, as the size of a firm increase, the inflow diversity also increases. This is because as a firm increases in size, it typically hires workers from more sectors in order to operate different divisions within the firm [163]. We are not primarily interested in this relationship, but the presence of a discontinuity in the relationship around the cutoff value of 50 employees. We find a positive discontinuity for the female inflow diversity at the cutoff value of 50, with treatment effect of 0.076 with a 95% confidence interval of $[0.053, 0.102]$. A small and negative discontinuity is found for the male case, but with an insignificant treatment effect. This shows that, for firms of around 50 employees, those that comply with EE Act have a larger inflow diversity of female workers than those who are exempt of the act. Although we can't as easily extrapolate our treatment effect for firms of different size, our results provide some evidence that the act is causing firms to hire female workers from more diverse industries. This may potentially be to enhancing their female representation.

As discussed in the previous subsection, not all labour mobility leads to positive economic impact, and the type of skills that flow into a firm matters. Therefore, we next investigate the impact of the act on gendered related- and unrelated variety of skilled labour inflows. We now only use the subset of skilled labour flows and the firm's to which they flow.

We apply an RD design taking the male and female related- and unrelated-variety of inflowing labour as the dependent variable in Eq. (6.7). The results are illustrated in Figure 6.5 (A) to (D), respectively. Focusing first on the related variety, we find a positive discontinuity for the female related variety at the cutoff, with a significant treatment effect of 0.008. However, no significant effect is observed for the male related variety. Secondly, for the unrelated variety of both male and female newly hired employees no significant discontinuity or treatment effect is observed. Our

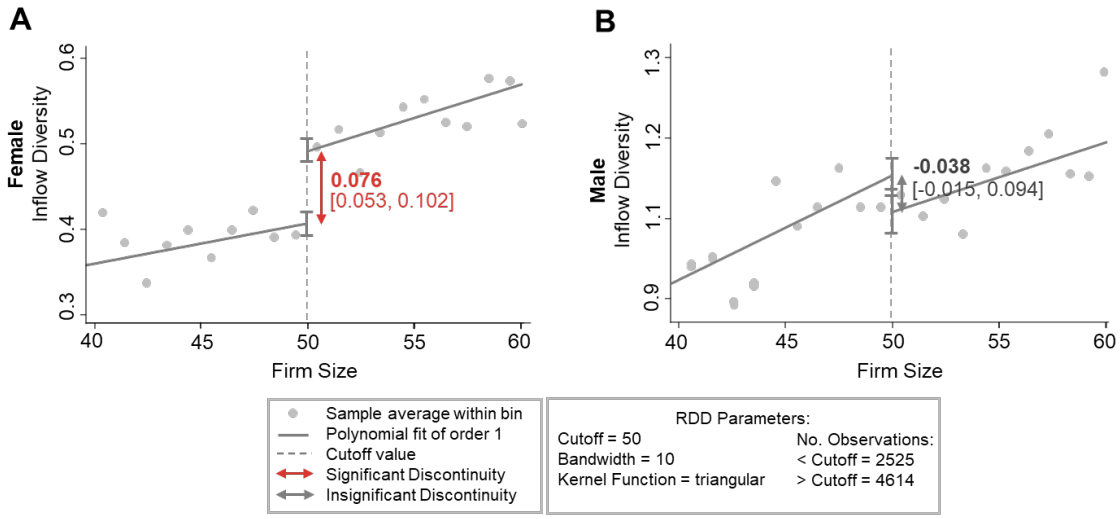


Figure 6.4: Regression discontinuity plots showing the impact of the EE Act for firms of size 50 on the (A) Female inflow variety and (B) Male inflow variety. Note that the RD design parameters are shown in the legend. Data points are binned for visualization purposes, with the exact number of observations used shown in the legend. The treatment effect, as well as its 95% confidence interval, is shown on each graph. The colour indicates non-overlapping intervals which is equivalent to a significant treatment effect.

results indicate, that the act has increased the related variety of newly hired skilled female employees for firms of size 50. However, this is not the case for the male related variety or the unrelated variety of both male and female employees.

So far, we have observed that the act enhanced the related variety of skilled female workers, and from the previous OLS model results that the general increase in related variety of workers is associated with an increase in labour productivity for firms. We therefore now investigate whether the act, potentially through enhancing the related variety of female inflows, has influenced the change in labour productivity of firms.

We again apply an RD design taking the change in firms' labour productivity as the dependent variable in Eq. (6.7). We use our skilled data-set and thereby include all firms that have hired skilled workers from other industries. The results are shown in Figure 6.6. Note that we do not find a significant discontinuity or treatment effect of the act on the firm's change in labour productivity. This suggests that the act-induced increase in female related variety did not as easily translate into increased labour productivity of firms. This could be that an organic increase in related variety is associated with firm growth, while the act-induced increased in related variety is primarily to meet quotas and therefore does not show an effect on the growth of labour productivity. This could be a potential reason for why the increased labour mobility does not show an effect on firm performance.

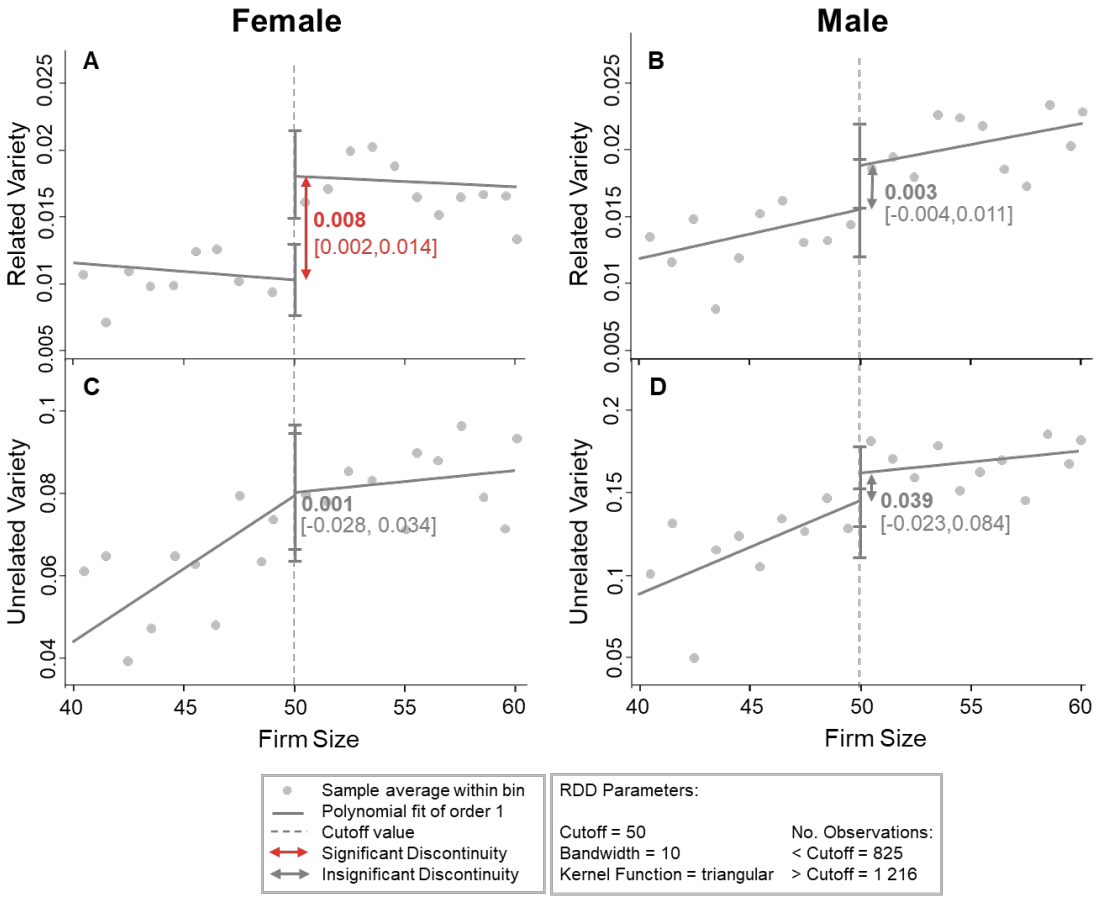


Figure 6.5: Regression discontinuity plots showing the impact of the EE Act for firms of size 50 on the (A) Female related and unrelated inflow diversity and (B) Male related and unrelated inflow diversity. Note that the RD design parameters are shown in the legend. Data points are binned for visualization purposes, with the exact number of observations used shown in the legend. The treatment effect, as well as its 95% confidence interval, is shown on each graph. The colour indicates non-overlapping intervals which is equivalent to a significant treatment effect.

In summary, our results show that the act increased the inflow variety of newly hired female workers for firms with approximately 50 employees. This shows that the act has caused firms to hire female workers from more diverse sectors. Furthermore, we also find that the act increased the related variety of newly hired skilled female workers, also for firms of the same size. Both from the literature and our OLS model, we know there is a positive relationship between related variety of newly hired workers and the growth of labour productivity in firms. However, we do not find a direct effect of the act on the labour productivity growth of firms. This could be that the artificial increase in the related variety of skilled female inflows by the act does not as easily translate into increased labour productivity, or that it may take a longer period to show effect.

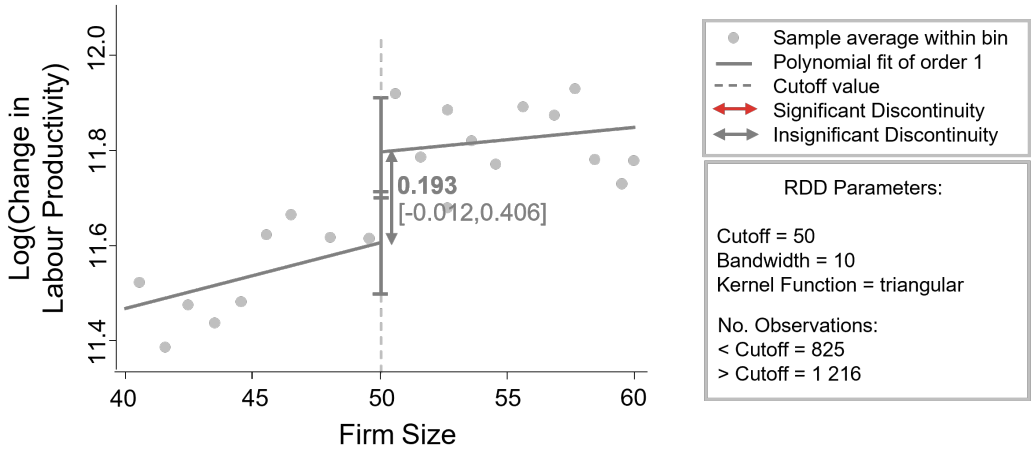


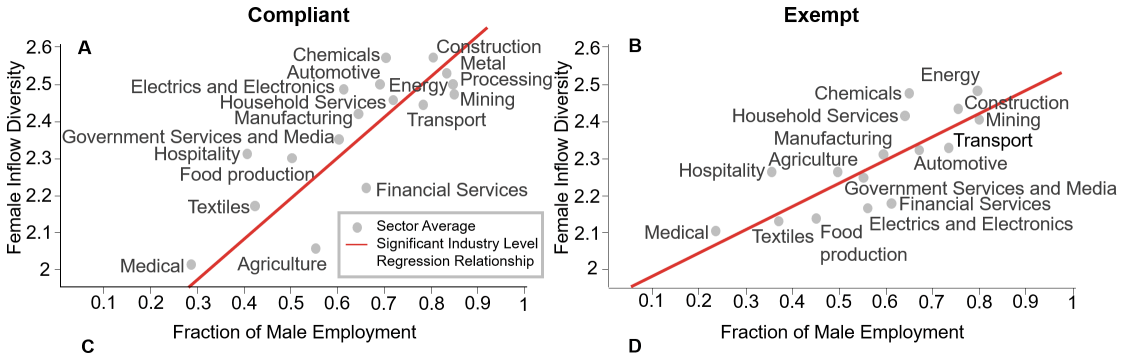
Figure 6.6: Regression discontinuity plot showing the impact of the EE Act for firms of size 50 on the change in labour productivity for all firms. Note that the RD design parameters are shown in the legend. Data points are binned for visualization purposes, with the exact number of observations used shown in the legend. The treatment effect, as well as its 95% confidence interval, is shown on each graph. The colour indicates non-overlapping intervals which is equivalent to a significant treatment effect.

6.6.3 Impact of the EE Act on industry labour inflow diversity

We are now interested in investigating if there is an association between the female inflow diversity of an industry and the average percentage of male workers within the industry. We previously hypothesised that male-dominant industries will be associated with a larger female inflow variety, as these industries have to hire more female workers from outside of their industries (as fewer female workers are available within their industry) to reach their quotas.

First, we compute each industry's male and female inflow diversity using Eq. (6.9). Note that we are now calculating the inflow diversity at an industry level. The relationship between the percentage of male employment and the female inflow diversity of an industry is shown in Figure 6.7[8]. We observe a positive and significant relationship, which is particularly pronounced in the case of the compliant group with a regression coefficient of 0.6936, compared to the exempt group with a regression coefficient of 0.2467. The difference between these coefficients (0.4469) is larger than the standard error of both the coefficients (0.0941 and 0.0864, respectively), and hence these relationships are significantly different. These results show that male-dominant

[8]Within the figure, industries are binned according to their fine c-sectors purely for visualisation purposes. However, the trend line, as well as the resulting regression tables quantifying the relationship and shown in Figure 6.7 (C) and (D), is calculated using the 388 individual industries.



C

Dependent Variable: Female Inflow Diversity	(1)	(2)
Fraction of Male Employment	0.6156***	0.6936***
	(5.4698)	(7.1579)
Log(Female Employment Size)		0.1001***
		(8.8236)
Change in Female Employment Size		-8.5574e-06**
		(-3.0082)
Female Average Wage		8.5355e-07**
		(3.1579)
Constant	1.7617***	1.7013***
	(27.361)	(22.458)
Observations	386	386
R-squared	0.2723	0.3672

Robust t-statistic in parentheses
*** p<0.01, ** p<0.05, * p<0.1

D

Dependent Variable: Female Inflow Diversity	(1)	(2)
Fraction of Male Employment	0.1633	0.2467***
	(1.3352)	(8.1886)
Log(Female Employment Size)		0.2116**
		(9.702)
Change in Female Employment Size		-1.93e-4**
		(-2.7145)
Female Average Wage		4.399e-07*
		(2.3471)
Constant	2.046***	1.904*
	(26.214)	(1.7877)
Observations	386	386
R-squared	0.0046	0.136

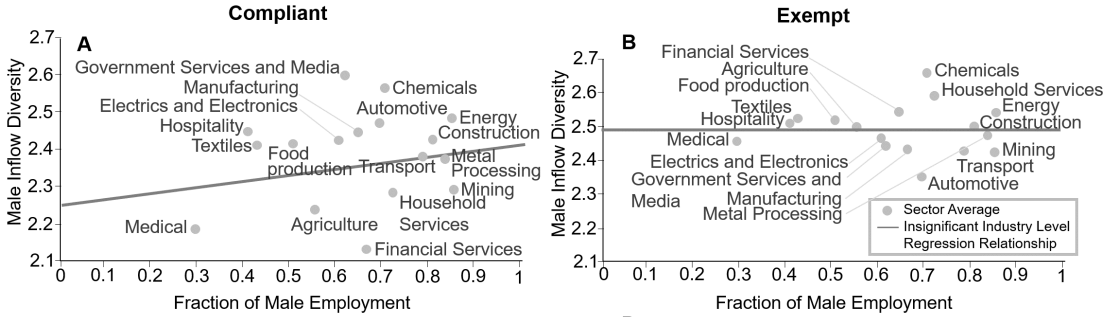
Robust t-statistic in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 6.7: (A) and (B) Graph showing the relationship between the percentage of male employment and the female inflow diversity over the period 2011–2014 for both the group of firms who comply with the EE Act and those who are exempt from the EE Act. Industries are binned into c-sectors for visualisation purposes. (C & D) Regression table quantifying the same relationship. We observe the strongest relationship between the percentage of male employment and the female inflow diversity for firms that comply with the EE Act.

industries have a higher female inflow diversity, especially amongst firms that comply with the act.

Similarly, the relationship between the percentage of male employment and the male inflow diversity of an industry for our treatment and control group is shown in Figure 6.8. Compared to the female case above, a weaker relationship between the percentage of male employment and the male inflow diversity is found in both the compliant and exempt groups. Furthermore, no significant difference is found between the compliant and exempt relationship in the male case, with low R-squared values. If we compare the male and female relationships with each other, we can conclude that the strongest relationship is found between the percentage of male employment and the female inflow diversity for the group that complies with the EE Act.

Previously we found that the act enhanced female inflow diversity for firms with roughly 50 employees. Furthermore, we found that the female inflow diversity is larger in male-dominant industries. Although we can not directly link these results,



C

Dependent Variable: Male Inflow Diversity	(1)	(2)
Fraction of Male Employment	0.1332***	0.1276
	(3.965)	(1.614)
Log(Male Employment Size)		0.0821***
		(9.061)
Change in Male Employment Size		-3.444e-06**
		(-2.645)
Male Average Wage		-2.0671e-07
		(-1.526)
Constant	1.9931***	2.2505***
	(37.169)	(17.025)
Observations	386	386
R-squared	0.0393	0.125

Robust t-statistic in parentheses
*** p<0.01, ** p<0.05, * p<0.1

D

Dependent Variable: Male Inflow Diversity	(1)	(2)
Fraction of Male Employment	0.0212	0.0668*
	(1.920)	(7.572)
Log(Male Employment Size)		0.01726***
		(17.356)
Change in Male Employment Size		-5.938e-05
		(-0.624)
Male Average Wage		-7.544e-07
		(-2.916)
Constant	2.0392***	2.4950*
	(28.942)	(6.492)
Observations	386	386
R-squared	0.0095	0.106

Robust t-statistic in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 6.8: (A) and (B) Graph showing the relationship between the percentage of male employment and the male inflow diversity over the period 2011–2014 for both the group of firms who comply with the EE Act and those who are exempt from the EE Act. Industries are binned into c-sectors for visualisation purposes. (C) and (D) Regression table quantifying the same relationship.

they suggest that the act has the greatest influence in enhancing the female inflow diversity in male-dominant industries.

6.7 Conclusion

This study aimed to investigate the EE Act's impact on gendered inter-industry labour mobility patterns. We find that the act increased the diversity of the sectoral origin of female labour inflows for firms. Furthermore, the act increased the related variety of newly hired skilled female workers for firms. However, unlike the findings in the literature where these types of flows positively influence firm performance, the artificial increase in the related variety of female workers did not translate to any significant impact of the act on changes in firms' labour productivity growth. Furthermore, we found that male-dominant industries display higher female inflow diversity.

We contribute to the affirmative action (AA) literature by showing how these policies can indirectly influence labour mobility patterns (increasing the total variety of all PDI workers and related variety of skilled PDI workers). As labour mobility patterns are crucial for knowledge diffusion and influence economic growth and resilience, industrial policy-makers need to understand how these AA policies influence labour mobility.

This work also contributes to the evolutionary economic geography literature, first by constructing the South African SRN. According to the author's knowledge, this is the first SRN constructed for an African country. We also develop a skill-based industry classification which groups industries according to their skill and knowledge overlap. Secondly, we add to the evidence that the general increase in the related variety of skilled workers is associated with an increase in a firm's labour productivity growth in South Africa. Although this result has been found for plants in Sweden, we show that it is also true for firms in South Africa. Furthermore, we show that AA-induced increases in the related variety of skilled female workers may not materialise into the same significant effect on firm performance. This work warrants further investigation into potential barriers to AA-induced labour flows from enhancing firm performance.

Beyond contributing to the economic literature, our paper also makes advances from a methodological perspective. Our policy evaluation approach is unique in that it combines an RD design with network analysis techniques. Within the network science literature, although much attention has been placed on detecting structural

anomalies [171], few researchers have extended these techniques to infer causal effects. Similarly, although much attention has been placed on investigating causal impact in the labour market regulation literature, few studies have focused on inferring causal impacts regarding labour flow network characteristics or structure.

6.7.1 Limitations and future work

Our study also has various limitations. First, we adopt a bandwidth of 10 employees in our regression discontinuity design. This is because this bracket of employees is grouped in the affirmative action act. However, the robustness of the study's results could be enhanced by ensuring that the discontinuity remains significant for slightly smaller or larger bandwidths. We mitigate the impact of the size of the bandwidth by implementing a triangular kernel function, ensuring that values closer to the cutoff have a higher weighting. However, as we no longer have access to the raw data, due to its access only being granted in the South African Treasury, we could not implement this robustness check.

Secondly, the author is aware that the results may be biased by the characteristics (e.g. the distribution of industries) of the firm's with the treatment and compliant groups. To provide more robust results, a more thorough investigation into the type of firms (more specifically, the number of firms in each industry and their hiring characteristics) in each group should be undertaken. This will ensure that the results are not due to the different types of industries within the two groups. We do not expect a significant difference distribution of industries within the two groups, particularly in our bandwidth (firms with 40-60 employees). However, this is a limitation of the study. Due to a lack of data access, we were unable to implement these robustness checks.

Furthermore, we wish to highlight that our results may be biased by the movement of firms in reaction to the act. For example, firms may deliberately down-scale or prevent the growth of their workforce (remaining at 49 employees) to not have to comply with the EE Act. This could be to prevent financial penalties incurred if the firm was to comply with the act. The bias may affect our results by reducing the value of the inflow diversity of firms with 49 or fewer employees. Thereby, also making our treatment effect larger than expected. With limited data on each firm's behaviour, we are unable to investigate the degree of this bias. However, when investigating the general relationship between the inflow diversity and the size of firms within our treatment group, we do not observe a significant jump in the inflow diversity value

closer to our cut-off point. This provides some confidence that this bias is not the sole reason for the treatment effect we observe

Thirdly, we highlight another labour market act in South Africa, the Labour Relations (LR) Act of 1995, which is compulsory for firms with more than 50 employees. As it has the same cutoff value, this LR act may also provoke the discontinuities we observe. The LR act provides regulations on how employment relationships should be dealt with. More specifically, the act gives rights to an employer and employee and dictates how each party conducts themselves in the employment relationship. The act applies to all individuals and does not differentiate between PDIs and other employees (*i.e.*, across gender or race). As our results are only significant for newly hired female workers and not for newly hired male workers, it is unlikely that this LR act is the primary driver of the discontinuities we observe. However, our results could be strengthened by running a regression discontinuity design on firms within industries that are exempt from the EE act and not the LR act (*e.g.*, the National Intelligent Services and the South African Secret Service). Ensuring that these industries do not lead to a discontinuity could provide further evidence that the EE act has, in fact, caused the discontinuity that we observe within our results.

Furthermore, we only consider the act's impact in a short period (2011-2014), even though the act has been implemented since 1998. We specifically consider this period, as there were no act amendments implemented during the period. Furthermore, the cutoff value of 50 employees was kept constant. We emphasise that our results are specific to this period and can not be generalised to other periods. Due to data availability, we have not been able to consider labour flows before the implementation of the act. However, if this data does become available, this could be interesting future work.

We also only consider labour flows if the worker comes from a prior industry. Therefore, we do not consider the inflow of new employees that enter the formal labour market for the first time. It would be interesting to evaluate what portion of the firm's labour flow are new to the labour market and if our results would still hold when including these labour flows. However, data on the type of skills or educational background of these workers would need to become available for this to be included.

In this study we only investigated the act's impact on gendered labour flows. However, the act also groups workers across races. The implications and effects of this are not captured in our study. This is because data containing a racial variable is not available. If this does become available, a fruitful and crucial future research

step would be to replicate our results along racial lines. This would provide more direct evidence to evaluate the act's intended impact.

Finally, our study investigates the causal impact of the EE act on the sectoral-diversity of inflowing labour. An interesting further research endeavour could consider investigating how the act has influenced the structure of the male and female labour pools. More specifically, how the modular structure of the male and female inter-industry labour flow network has changed before and after the act's implementation. This could highlight how the act has influenced both knowledge specialisation (through communities becoming more isolated) or the generalisation and merging of skills (through communities becoming more connected). As these structures influenced knowledge diffusion and regional industrial diversification, these are crucial insights to ensure that the AA act does not influence labour mobility patterns in a way that restricting economic growth.

To do this, one could adopt the modular comparison techniques (developed in Chapter 3 and 4) with an Interrupted Time Series Design (ITSD) [42]. An ITSD is an quasi-experimental technique that is used when time-series data (*i.e.*, data before and after the policy implementation) is available. This would enable an investigation of whether there is any causal link between the act's implementation and significant changes in these networks modular structure. However, data from before the act's implementation will need to become publicly available for this research to become viable.

6.7.2 Policy implications

Our study also has clear policy implications. First, our study warrants that governments should keep track of the impact of changes in inter-industry labour flow patterns when implementing AA policies. This is because inter-industry labour flow patterns are crucial for the diffusion of knowledge and, thereby, industrial diversification and economic growth. As our study showed, AA can enhance the sectoral diversity of newly hired PDIs. This could potentially widen (reduce) the modular structure of industrial clusters or labour pools within the labour market. Although this could be advantageous to allow knowledge to diffuse more easily, it could also reduce the specialisation of knowledge which is key to fostering competitive advantage within an economy. Governments, therefore, need to track these implications carefully when introducing AA policies.

Furthermore, we find that the act is causing firms to recruit female workers from more diverse sectoral backgrounds. This shows that firms are actively implementing

new recruitment strategies to find and attract female candidates outside their own industry or sector. Policy could aid these firm search efforts. This could include helping firms to hire workers from specifically related industries and not from unrelated industries, which is more likely to enhance their performance. Our SRN, and our new industry classification, provide a novel tool to quantify the relatedness between industries. Government agencies and NGOs could deploy insights from this network to aid firm search efforts in related industries. Specific support could include, for example, large career fairs targeting several related industries or the provision of bursaries to candidates in related industry fields. Furthermore, higher education institutions could be encouraged to form knowledge partnerships with firms in related industries. These knowledge partnerships could include training courses or schemes to create work tasters, pre-recruitment training and guaranteed interviews for candidates in related industries [126].

Chapter 7

Conclusion

In this thesis, we developed and implemented various network sciences tools to model and analyse inter-industry labour flow networks. Here, we summarise a few key results and highlight how they contribute to the literature. We also indicate the broader implications of our work and show some potential future research avenues. At the end of each chapter, we provide a more thorough conclusion of the research findings, their limitations and potential future work.

In the first part of this thesis, we contribute to the literature on modular network comparison. Comparing the modular structure of different networks is challenging, and few techniques have been developed for this purpose. In Chapter 3, we propose the Bi-directional distance, which compares the global modular structure of two node-aligned graphs by assessing the fit of one's partition to the topology of the other network (and vice versa). This technique is particularly advantageous as it incorporates information on the underlying network structure within the comparison, unlike popular set-theory-based measures (*e.g.*, NMI), which equate modular structure with the network's partition. This enables the measure to identify similarities and differences between different networks' modular structures that are not captured by their partitions alone. The two-dimensional structure of the distance also enables the measure to indicate when one network's partition describes the modular structure of the other well, but not vice versa. This typically occurs when one network is a sparser or nested version of the other. Furthermore, as the framework is agnostic to a community detection algorithm (although limited to those that adopt quality function optimisation), it can be more broadly applied to compare communities that are defined different.

In Chapter 4, we propose another modular network comparison technique. This technique is aimed at comparing a single community across different node-aligned undirected graphs. Although understanding how a single community differs across

networks or across time is key to understanding parts of complex systems, no modular community detection algorithm takes (or can be adapted to take) a single community perspective. Furthermore, local quality functions fail to capture the influence of both the inner- and outer-community edge density and their connectivity. We address this gap in the network toolbox by developing the maximum retention distance. This metric compares the modular structure of a pre-defined (connected) community by comparing the largest difference in the degree to which the community can conserve dynamics, starting within the community, across two graphs. By comparing a dynamic process on the graph, the measure captures both the influence of community edge density and connectivity simultaneously. This enables the measure to capture the influence of node roles on the modular structure of the community (e.g. if core nodes or peripheral nodes are connected outward to other communities). Furthermore, by taking a local approach and not employing a null-model comparison to quantify the modular structure of the community, the measure can be used to compare communities across a pair or set of different networks.

Both our local and global modular comparison techniques have enhanced the network toolbox, particularly for industrial policy-makers. This is because our measures reveal the similarities (and differences) in the structure of knowledge flows in labour flow networks of different European countries. The similarities we find enable policy-makers to uncover universal structures in these networks and show where portability of networks (or parts of network) across contexts is possible. On the other hand, the differences we find provide insight into hidden growth opportunities: linkages and clustering patterns present in one country which suggest potential unseen industrial diversification paths for another. These novel insights are vital to allow for the better modelling and understanding of industrial diversification processes of different regions.

In the second half of the thesis, we focus on using network techniques more broadly to model and analyse inter-industry labour flow networks. We mainly focus on two related questions in the economic geography literature.

In Chapter 5, we contribute to the research question in economic geography on whether attracting multinational enterprises (MNEs) into a host region can enhance its industrial diversification opportunities. In other words, can 'imported' skills and know-how unavailable locally be used as 'stepping-stones' to enhance the region's industrial diversification path? More specifically, we investigate if MNEs enhance the entry and survival of related domestic industries (particularly Irish manufacturing and

exporting industries) through knowledge spillovers in Irish regions. We find that so-called 'overlapping industries' (industries that contain both domestic and MNE firms) positively influence the domestic economy by both increasing related domestic firm entry and survival. In contrast, we find that 'exclusive MNE' industries negatively impact a host region. They reduce related domestic industry entry and survival. Our results show that domestic industries cannot 'leap' into these MNE-dominated industries, likely due to a know-how gap between the two types of firms. In contrast to these results, we also found evidence of a cluster of domestic industries entering into MNE-exclusive industries in the last period of our analysis. We hypothesise that this might be explained by government schemes designed to stimulate and boost domestic-MNE links in response to Brexit.

This work also contributes methodologically by introducing a new type of cohesion measure to capture the influence of higher-order linkages or a cluster of related industries surrounding an industry. More specifically, it considers the connectivity of related industries and their connectivity to other industries in the region. Our results show that higher-order linkages matter. Specifically, in the recovery period (a time of economic hardship), higher-order linkages to overlapping industries enhanced domestic industry entry and survival, in contrast to purely direct linkages. Therefore, the key to domestic industrial diversification is to enter into industries that are surrounded by 'clusters' or groups of existing interconnected, overlapping industries.

In Chapter 6, we investigate another related economic question: How do affirmative action policies indirectly influence patterns of inter-industry labour mobility? More specifically, we investigate how the Employment Equity (EE) act has impacted the sectoral diversity of male and female workers in South Africa. Our results show that the act enhanced the sectoral diversity of newly hired female workers but had no significant impact on the male case. Furthermore, amongst the inflow of skilled workers, we found that the act enhanced the inflow of women from related industries, but not those from unrelated industries. However, unlike the findings in the literature, where these inflows enhance firm performance, we find no significant impact of the act on the growth of a firm's labour productivity. We also find a positive and significant relationship between the percentage of male workers within an industry and the industry's sectoral diversity of newly hired female workers. Although we show no causal relation, this result suggests that the act may have a more considerable impact on male-dominant industries.

This chapter also adds evidence of the use of quasi-experimental techniques to evaluate the causal impact of policies on network structure. More specifically, we

illustrate using a regression discontinuity design on changes in a local entropy-based centrality measure on a network. Recently, a large focus has been placed on causal policy evaluation within the economic literature. Similarly, network science has seen increased techniques designed to detect network anomalies. However, only a few studies have combined these techniques.

Overall, the tools we developed in this thesis contribute to a greater research drive to enhance evidence-based tools for industrial policy. These policies are aimed at enhancing regional industrial diversification and regional economic growth. By comparing the modular structure of the underlying economic landscapes across different countries, governments can better assess how their landscapes differ and which countries should adopt similar policies. Secondly, by developing a cohesion measure and modelling the impact of MNEs on domestic industry dynamics, governments can use these tools to gain insight into which MNE industries will most positively influence their host regions. Finally, when implementing affirmative action, governments should be weary that it enhances sectoral diversity of labour flows. This could potentially broaden labour pools but could also prevent knowledge specialisation. Although our modelling tools and findings are not conclusive on their own, they contribute to a greater package of analysis which can be used to develop informed and strategic government policies and investments.

7.1 Future work

In chapter 3, we proposed the bi-directional distance to compare the modular structure of two node-aligned networks using a descriptive quality function. However, this technique can be extended to adopt other community detection techniques or to compare other meso-scale structures.

First, our bi-directional framework could be extended to adopt inferential community detection algorithms. As these techniques describe a precise generative model with a corresponding community structure, adopting these techniques would allow our measure to compare the generative model that assembles the modular structure of networks. Taking a partition swapping approach, one could compare the posterior distributions for each network's own partition and the second graph's obtained partition. One could use a posterior odds ratio or the descriptive length as the distance function to compare the two partitions on each graph.

Furthermore, our framework could be extended to compare other meso-structures. For example, the core-periphery structure of two networks. As this structure can

also be evaluated through quality functions, we could adopt one of these quality functions to compare this structure across different node-aligned networks. Similarly, our framework could compare the degree of associativity across different networks. As this is often measured using the modularity function, we could adopt this function in our framework. As our measure considers the underlying structure (and thereby the underlying group ties), this measure provides a unique way to capture the overlap in group association and their tendency to link together.

In chapter 4, we developed the maximum retention distance, a single-community comparison technique. Although this technique allows for parts of communities to be compared, it can only be applied in some networks. For example, the network needs to be undirected and the community connected in both graphs. We could extend this measure to the directed case or cases where the connectedness requirement is not met. This could potentially include adapting a random walk transportation step to isolated nodes, as demonstrated in the PageRank centrality measure.

Furthermore, the retention distance only compares the topology of the community and does not capture where it lies in the broader network. A fruitful area of future work could develop a different metric to capture the distances between the community and other communities in the network. Evaluating how the distances between communities differ across networks or changes in time can add key information to understand how networks differ or are changing. For example, in the case of our SRN, this could show the process of skill clusters merging or separating as their knowledge becomes more specialised and distinct from each other.

In chapter 5, we investigated whether MNEs positively enhance domestic industrial diversification through knowledge spillovers. We constructed a higher-order cohesion measure that captured the knowledge spillovers occurring from a cluster of inter-connected related industries. We showed that these types of knowledge spillovers do occur and matter. These results provide evidence that it is not only directly related industries that influence industrial diversification processes. This provides a rationale for the development of industrial diversification modelling techniques that consider more of the underlying SRN structure. Currently, most techniques take a local approach (where only neighbouring industries are considered). An interesting extension could consider how the presence of existing industries in an industry's industrial cluster or skill-basin influences its dynamics. Recall that the SRN has a modular structure, where communities represent skill basins. These skill basins are seen as labour pools where workers can freely move between these industries. Knowledge is thereby also more easily transferred amongst these industries. Interesting

further research questions could include: How does the number and size of MNEs in these skill-basins influence the dynamics of the domestic industries present? Furthermore, does the modular structure or isolation of these skill basins affect the degree of spillovers between these firms?

Furthermore, in our analysis, we found evidence of a cluster of domestic industries entering into MNE-exclusive industries. These entries are particularly interesting as they enter into industries unrelated to that of the industrial basket. Although we showed no statistical relationship, we hypothesise that this might be explained by government schemes designed to stimulate and boost domestic-MNE links in response to Brexit. It would be interesting to investigate if there is a causal relationship between domestic entry and these schemes. One could also investigate the survival rate of these new entries compared to entries in other periods, providing empirical evidence on the long-lasting economic benefits from these dynamics (and potentially these schemes). This work could contribute to related literature that investigates unrelated diversification or 'leap-frogging' - a process known to be rare and where little is known about the conditions required for firms to 'leap' into these more distant sectors [175].

Finally, in chapter 6, we demonstrated the use of a regression discontinuity design with a network centrality measure to show how the EE act influenced the sectoral diversity of newly hired workers of firms in South Africa. Future research could consider how the act has influenced the structure of labour pools in the labour market. Recall that these labour pools both enable knowledge specialisation, which enhances a region's competitive advantage, but also restricts knowledge diffusion and industrial diversification opportunities. Therefore, understanding how these structures are changed by the act is an important research endeavour. To investigate this, one could combine an Interrupted Time Series Design (a quasi-experimental technique that specifically uses time-series data) with some of the modular comparison techniques developed in the first part of this thesis to consider how the modular structure of the male and female inter-industry labour flow networks has changed in response to the act. This research could also contribute to the current research endeavour of bringing together quasi-experimental techniques with network anomaly detection.

Furthermore, a fruitful and important future research step would be to replicate our results in this chapter along racial lines. This is because the AA act also groups workers across races. This would provide more direct evidence to evaluate the act's intended impact. However, data containing a racial variable would need to become available for this research to become plausible.

Appendix A

SI: Global modular comparison

A.1 Technical details regarding SBM construction

In Figure 3.3 we can see the performance of our bi-directional measure in three different families of SBMs. Here, we summarize how these three different families were generated. All networks consist of 120 nodes and are generated according to the following parameters:

- All networks in family A consist of 4 identical blocks of size 30, and the SBM parameters are $P_{ij} = 0.20$ and $P_{ii} = \{0.80, 0.60, 0.45, 0.35, 0.20\}$ for each successive network. Since for the last network, $A5$, both probabilities are the same, we do not have any a priori community structure, and we can use this network as a point of comparison.

- For all networks in family B, $P_{ij} = 0.20$ and $P_{ii} = 0.80$. But each successive networks splits some of the blocks of the previous one. So $B2$ split one of the original blocks into two, $B3$ splits the other original block to obtain 4 identical medium blocks. $B4$ splits one of these 4 blocks into 3 smaller ones, and $B5$ repeats that procedure on the other 3 medium blocks.

- $C1$ and $C2$ are produced by again using $P_{ij} = 0.20$ and $P_{ii} = 0.80$, with the first one split into 3 blocks and the second one into 4 blocks. $C3$ has the same blocks as $C2$, but $P_{2,3} = P_{3,2} = 0.40$.

Furthermore, in Figure 3.6 we illustrate the use of the BiDir distance using a multi-resolution community detection algorithm. Here, we provide more detail regarding how the three nested SBMs were constructed. All three of the networks consists of 300 nodes and have implanted communities at 3-levels. As illustrated in the adjacency matrices of each network, each network contains regions with one of 4 different shades

of grey. These regions correspond to 4 different values of P_{ij}. From darkest to lightest the regions correspond to the following four values of $P_{ij} \in \{0.8, 0.53, 0.27, 0.15\}$. The size of the different regions for each network are given as follows:

- Network A consists of 12 identical blocks of size 25. Two of these blocks are then merged together to form 6 identical blocks of size 50. Similarly, two of these blocks are then grouped together to form 3 identical blocks of size 100.
- Network B, consists of 18 blocks of size $20, 10, 10, 25, 10, 25, 20, 10, 10, 25, 10, 25, 20, 10, 10, 25, 10, 25$, respectively. Two consecutive blocks are then joined together to create 9 blocks of size $30, 35, 35, 30, 35, 35, 30, 35, 35$, respectively. Finally, three consecutive blocks are joined together to create three identical blocks of size 100 - similar to Network A.
- Network C, consists of 12 identical blocks of size 25, similar to Network A. These blocks are merged together to form four communities contain ing three blocks of size 75. Finally, two 75-sized blocks are joined together to form two communities of size 150.

Appendix B

SI: Single-community comparison

B.0.1 The evaluation of other summary statistics

In chapter 4 we introduced the maximum retention distance. Although we adopt the maximum value, there are various other summary statistics that can also be used to reduce the retention distance curve into a single-valued distance. Here, we investigate the merits of three other summary statistics, namely taking the retention distance curve at $t = 1$, comparing the largest eigenvalues and finally taking the integral of the retention distance curve. We show that each of these summary statistics are problematic by either not capturing both the influence of edge density and edge connectivity when comparing the communities or not adhering to the normal distance behaviour [195, 118].

B.0.1.1 The retention distance at time = 1

First, we consider using the retention distance obtained at $t = 1$. Recall that this is equivalent to local modularity [44]. We previously showed that this measure only captures the structure of the boundary nodes, but ignores the connectivity of the rest of the community. We again illustrate this using toy networks.

In Figure B.1, we illustrate 4 toy networks A-D, where we focus on comparing the modular structure of the blue community in network A to its' structure in each successive network. We observe, that to obtain the community structure in networks B and C, we remove 2 inner-community edges or add 3 inner-community edges, respectively. The community in D has a completely different structure - a star graph with a highly heterogeneous degree distribution.

Here, we focus on comparing the blue community across graph A and B. As graph A has 2 additional inner-community edges it is more modular. We show the resulting distances obtained in Figure B.1 (I). We observe, that the retention distance at $t = 1$

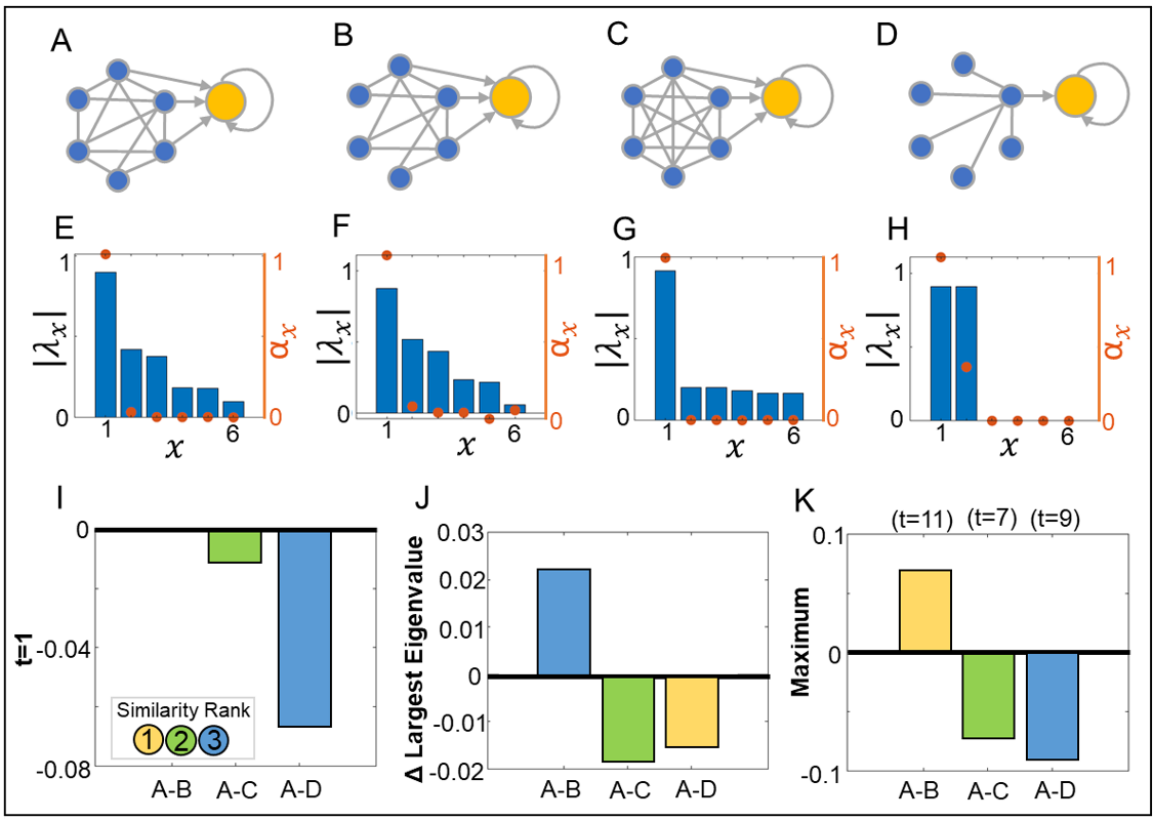


Figure B.1: Comparing networks using the retention distance function with different summary statistics. (A-D) Four toy networks, where we compare the blue community in A to its structure in each successive network. (E-H) The eigenvalue and coefficients of the system matrix of each corresponding toy network (above the graph). (I-K) The retention distance for each comparison using (I) the retention at $t = 1$, (J) comparing the largest eigenvalue and (K) adopting the maximum distance.

is unable to capture a difference between the community in A and B. This is as the boundary nodes of the two graphs are the same. The measure only evaluates their inner-community strength of the boundary nodes and ignores the structure of the rest of the nodes in the community. This measure is therefore unable to capture the influence of community edge density and connectivity when comparing a community across a network pair.

B.0.1.2 Comparing the largest eigenvalue

It is well-known result that the dynamics acting on a graph can be fully described by the eigenvalues and eigenvectors of its system matrix (*e.g.*, the graph laplacian). Each eigenvalue defines the exponential rate of decay along its corresponding eigenvector. With the assumption that all eigenvalues are orthogonal and less than one [123], the exponential nature of the decay means that the largest (or second largest) eigenvalue

can offer a sufficiently accurate approximation of the relaxation time of the diffusion dynamics [80]. Here, we investigate whether we can compare the largest eigenvalue of the Q matrix as way to summarise the retention distance curve.

In Section 4.4.3, we showed how we can redefine the retention function in terms of the eigenvalues and eigenvectors of its system matrix (Q). Here, we saw that the retention function is mainly dominated by the decay of the largest eigenvalue. If the latest eigenvalue closer to 1, the community is isolated from the rest of the graph. On the other hand, the smaller its value, the faster the rate of decay and therefore the more connected the community is to the rest of the network.

We also showed in Section 4.5.3.1, that for a special case graph which has a constant inner- and outer-community strength the retention function can exactly be defined using only the largest eigenvalue. From these results, it seems apparent that comparing the largest eigenvalue would be a good summary statistic to compare the modular structure of a community across a graph pair.

However, this metric performs less well when we consider more heterogeneous community connectivity (*e.g.* a community with a power-law degree distribution). This is because these structures display a more varied eigenspectra. We illustrate how this can fail to capture community connectivity using our same toy networks in Figure B.1.

We again focus on comparing the blue community in A to its structure in each of the other toy networks. Recall, that network B and C are produced by removing and adding inner-community edges, respectively. On the other hand, D is has a completely different star-like inner-community structure. We illustrate the magnitude of the eigenvalues and the coefficient α_i (as defined in Eq (4.5)) for each of the graphs A-D in E-H, respectively. For graph D, we observe (in sub-figure H) that both the first and second eigenvalue (as well as their corresponding coefficients) are large. Therefore, both of these values are needed to well describe the dynamics of a random walker on graph D, the dynamics will not be well described by only the largest eigenvalue.

We show how this can obscure the ranking of our toy network comparison when adopting the difference in the largest-eigenvalue as our distance measure. In sub-figure B.1 (J), we illustrate the distances obtained. We observe that the metric obtains a larger distance between the modular structure of the blue community in graph A-C than A-D. However, we can easily see that C has a more similar modular structure than D. The measure fails to capture this similarity, as it only considers the largest eigenvalue (ignoring key information in the second eigenvalue).

We could also compare the second (and third) largest eigenvalue. However, this increases the dimensionality of the measure. It is also unclear how to summarise these values into a single value. For example, it would be difficult to take a weighted average as the eigenvalues would take different weights in different graph structures.

As comparing the largest eigenvalue is unable to pick up differences in edge density and connectivity when comparing the community modular structure - especially when the community connectivity is heterogeneous (community displays power-law degree distribution) it is not a well-suited summary statistic for our application.

B.0.1.3 Taking the integral

Another obvious summary statistic is to take the area under the retention distance curve. However, the measure does not comply with the Edge-submodularity property [118] (*i.e.*, a change in a sparse network is more important than a similar change in a denser network of the same size) and therefore does not hold to regular network distance behaviour. Furthermore, this summary statistic is also problematic, as the area under two retention distance curves are often equivalent even when the curves have different shapes. Here, we illustrate both of these problematic properties.

First, we show that the area under the retention distance curve is equivalent to comparing the expected first-mean-passage time (FMPT) of the walker (starting on any node within the community) to be absorbed by RON. In other words, the difference in the expected number of steps a walker takes before it escapes.

To show this, we define the fundamental matrix as $Z = \sum_{k=0}^{\infty} Q^k = (I - Q)^{-1}$. Recall, that $Q^k(i, j)$ is the probability that the walker will start on i and be at j after k steps (before being absorbed). A well-known result is that the FMPT can be given by $Z\mathbf{1}$. The i-th element of $Z\mathbf{1}$ is the expected number of steps a random walker starting on node i will take before being absorbed. The expected FMPT for a node starting within the community can then be given as $\frac{1}{n_c}\mathbf{1}Z\mathbf{1}$.

The expected FMPT can be written in terms of the retention as:

$$\begin{aligned}\frac{1}{n_c}\mathbf{1}Z\mathbf{1} &= \frac{1}{n_c}\mathbf{1}\sum_{t=0}^{\infty} Q^t\mathbf{1} \\ &= \sum_{t=0}^{\infty}\frac{1}{n_c}\mathbf{1}Q^t\mathbf{1} \\ &= \sum_{t=0}^{\infty}\rho(C, G, t)\end{aligned} \tag{B.1}$$

Hence, the integral under the retention curve is the expected FMPT.

This is also equivalent to the random walker centrality [156] of RON. The random walk centrality is a measure, often applied to communication networks, that captures the ability of a node to efficiently receive information from all other nodes.

Next, we show that the area under the retention distance curve does not comply with the Edge-submodularity property. We start by again considering the special case community with constant inner- and outer-community strength (within both graphs). Note that this graph is stochastically equivalent to an ER structure community (with constant inner- and outer-edge probability), and thereby represent a community within a SBM graph. For this case, we showed that the retention function for a community C in graph $G1$ is given as $\rho(C, G1, t) = (\frac{k_{in}}{k_{in}+k_{out}})^t$.

Now we consider a perturbation which increases the inner-community node strength (k_{in}) by x, resulting in a new graph $G2$ with inner-community node strength of $k_{in}+x$. The distance (using the integral which is equivalent to the expected FMPT) between the community in $G1$ and $G2$ is given by:

$$\begin{aligned}\frac{1}{n_c}\mathbf{1}Z(C,G1)\mathbf{1}' - \frac{1}{n_c}\mathbf{1}Z(C,G2)\mathbf{1}' &= \frac{1}{n_c}\mathbf{1}(I-Q_1)^{-1}\mathbf{1}' - \frac{1}{n_c}\mathbf{1}(I-Q_2)^{-1}\mathbf{1}' \\ &= \frac{1}{1-\frac{k_{in}}{k_{in}+k_{out}}} - \frac{1}{1-\frac{k_{in}+x}{k_{in}+x+k_{out}}} \\ &= \frac{k_{in}+k_{out}}{k_{out}} - \frac{k_{in}+x+k_{out}}{k_{out}} \\ &= \frac{-x}{k_{out}}\end{aligned} \tag{B.2}$$

Note, that the distance only depends on the size of x and the outer-strength (k_{out}) of the nodes in the community. This distance ignores the influence of the inner-node strength (k_{in}). Hence, a change in a community of low inner-edge density will have the same impact as a change in a high inner-edge density community. Hence, the "Edge-Submodularity" property is not held.

Thirdly, we show that the area under two retention distance curves are often equivalent even when the curves have different shapes. We see this happening when communities differ both by their inner-community density and their connectivity. We elaborate on this property by showing, through toy networks, where the integral is not able to capture differences in a community's modular structure.

In Figure B.2 A-C, we illustrate three networks where we compare the blue community in A to its' structure in B and C. B differs from A, in that it has two more inner-community edges - one of the edges (e(2,4)) is deep in the community and the other (e(3,6)) closer to the front (community boundary). C differs from A in that it has 3 more inner-community edges, and is therefore slightly more modular. These

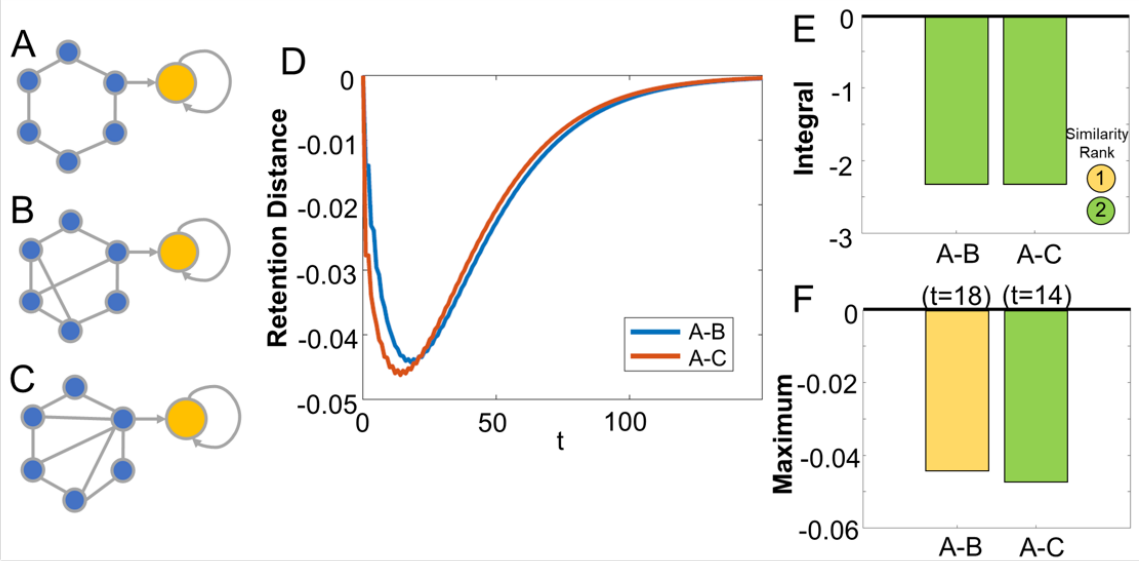


Figure B.2: Comparing a community across a set of toy networks using the integral and the maximum of the retention distance function. (A-C) Toy networks, where we compare the blue community in A to its structure in each successive network. (D) The resulting retention distance when comparing the community in A to its structure in B and C. (E & F) The distances obtained when summarising the curve in D using the integral and maximum, respectively.

edges are positioned close to the front of the community. In D, we show the resulting retention distance curve for both network pairs. We observe that A-C has a larger but earlier peak. In contrast, A-B has a smaller but later peak. The reason for this difference is that the networks differ both in edge-density and in their connectivity (more specifically the depth at which edges are added). In E, we summarise the curve by taking its integral. We observe that the area under both curves is equal. Therefore, by using the integral, we are unable to capture the differences in the community's modular structure across these network pairs.

For comprehensiveness, we also show the maximum retention distance obtained for the previous toy networks (shown in Figure B.1 and Figure B.2) used when evaluating the merits of other summary statistics. We observe that the distance can correctly capture the difference between the community's modular structure in both toy network sets.

Appendix C

SI: Ireland

C.1 Data descriptive

In this study we adopt an export firm employment dataset. This dataset is derived from the Irish Department of Business, Enterprise, and Innovation Annual Business Survey of Economic Impact and covers a large subset of exporting firms within the Irish economy assisted by government agencies. The dataset contains the employment size of industries (broken down into 4-digit NACE 2 industry codes) across regions (defined by NUTS level 3) and years. Furthermore, the dataset is broken down into employees who work for Irish firms and those who work for firms predominately owned through foreign investment. In table C.1 we show the number of MNE and domestic industries within each year and each Irish region that we use as observations within our analysis. We observe that there is a higher number of domestic industries compared to MNE industries in all regions and time periods within our data set.

Furthermore, in our study, we investigate how the cohesion to domestic and MNE industries is associated with domestic industry entry and survival. We investigate

Table C.1: Descriptors of the number of industries across time, within regions, and by ownership type within our sample data

	All		Border		Dublin		Mid East		Mid West		Midlands		South East		South West		West	
Year	DOM	MNE	DOM	MNE	DOM	MNE	DOM	MNE	DOM	MNE	DOM	MNE	DOM	MNE	DOM	MNE	DOM	MNE
2006	334	180	141	42	221	94	151	55	124	55	79	34	122	40	175	60	153	42
2007	346	181	142	41	221	94	151	52	130	55	82	33	124	38	175	61	157	41
2008	352	177	143	40	226	92	160	51	129	53	83	33	129	38	174	65	157	40
2009	361	176	146	40	230	90	158	52	128	50	85	32	129	38	176	63	156	38
2010	358	175	143	38	228	90	158	51	129	50	90	31	127	36	173	63	146	39
2011	358	177	144	38	225	92	161	50	128	47	89	31	129	37	172	61	140	37
2012	357	178	132	39	225	97	157	51	127	46	84	30	127	38	176	65	141	36
2013	358	180	133	38	228	102	154	50	120	45	84	31	122	35	173	66	141	36
2014	359	186	136	37	227	110	159	50	126	46	85	29	124	36	176	67	145	36
2015	359	190	140	36	232	113	159	47	127	49	88	27	125	38	166	67	144	37
2016	365	196	139	35	234	116	158	48	123	49	85	25	126	40	168	72	148	38
2017	366	197	143	37	233	116	161	49	127	50	86	28	127	40	165	74	147	39
2018	365	202	140	37	233	123	161	52	126	51	86	29	127	40	166	76	150	44
2019	363	203	141	36	230	123	162	52	124	52	85	30	126	43	164	74	149	45

the cohesion to three mutually exclusive subsets of industries, namely the so-called exclusive MNE industries (industries in which only MNEs are active), the exclusive domestic industries (industries in which only domestic firms are active) and the overlapping industries (those in which both are active).

In Table C.2 we show the descriptive statistics for both the presence and employment size of domestic industries and MNE industries, as well as those of the three different sets of industries mentioned. We show the descriptive variables for all regions and time periods together. In the table, X represents a binary variable indicating whether an industry is present and Emp represents the number of employees within an industry.

We observe that the presence of domestic industries is higher than that of MNE industries. We also find that there are, on average, the most exclusive domestic industries present and the least exclusive MNE industries present. What is particularly interesting is that we find that the employment size of multinational industries is larger than that of domestic industries. The set of industries with the largest number of employees on average is found to be the exclusive MNE industries. It is important to remember that this dataset specifically considers a set of industries that are supported by government agencies and not all industries (or employment) within Ireland.

C.2 Extended economic model results

In this section, we show an extension of the results shown and discussed in Section 5.6.1 and 5.6.2.

Table C.2: Descriptive on industry presence and size for different ownership-type industries

Variable	N	Mean	SC	Min	Max
X_D	51632	0.3208	0.4668	0	1
X_M	51632	0.1130	0.3166	0	1
X_{exclD}	51632	0.2476	0.4316	0	1
X_{exclM}	51632	0.0398	0.1955	0	1
$X_{overlap}$	51632	0.0732	0.2605	0	1
Emp_D	51632	42.1126	189.3348	0	8597
Emp_M	51632	47.4123	464.8045	0	23899
Emp_{exclD}	51632	34.0544	144.7844	0	4632
Emp_{exclM}	51632	39.3541	418.1570	0	23008
$Emp_{overlap}$	51632	16.1163	237.0211	0	17194

First, as an extension of the results shown in Table 5.3 and Table 5.4, we show how the cohesion, combining both WC and SC together, to domestic and MNE firms is associated with the entry of new domestic firms. Our results are found in Table C.3. We observe that our results are consistent with those found when considering each cohesion measure separately. Although the values of the coefficient slightly change, they do not loose significance or change sign. As our cohesion measures remain significant when controlling for the corresponding cohesion measure, this demonstrates empirically that our two measures pick up different dimensions of cohesion.

Similarly, we extend the results shown in Table 5.5, and Table 5.6, by investigating how the cohesion (using both the WC and SC) to domestic and MNE industries is associated with the exit of new domestic firms within a region. As before, and shown in Table C.4, our cohesion measures remain significant when controlling for the other cohesion measure, and our prior results therefore still hold.

Table C.3: Panel probit regression results for domestic industry entrance between 2006 – 2019 and their weighted closeness and strategic closeness measures as independent variable

Baseline period	2006-2009					2010-2014					2015-2019				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
MNE industry	-0.439	-0.429	-0.382	-0.400	-0.307	0.021	0.022	0.063	0.011	0.018	1.673***	1.640***	1.731***	1.635***	1.573***
	(0.315)	(0.315)	(0.329)	(0.332)	(0.345)	(0.306)	(0.308)	(0.308)	(0.306)	(0.312)	(0.451)	(0.453)	(0.460)	(0.496)	(0.510)
WC_{excD}		0.065			0.043		0.036*			0.034*		0.055			0.095
		(0.070)			(0.075)		(0.018)			(0.015)		(0.116)			(0.125)
WC_{excM}			-0.143*		-0.138*			0.286		0.357			0.268		0.189
			(0.125)		(0.086)			(0.206)		(0.225)			(0.258)		(0.306)
$WC_{overlap}$				0.504***	0.506***				0.114	0.053				0.105	0.105
				(0.133)	(0.139)				(0.131)	(0.146)				(0.156)	(0.163)
SC_{excD}		-100.009			59.815		575.357**			667.783**		72.406			209.336
		(238.468)			(249.803)		(311.964)			(324.805)		(421.255)			(440.398)
SC_{excM}			-351.830*		-412.785*			-182.456		-223.576			-339.799		53.867
			(304.387)		(346.299)			(342.552)		(350.216)			(477.790)		(562.256)
$SC_{overlap}$				686.059**	715.664**				293.107	413.751				1183.7**	1331.5**
				(285.401)	(306.401)				(344.324)	(370.886)				(559.926)	(627.794)
Region FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Constant	-8.841	-20.284	-8.033	-15.420	-23.395	-42.689	-21.395	-33.742	-23.587	-43.861	-16.145	-20.491	-8.576	-24.065	-17.988
	(8.140e+06)	(8.139e+06)	(4.398e+06)	(8.139e+06)	(8.139e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.755e+06)	(5.028e+06)	(5.755e+06)	(5.755e+06)
N	2522	2522	2522	2522	2522	2494	2494	2494	2494	2494	2507	2507	2507	2507	2507
AUC	0.9480	0.9482	0.9485	0.9520	0.9524	0.9700	0.9723	0.9709	0.9702	0.733	0.9814	0.9814	0.9814	0.9823	0.9825

Notes: Robust standard error in parenthesis; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.4: Panel probit regression results for domestic industry exits between 2006 – 2019 and their weighted closeness and strategic closeness measures as independent variable

Baseline period	2006-2009					2010-2014					2015-2019				
	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)	(1)	(2)	(3)	(4)	(5)
MNE industry	-0.607**	-0.545**	-0.597**	-0.621**	-0.632**	-0.140	-0.138	-0.144	-0.029	-0.061	0.025	0.020	0.079	0.0672	0.095
	(0.271)	(0.283)	(0.272)	(0.280)	(0.298)	(0.255)	(0.256)	(0.258)	(0.267)	(0.269)	(0.334)	(0.339)	(0.336)	(0.338)	(0.344)
WC_{excD}		-0.381***			-0.457***		-0.074			-0.122		-0.129			-0.077
		(0.145)			(0.151)		(0.086)			(0.095)		(0.096)			(0.104)
WC_{excM}			0.026		0.176			0.325**		0.393*			-0.346		-0.296
			(0.224)		(0.245)			(0.179)		(0.202)			(0.390)		(0.199)
$WC_{overlap}$				-0.242**	-0.418**				-0.355**	-0.365**				0.031	0.013
				(0.151)	(0.196)				(0.138)	(0.147)				(0.096)	(0.104)
SC_{excD}		297.217*			346.128		34.66			-41.043		-122.09			-351.48
		(204.274)			(399.523)		(330.05)			(349.99)		(334.72)			(351.96)
SC_{excM}			175.719		166.258			335.05*		306.4*			419.51		446.99
			(422.228)		(466.682)			(292.56)		(210.41)			(288.66)		(527.32)
$SC_{overlap}$				268.346	524.423				-363.83*	-382.07*				-579.99**	-534.61**
				(354.404)	(389.187)				(304.86)	(306.62)				(319.5)	(324.44)
Region FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Industry FE	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Constant	-1.193*	1.159	-1.203*	-1.034	1.898	-21.305	-20.389	-14.757	-9.565	-15.762	-15.163	-12.546	-16.339	-14.621	-16.49
	(0.622)	(0.986)	(0.623)	(0.641)	(1.076)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)	(3.32e+06)
N	1166	1166	1166	1166	1166	1194	1194	1194	1194	1194	1181	1181	1181	1181	1181
AUC	0.9466	0.9511	0.9466	0.9451	0.9515	0.9478	0.9477	0.9506	0.9548	-0.9562	0.9649	0.9660	0.9659	0.9657	0.9668

Notes: Robust standard error in parenthesis; * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

References

[1] D. Acemoglu, U. Akcigit, and W. Kerr. Networks and the macroeconomy: An empirical exploration. Technical report, National Bureau of Economic Research, 2015.

[2] R. Agarwal, M. Ganco, and R.H. Ziedonis. Reputations for toughness in patent enforcement: Implications for knowledge spillovers via inventor mobility. *Strategic Management Journal*, 30(13):1349–1374, 2009.

[3] B. Aitken, A. Harrison, and R.E. Lipsey. Wages and foreign ownership: A comparative study of Mexico, Venezuela, and the United States. *Journal of International Economics*, 40(3-4):345–371, 1996.

[4] J. Alcacer and M. Delgado. Spatial organization of firms and location choices through the value chain. *Management Science*, 62(11):3213–3234, 2016.

[5] A. Alexander-Bloch, R. Lambiotte, B. Roberts, J. Giedd, N. Gogtay, and E. Bullmore. The discovery of population differences in network community structure: new methods and applications to brain functional networks in schizophrenia. *Neuroimage*, 59(4):3889–3900, 2012.

[6] P. Almeida and B. Kogut. Localization of knowledge and the mobility of engineers in regional networks. *Management Science*, 45(7):905–917, 1999.

[7] R. Andersen and K.J. Lang. Communities from seed sets. In *Proceedings of the 15th International Conference on World Wide Web*, pages 223–232, 2006.

[8] M. Andersson and P. Thulin. Labor mobility and spatial density, 2011.

[9] A. Ando and F.M. Fisher. Near-decomposability, partition and aggregation, and the relevance of stability discussions. *International Economic Review*, 4(1):53–67, 1963.

[10] C. Antonelli. The business governance of localized knowledge: an information economics approach for the economics of knowledge. *Industry and Innovation*, 13(3):227–261, 2006.

[11] J.M. Arnold and B.S. Javorcik. Gifted kids or pushy parents? Foreign direct investment and plant productivity in indonesia. *Journal of International Economics*, 79(1):42–53, 2009.

[12] M. Ayyagari and R. Kosová. Does FDI Facilitate Domestic Entry? Evidence from the Czech Republic. *Review of International Economics*, 18(1):14–29, 2010.

[13] A. Bahadur, E. Lovell, E. Wilkinson, and T. Tanner. Resilience in the SDGs: Developing an indicator for Target 1.5 that is fit for purpose. Technical report, Overseas Development Institute, London, 2015.

[14] P.A. Balland, D. Rigby, and R. Boschma. The technological resilience of US cities. *Cambridge Journal Of Regions, Economy And Society*, 8(2):167–184, 2015. Publisher: Oxford University Press.

[15] R. Balsvik. Is labor mobility a channel for spillovers from multinationals? Evidence from Norwegian manufacturing. *The Review of Economics and Statistics*, 93(1):285–297, 2011.

[16] A.L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[17] M. Barahona and L.M. Pecora. Synchronization in small-world systems. *Physical Review Letters*, 89.5:54101, 2002.

[18] S. Barrios, S. Dimelis, H. Louri, and E. Strobl. Efficiency spillovers from foreign direct investment in the EU periphery: A comparative study of Greece, Ireland, and Spain. *Review of World Economics*, 140(4):688–705, 2004.

[19] F. Barry. Diversifying external linkages: The exercise of Irish economic sovereignty in long-term perspective. *Oxford Review of Economic Policy*, 30(2):208–222, 2014.

[20] F. Barry. Outward-oriented economic development and the Irish education system. *Irish Educational Studies*, 33(2):213–223, 2014.

[21] F. Barry, H. Görg, and E. Strobl. Foreign direct investment, agglomerations, and demonstration effects: An empirical investigation. *Review of World Economics*, 139(4):583–600, 2003.

[22] F. Barry, H. Görg, and E. Strobl. Foreign direct investment and wages in domestic firms in Ireland: Productivity spillovers versus labour-market crowding out. *International Journal of the Economics of Business*, 12(1):67–84, 2005.

[23] N. Bassett-Jones. The paradox of diversity management, creativity and innovation. *Creativity and Innovation Management*, 14(2):169–175, 2005.

[24] G.S. Becker and K.M. Murphy. The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, 107(4):1137–1160, 1992.

[25] G. Békés, J. Kleinert, and F. Toubal. Spillovers from multinationals to heterogeneous domestic firms: Evidence from hungary. *The World Economy*, 32(10):1408–1433, 2009.

[26] M. Blomström and A. Kokko. Multinational corporations and spillovers. *Journal of Economic Surveys*, 12(3):247–277, 1998.

[27] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[28] L.A. Booysen and S.M. Nkomo. New developments in employment equity and diversity management in South Africa. In *International Handbook on Diversity Management at Work*. Edward Elgar Publishing, 2014.

[29] R. Boschma. Towards an evolutionary perspective on regional resilience. *Regional Studies*, 49(5):733–751, 2015. Publisher: Routledge.

[30] R. Boschma, R. Eriksson, and U. Lindgren. How does labour mobility affect the performance of plants? The importance of relatedness and geographical proximity. *Journal of Economic Geography*, 9(2):169–190, 2009.

[31] R. Boschma and C. Gianelle. Regional branching and Smart Specialisation policy. *JRC Technical Reports*, (06/2104), 2013.

[32] R. Boschma and S. Iammarino. Related variety, trade linkages, and regional growth in Italy. *Economic Geography*, 85(3):289–311, 2009.

[33] R. Boschma, A. Minondo, and M. Navarro. The emergence of new industries at the regional level in spain: A proximity approach based on product relatedness. *Economic Geography*, 89(1):29–51, 2013.

[34] P. Breathnach, C. van Egeraat, and D. Curran. Regional economic resilience in Ireland: The roles of industrial structure and foreign inward investment. *Regional Studies*, 2(1):497–517, 2015.

[35] J. Brennan and M. Minihan. Market diversification the key to post-Brexit success. *The Irish Times*, 2017.

[36] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998.

[37] Javorcik B.S. Does foreign direct investment increase the productivity of domestic firms? In search of spillovers through backward linkages. *American Economic Review*, 94(3):605–627, 2004.

[38] G. Burger. The proper tool for BEE fronting removal. *Accountancy SA*, pages 38–43, 2014.

[39] R. Burger and R. Jafta. Affirmative action in South Africa: An empirical assessment of the impact on labour market outcomes. CRISE Working Paper 76, 2010.

[40] K.L. Calvert, M.B. Doar, and E.W. Zegura. Modeling internet topology. *IEEE Communications*, 35(6):160–163, 1997.

[41] R. Catini, D. Karamshukb, O. Pennera, and M. Riccaboni. Identifying geographic clusters: A network analytic approach. *Research Policy*, 44:1749–1762, 2015.

[42] M.D. Cattaneo, N. Idrobo, and R. Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press, 2019.

[43] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly. Metrics for community analysis: A survey. *ACM Computing Surveys (CSUR)*, 50(4):1–37, 2017.

[44] Jiyang Chen, Osmar Zaïane, and Randy Goebel. Local community identification in social networks. In *Proceedings of the International Conference on Advances in Social Network Analysis and Mining*, pages 237–242. IEEE, 2009.

[45] D.A. Chimhandamba. *Black economic empowerment and firm competitiveness.* PhD thesis, University of Pretoria, 2010.

[46] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[47] W.M. Cohen and D.A. Levinthal. Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, pages 128–152, 1990.

[48] S.M. Collins. Blacks on the bubble: The vulnerability of Black executives in White corporations. *The Sociological Quarterly*, 34(3):429–447, 1993.

[49] T. Conefrey, G. O'Reilly, and G. Walsh. Modelling external shocks in a small open economy: The case of ireland. *National Institute Economic Review*, 244(1):R56–R63, 2018.

[50] N. Cortinovis, R. Crescenzi, and F. van Oort. Multinational enterprises, industrial relatedness and employment in European regions. *Journal of Economic Geography*, 20(5):1165–1205, 2020.

[51] E. Costenbader and T.W. Valente. The stability of centrality measures when networks are sampled. *Social Networks*, 25(4):283–307, 2003.

[52] R. Crescenzi, L. Gagliardi, and S. Iammarino. Foreign multinationals and domestic innovation: Intra-industry effects and firm heterogeneity. *Research Policy*, 44(3):596–609, 2015.

[53] R. Crescenzi, D. Luca, and S. Milio. The geography of the economic crisis in Europe: National macroeconomic conditions, regional structural factors and short-term economic performance. *Cambridge Journal Of Regions, Economy And Society*, 9(1):13–32, 2016.

[54] R. Crescenzi, C. Pietrobelli, and R. Rabellotti. Innovation drivers, value chains and the geography of multinational corporations in Europe. *Journal of Economic Geography*, 14(6):1053–1086, 2014.

[55] N. Crespo and M.P. Fontoura. Determinant factors of FDI spillovers – What do we really know? *World Development*, 35(3):410–425, 2007.

[56] Z. Csáfordi, L. László, B. Lengyel, and K.M. Kiss. The effect of labor flows, ownership and skill-relatedness on firm productivity. In *Proceedings of International Academic Conferences*, number 4006263. International Institute of Social and Economic Sciences, 2016.

[57] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[58] J.N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing discrete-time finite markov chains. *Journal of Applied Probability*, 2(1):88–100, 1965.

[59] G. de Anda-Jáuregui. Guideline for comparing functional enrichment of biological network modular structures. *Applied Network Science*, 4(1):1–17, 2019.

[60] M. Delgado, M.E. Porter, and S. Stern. Defining clusters of related industries. *Journal of Economic Geography*, 16(1):1–38, 2016.

[61] J.C. Delvenne, S.N. Yaliraki, and M. Barahona. Stability of graph communities across time scales. *Proceedings of the National Academy of Sciences*, 107(29):12755–12760, 2010.

[62] Department of Jobs, Enterprise and Innovation. Ireland's Smart Specialisation Strategy for Research and Innovation. Technical report, Department of Jobs, Enterprise and Innovation, Dublin, Ireland, 2014.

[63] Department of Jobs, Enterprise and Innovation. Building stronger business, 2017.

[64] Department of Public Service and Administration. Green paper on employment and occupational equity: Policy proposals. Working paper, South Africa, 1996.

[65] M. Di Ubaldo, M. Lawless, and I. Siedschlag. Productivity spillovers from multinational activity to indigenous firms in ireland. Technical report, ESRI Working Paper, 2018.

[66] D. Diodato. A network-based method to harmonize data classifications. Technical report, Utrecht University, Department of Human Geography and Spatial Planning, 2018.

[67] D. Diodato, F. Neffke, and N. O'Clery. Why do industries coagglomerate? How marshallian externalities differ by industry and have evolved over time. *Journal of Urban Economics*, 106:1–26, 2018.

[68] D. Diodato and A. Weterings. The resilience of regional labour markets to economic shocks: Exploring the role of interactions among firms and workers. *Journal of Economic Geography*, 15(4):723–742, 2014.

[69] N.P. Dongwana. *The impact and related costs of implementing changes in the Broad-Based Black Economic Empowerment (BBBEE) codes of good practice on companies listed on the Johannesburg Stock Exchange (JSE).* PhD thesis, University of Witwatersrand, Johannesburg, 2016.

[70] C. Donnat and S. Holmes. Tracking network dynamics: A survey of distances and similarity metrics. *ArXiv Preprint 1801.07351*, 2018.

[71] Z. Elekes, R. Boschma, and B. Lengyel. Foreign-owned firms as agents of structural change in regions. *Regional Studies*, 53(11):1603–1613, November 2019.

[72] Enterprise Ireland. Market discovery fund, 2016.

[73] P. Erdős and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[74] R.H. Eriksson. Localized spillovers and knowledge flows: How does proximity influence the performance of plants? *Economic Geography*, 87(2):127–152, 2011.

[75] G. Esping-Andersen and M. Regini. *Why deregulate labour markets?* OUP Oxford, 2000.

[76] J. Essletzbichler. Relatedness, industrial branching and technological cohesion in US metropolitan areas. *Regional Studies*, 49(5):752–766, 2015.

[77] Eurostat. Correspondance table NACE Rev 1.1 to NACE Rev 2. Technical report, European Commission, 2008.

[78] S. Fainshmidt, A. Nair, and M.R. Mallon. MNE performance during a crisis: An evolutionary perspective on the role of dynamic managerial capabilities and industry context. *International Business Review*, 26(6):1088–1099, 2017.

[79] M. Farjoun. Beyond industry boundaries: Human expertise, diversification and resource-related industry groups. *Organization Science*, 5(2):185–199, 1994.

[80] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[81] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[82] U. Fratesi and A Rodríguez-Pose. The crisis and regional employment in Europe: What role for sheltered economies? *Cambridge Journal Of Regions, Economy And Society*, 9(1):33–57, 2016.

[83] K. Frenken and R.A. Boschma. A theoretical framework for evolutionary economic geography: Industrial dynamics and urban growth as a branching process. *Journal of Economic Geography*, 7(5):635–649, 2007.

[84] K. Frenken, F. Van Oort, and T. Verburg. Related variety, unrelated variety and regional economic growth. *Regional Studies*, 41(5):685–697, 2007.

[85] I. Gaddis and S. Klasen. Economic development, structural change, and women's labor force participation: A reexamination of the feminization u hypothesis. *Journal of Population Economics*, 27(3):639–681, 2014.

[86] S. Girma and K. Wakelin. Are there regional spillovers from FDI in the UK? In *Trade, Investment, Migration and Labour Market Adjustment*, pages 172–186. Springer, 2002.

[87] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[88] C. Godsil and G.F. Royle. *Algebraic graph theory*, volume 207. Springer Science & Business Media, 2001.

[89] B.H. Good, Y.A. De Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4):046106, 2010.

[90] H. Görg and D. Greenaway. Much ado about nothing? Do domestic firms really benefit from foreign direct investment? *The World Bank Research Observer*, 19(2):171–197, 2004.

[91] H. Görg and E. Strobl. Multinational companies and productivity spillovers: A meta-analysis. *The Economic Journal*, 111(475):F723–F739, 2001.

[92] H. Görg and E. Strobl. Multinational companies and indigenous development: An empirical analysis. *European Economic Review*, 46(7):1305–1322, 2002.

[93] H. Görg and E. Strobl. Multinational companies, technology spillovers and plant survival. *Scandinavian Journal of Economics*, 105(4):581–595, 2003.

[94] H. Görg and E. Strobl. Foreign direct investment and local economic development: Beyond productivity spillovers. *Does Foreign Direct Investment Promote Development*, pages 137–55, 2005.

[95] H. Görg and E. Strobl. Spillovers from foreign firms through worker mobility: An empirical investigation. *Scandinavian Journal of Economics*, 107(4):693–709, 2005.

[96] O.A. Guerrero and R.L. Axtell. Employment growth through labor flow networks. *PloS One*, 8(5):p.e60808, 2013.

[97] L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE transactions on computer-aided design of integrated circuits and systems*, 11(9):1074–1085, 1992.

[98] P. Hamilton. Market diversification the key to post-Brexit success. *The Irish Times*, 2018.

[99] R.W. Hamming. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160, 1950.

[100] R. Harris and C. Robinson. Foreign ownership and productivity in the united kingdom estimates for UK manufacturing using the ARD. *Review of Industrial Organization*, 22(3):207–223, 2003.

[101] R. Hausmann and C.A. Hidalgo. The network structure of economic output. *Journal of Economic Growth*, 16(4):309–342, 2011.

[102] R. Hausmann, J. Hwang, and D. Rodrik. What you export matters. *Journal of Economic Growth*, 12(1):1–25, 2007.

[103] C.A. Hidalgo, P. Balland, R. Boschma, M. Delgado, M. Feldman, K. Frenken, E. Glaeser, C. He, D.F. Kogler, and A. Morrison. The principle of relatedness. In *Proceedings of the International Conference on Complex Systems*, pages 451–457. Springer, 2018.

[104] C.A. Hidalgo, B. Klinger, A.L. Barabási, and R. Hausmann. The product space conditions the development of nations. *Science*, 317(5837):482–487, 2007.

[105] P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[106] F.M. Horwitz. An analysis of skills development in a transitional economy: the case of the South African labour market. *The International Journal of Human Resource Management*, 24(12):2435–2451, 2013.

[107] S. Iammarino and P. McCann. *Multinationals and economic geography: Location and technology, innovation*. Edward Elgar, Cheltenham, UK, 2013.

[108] G.W. Imbens and T. Lemieux. Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635, 2008.

[109] P. Jaccard. The distribution of the flora in the alpine zone 1. *New Phytologist*, 11(2):37–50, 1912.

[110] A.B. Jaffe. Characterizing the "technological position" of firms, with application to quantifying technological opportunity and research spillovers. *Research Policy*, 18(2):87–97, 1989.

[111] F. Jaumotte. Labour force participation of women. *OECD Economic Studies*, 2003(2):51–108, 2004.

[112] U. Kang and C. Faloutsos. Beyond 'caveman communities': Hubs and spokes for graph compression and mining. In *Proceedings of the 11th International Conference on Data Mining*, pages 300–309. IEEE, 2011.

[113] A. Kerr. Job flows, worker flows and churning in South Africa. *South African Journal of Economics*, 86:141–166, 2018.

[114] J. Kim and G. Marschke. Labor mobility of scientists, technological diffusion, and the firm's patenting decision. *RAND Journal of Economics*, pages 298–317, 2005.

[115] S. Kirkpatrick, C.D. Gelatt Jr, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[116] A. Kokko. Technology, market characteristics, and spillovers. *Journal of Development Economics*, 43(2):279–293, 1994.

[117] A. Kokko, R. Tansini, and M.C. Zejan. Local technological capability and productivity spillovers from FDI in the Uruguayan manufacturing sector. *The Journal of Development Studies*, 32(4):602–611, 1996.

[118] D. Koutra, J.T. Vogelstein, and C. Faloutsos. Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 162–170. SIAM, 2013.

[119] M.C. Kruger and E.P. Kleynhans. Effect of Black Economic Empowerment on profit and competitiveness of firms in South Africa. *Professional Accountant*, 14(1):1–10, 2014.

[120] F.A. Kurtulus. Affirmative action and the occupational advancement of minorities and women during 1973–2003. *Industrial Relations: A Journal of Economy and Society*, 51(2):213–246, 2012.

[121] F.A. Kurtulus. The impact of affirmative action on the employment of minorities and women over three decades: 1973-2003. *Upjohn Institute Working Paper*, 2015.

[122] R. Lambiotte, J.C. Delvenne, and M. Barahona. Dynamics and modular structure in networks. *ArXiv preprint ArXiv:0812.1770*, 2008.

[123] R. Lambiotte and M. Schaub. *Modularity and dynamics on complex networks*. Cambridge University Press, 2021.

[124] R. Lambiotte, R. Sinatra, J.C. Delvenne, T.S. Evans, M. Barahona, and V. Latora. Flow graphs: Interweaving dynamics and structure. *Physical Review E*, 84(1):017102, 2011.

[125] A. Lancichinetti, F. Radicchi, and J.J. Ramasco. Statistical significance of communities in networks. *Physical Review E*, 82(1):046110, 2010.

[126] M. Landabaso, P. McCann, and R. Ortega-Argilés. Smart specialisation in european regions: Issues of strategy, institutions and implementation. *European Journal of Innovation Management*, 2014.

[127] J.S. Leonard. The impact of affirmative action regulation and equal employment law on black employment. *Journal of Economic Perspectives*, 4(4):47–63, 1990.

[128] J. Leskovec, K.J. Lang, and M. Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th International Conference on World Wide Web*, pages 631–640, 2010.

[129] A. Lo Turco and D. Maggioni. Local discoveries and technological relatedness: The role of MNEs, imports and domestic capabilities. *Journal of Economic Geography*, 19(5):1077–1098, 2019.

[130] F. Luo, J.Z. Wang, and E. Promislow. Exploring local community structures in large networks. *Web Intelligence and Agent Systems: An International Journal*, 6(4):387–400, 2008.

[131] J. March. Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 1991.

[132] J.R. Markusen and A.J. Venables. Foreign direct investment as a catalyst for industrial development. *European Economic Review*, 43(2):335–356, 1999.

[133] A. Marshall. *Principles of economics*. Macmillan, London, 1920.

[134] R. Martin. Regional economic resilience, hysteresis and recessionary shocks. *Journal of Economic Geography*, 12(1):1–32, 2012. Publisher: Oxford University Press.

[135] R. Martin and B. Gardiner. The resilience of cities to economic shocks: A tale of four recessions (and the challenge of Brexit. *Papers in Regional Science*, 98(4):1801–1832, 2019.

[136] R. Martin and P. Sunley. On the notion of regional economic resilience: Conceptualization and explanation. *Journal of Economic Geography*, 15(1):1–42, 2015.

[137] N. Masuda, M.A. Porter, and R. Lambiotte. Random walks and diffusion on networks. *Physics Reports*, 716:1–58, 2017.

[138] H. McGuirk, H. Lenihan, and M. Hart. Measuring the impact of innovative human capital on small firms' propensity to innovate. *Research Policy*, 44(4):965–976, 2015.

[139] M. Meilă. Comparing clusterings by the variation of information. In *Learning theory and kernel machines*, pages 173–187. Springer, 2003.

[140] C. Miller. The persistent effect of temporary affirmative action. *American Economic Journal: Applied Economics*, 9(3):152–90, 2017.

[141] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.

[142] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[143] K. Mzilikazi. *The impact of Broad Based Black Economic Empowerment compliance on profitability of companies listed in the Johannesburg Stock Exchange: a cross industry analysis*. PhD thesis, University of Witwatersrand, Johannesburg, 2016.

[144] S. Nagpal, K.D. Baksi, B.K. Kuntal, and S.S. Mande. NetConfer: A web application for comparative analysis of multiple biological networks. *BMC Biology*, 18:1–12, 2020.

[145] F. Neffke, M. Hartog, R. Boschma, and M. Henning. Agents of structural change: the role of firms and entrepreneurs in regional diversification. *Economic Geography*, 94(1):23–48, 2018.

[146] F. Neffke and M. Henning. Skill relatedness and firm diversification. *Strategic Management Journal*, 34(3):297–316, 2013.

[147] F. Neffke, M. Henning, and R. Boschma. How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. *Economic Geography*, 87(3):237–265, 2011.

[148] F. Neffke, M. Henning, R. Boschma, K.J. Lundquist, and L.O. Olander. The dynamics of agglomeration externalities along the life cycle of industries. *Regional studies*, 45(1):49–65, 2011.

[149] F. Neffke, A. Otto, and A. Weyh. Inter-industry labor flows. *Journal of Economic Behavior & Organization*, 142(1):275–292, 2017.

[150] R.R. Nelson and S.G. Winter. *An evolutionary theory of economic change*. The Belknap Press of Harvard University Press, Cambridge, 1982.

[151] M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[152] M.E.J. Newman. The physics of networks. *Physics Today*, 61(11):33–38, 2008.

[153] M.E.J. Newman, A.L. Barabási, and D.J. Watts. *The structure and dynamics of networks.* Princeton university press, 2006.

[154] M.E.J. Newman, G.T. Cantwell, and J.G. Young. Improved mutual information measure for clustering, classification, and community detection. *Physical Review E*, 101(4):042304, 2020.

[155] M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[156] J.D. Noh and H. Rieger. Random walks on complex networks. *Physical Review Letters*, 92(11):118701, 2004.

[157] B. Nooteboom. Innovation and inter-firm linkages: New implications for policy. *Research Policy*, 28(8):793–805, 1999.

[158] N. O'Clery. A tale of two clusters: The evolution of ireland's economic complexity since 1995. *Science*, 317:482–487, 2016.

[159] N. O'Clery, J.C. Chaparro, A. Gomez-Lievano, and E. Lora. Skill diversity as the foundation of formal employment creation in cities. Technical report, Working Paper at Center for International Development at Harvard, 2018.

[160] N. O'Clery, R.P. Curiel, and E. Lora. Commuting times and the mobilisation of skills in emergent cities. *Applied Network Science*, 4(1):118, 2019.

[161] N. O'Clery and S. Kinsella. Modular structure in labour networks reveals skill basins. *Research Policy*, 51(5):104486, 2022.

[162] S. O'Connor, E. Doyle, and S. Brosnan. Clustering in Ireland: Development cycle considerations. *Regional Studies*, 4(1):263–283, 2017.

[163] W.Y. Oi and T.L. Idson. Firm size and wages. *Handbook of Labor Economics*, 3:2165–2214, 1999.

[164] E. O'Leary and C. van Egeraat. Introduction: Rethinking irish economic development. *Administration*, 66(1):85–87, 2018.

[165] J.P. Onnela, D.J. Fenn, S. Reid, M.A. Porter, P.J. Mucha, M.D. Fricker, and N.S. Jones. Taxonomies of networks from community structure. *Physical Review E*, 86(3):036104, 2012.

[166] M. Oosthuizen. The post-apartheid labour market: 1995-2004. Working Paper 06/103, World Institute for Development Economic Research (UNU-WIDER), 2006.

[167] G.K. Orman, V. Labatut, and H. Cherifi. Comparative evaluation of community detection algorithms: A topological approach. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(08):P08001, 2012.

[168] Neave O'Clery, Samuel Heroy, François Hulot, and Mariano Beguerisse-Díaz. Unravelling the forces underlying urban industrial agglomeration. In *Handbook of Cities and Networks*, pages 472–492. Edward Elgar Publishing, 2021.

[169] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[170] L.L. Pasinetti. Structural change and economic growth: a theoretical essay on the dynamics of the wealth of nations. 1983.

[171] L. Peel and A. Clauset. Detecting change points in the large-scale structure of evolving networks. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[172] L. Peel, D.B. Larremore, and A. Clauset. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548, 2017.

[173] D. Pieterse, E. Gavin, and C.F. Kreuser. Introduction to the South African Revenue Service and National Treasury Firm-Level Panel. *South African Journal of Economics*, 86:6–39, 2018.

[174] S. Pinch and N. Henry. Paul krugman's geographical economics, industrial clustering and the british motor sport industry. *Regional Studies*, 33(9):815–827, 1999.

[175] F. L. Pinheiro, A. Alshamsi, D. Hartmann, R. Boschma, and C.A. Hidalgo. Shooting low or high: Do countries benefit from entering unrelated activities? *Papers in Evolutionary Economic Geography*, 18(07), 2018.

[176] L. Pizzati and B. Funck. *Labor, employment, and social policies in the EU enlargement process: Changing perspectives and policy options.* World Bank Publications, 2002.

[177] P. Pons and M. Latapy. Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, pages 284–293. Springer, 2005.

[178] M.A. Porter, P.J. Mucha, M.E.J. Newman, and A.J. Friend. Community structure in the united states house of representatives. *Physica A: Statistical Mechanics and its Applications*, 386(1):414–438, 2007.

[179] M.E. Porter. *Clusters and the new economics of competition*, volume 76. Harvard Business Review Boston, 1998.

[180] M.E. Porter. *Competitive advantage of nations: Creating and sustaining superior performance*, volume 2. Simon and Schuster, 2011.

[181] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[182] W.K. Roche, P.J. O'Connell, and A. Prothero. *Austerity and recovery in Ireland: Europe's poster child and the Great Recession.* Oxford University Press, 2016.

[183] M. Rosvall, D. Axelsson, and C.T. Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

[184] M. Rosvall and C.T. Bergstrom. Mapping change in large networks. *PloS one*, 5(1):e8694, 2010.

[185] M.T. Schaub, J.C. Delvenne, R. Lambiotte, and M. Barahona. Structured networks and coarse-grained descriptions: A dynamical perspective. *Advances in Network Clustering and Blockmodeling*, pages 333–361, 2019.

[186] M.T. Schaub, J.C. Delvenne, M. Rosvall, and R. Lambiotte. The many facets of community detection in complex networks. *Applied Network Science*, 2(1):1–13, 2017.

[187] M.T. Schaub, J.C. Delvenne, S.N. Yaliraki, and M. Barahona. Markov dynamics as a zooming lens for multiscale community detection: Non clique-like communities and the field-of-view limit. *PloS One*, 7(2), 2012.

[188] M.T. Schaub, N. O'Clery, Y.N. Billeh, J.C. Delvenne, R. Lambiotte, and M. Barahona. Graph partitions and cluster synchronization in networks of oscillators. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(9):094821, 2016.

[189] J. Simmie and R. Martin. The economic resilience of regions: Towards an evolutionary approach. *Cambridge Journal of Regions, Economy and Society*, 3(1), 2010.

[190] H.A. Simon and A. Ando. Aggregation of variables in dynamic systems. *Econometrica*, pages 111–138, 1961.

[191] J. Song, P. Almeida, and G. Wu. Learning–by–hiring: When is mobility more likely to facilitate interfirm knowledge transfer? *Management Science*, 49(4):351–365, 2003.

[192] N. Spaull and H. Van Broekhuizen. The "Martha Effect": The compounding female advantage in South African higher education. Working Paper 14, Stellenbosch University, 2017.

[193] D. Straulino, M.S. Landman, and N. O'Clery. A bi-directional approach to comparing the modular structure of networks. *EPJ Data Science*, 10(1):13, 2021.

[194] I. Szakálné Kanó, B. Lengyel, Z. Elekes, and I. Lengyel. Agglomeration, foreign firms and firm exit in regions under transition: The increasing importance of related variety in Hungary. *European Planning Studies*, 27(11):2099–2122, 2019.

[195] M. Tantardini, F. Ieva, L. Tajoli, and C. Piccardi. Comparing methods for comparing networks. *Scientific Reports*, 9(1):1–19, 2019.

[196] The Department of Labour. The employment equity act. The South African labour law manual, South Africa, 1998.

[197] The Public Service Commission. Gender mainstreaming initiative in the public service. *Pretoria: PSC*, 2006.

[198] V.A. Traag, L. Waltman, and N.J. Van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

[199] L.N. Trefethen. *Approximation theory and approximation practice*, volume 128. SIAM, 2013.

[200] J.C. Visagie. The influence of affirmative action on SMME culture in south africa. *Participation and Empowerment: An International Journal*, 1999.

[201] D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

[202] J.J Whang, D.F. Gleich, and I.S. Dhillon. Overlapping community detection using seed set expansion. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 2099–2108, 2013.

[203] J. Xiao, R. Boschma, and M. Andersson. Resilience in the European Union: The effect of the 2008 crisis on the ability of regions in Europe to develop new industrial specializations. *Industrial and Corporate Change*, 27(1):15–47, 2018.

[204] J. Xie, S. Kelley, and B.K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):1–35, 2013.

[205] J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.

[206] Y.W. Yu, J.C Delvenne, S.N. Yaliraki, and M. Barahona. Severability of mesoscale components and local time scales in dynamical networks. *ArXiv Preprint ArXiv:2006.02972*, 2020.

[207] S.A. Zahra and G. George. Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review*, 27(2):185–203, 2002.