



**University of Cagliari**

**PhD in Physics**

# Statistical Physics of Network Communities in Economic Systems

FIS/03

**Tutor:**

Prof. Alberto Devoto

**Co-tutor:**

Prof. Paolo Ruggerone

**Coordinator:**

Prof. Paolo Ruggerone

**PhD Candidate:**

Federica Cerina

**XXVII Cycle - Academic year 2013/2014**

# Introduction

In the last decade, the study of big networked systems has received a great deal of attention thanks to the increased availability of large datasets and the technology to analyze them. At the present day, complex networks have made their official debut in many fields: brain is made of millions of interconnected neurons; ecosystems consists of several species whose interdependencies can be mapped onto a network; social systems can be well represented by graphs whose edges describe the interactions between individuals and so on.

The dynamic nature and the big sizes of complex networks have attracted the attention of the physicists, who are currently contributing to the modeling of 'complex systems' by using tools and methodologies developed in statistical mechanics and theoretical physics.

During the past 30 years, physicists have achieved important results in the field of phase transitions, statistical mechanics, nonlinear dynamics, and disordered systems. In these fields, power laws, scaling, and unpredictable (stochastic or deterministic) time series are present and the current interpretation of the underlying physics is often obtained using these concepts.

Statistical mechanics has long studied how the interactions happening on a microscopical scale can affect a system's macroscopical behaviour. The re-

cent discovery of the importance of network representation has given a boost to the study of models whose most relevant aspect lies in the "universality" of the critical exponents: indeed, many phenomena, despite the difference regarding the nature of the single components, share a common set of critical exponents. Scale invariance allows to explain how, near the transition, the macroscopical behaviour of a system does not depend on the form of the microscopical interaction but just on the *dimensionality* of the system itself and order parameters' simmetries.

In particular, many complex systems are organized as networks into space. Transports, Internet, telecommunications and social networks are all examples of networks where space plays a relevant role and topology alone cannot convey all the information. Thus, it is easy to see how understanding and charachterizing the structure of such networks is crucial in many areas, ranging from urban studies to epidemiology. An important consequence of spatial dependency in networks is, for instance, the cost (in terms of money or other) associated to the length of the connections, heavily affecting network topology itself.

Thank to the sophisticated techniques of geo-localization we are now able to trace people and goods movements but, to extract meaningful information from this enormous quantity of data on mobility, we need appropriate tools. In this context, the study of the effect of space in networks has undertaken the important role to unravel regularities and behaviours from data and supply suitable models.

This holds true especially for community detection. Finding meaningful communities in a networks is still a difficult task but essential to unveil

functional relations between the parts. In spatial networks this is even more tricky since communities can be affected by spatial propinquity or other geographical factors.

Standard community detection suffers from various drawbacks, one of them being the community definition that does not provide any information about the importance of a node inside its own community. Nodes of a community do not have all the same importance for the community stability: the removal of a node in the "core" of a network affects the partition much more than the deletion of a node in the periphery.

This work intends to study community detection and develop new methods and algorithms to be applied to the study of the global market and its functioning, in order to understand the causes of critical events, like the recent global financial crisis.

Economics systems, indeed, exhibit several of the properties that characterize complex systems and seem to be an ideal testing ground. They are open systems in which many sub-units interact nonlinearly in the presence of feedback. In these systems, the governing rules are rather stable and the time evolution of the system is continuously monitored. It is now possible to develop models and to test their accuracy and predictive power using available data, since large databases exist even for high-frequency data. Even if the correlation may not be blatant, space is not to be looked over in economy, too. It has been proved, indeed, that trade and economic exchanges between countries are necessarily driven by geographical proximity or impeded by natural barriers. Also, space provides the framework for all data collections, samples and surveys. One could say that geography provides context to sta-

tistical data.

Recently, a growing number of physicists have attempted to analyze and model financial markets and, more generally, economic systems. The interest of this community in economic systems has roots that date back to 1936, when Majorana wrote a pioneering paper on the essential analogy between statistical laws in physics and in the social sciences (translation from the original italian paper in Majorana and Mantegna (2006)).

This unorthodox point of view was considered of marginal interest until recently. Since 1990, the physics research activity in this field has become less episodic and a research community has begun to emerge. The research activity of this group of physicists is complementary to the most traditional approaches of economy and mathematical finance. One characteristic difference is the emphasis that physicists put on the empirical analysis of economic data. Another is the background of theory and method in the field of statistical physics developed over the past 30 years that physicists bring to the subject. The concepts of scaling, universality, disordered frustrated systems, and self-organized systems might be helpful in the analysis and modeling of economic systems (Mantegna et al. (2000)).

Before discussing the main part of the work, some brief digressions are needed: **Chapter 1** will provide an introduction to graph theory and some basic definitions; **Chapter 2** will give a rapid overview on community detection and, in particular, modularity based methods while **Chapter 3** will describe at length the data used in our work. In **Chapter 4** we will thoroughly show results and applications and, finally, in **Chapter 5** we will discuss results, draw conclusions and set the future work.

# Contents

# Chapter 1

# Introduction to graphs

In general terms, a *network* is defined as any system that can be represented as a mathemathical abstract object called *graph*, whose *vertices* (or *nodes*) are the elements of the system and its *edges* (or *links*) identify the relations between them.

It's clear to see how such a definition applies to a variety of cases and, in that sense, networks are a valid and convenient method for representing relations in complex systems, where the interactions occur among a great number of players.

Graph teory dates back to Euler's solution of the puzzle of Königsberg's bridges in 1736. Since then a lot has been learned about graphs and their mathematical properties. In the 20th century they have also become extremely useful as representation of a wide variety of systems in different areas. Biological, social, technological, and information networks can be studied as graphs, and graph analysis has become crucial to understand the features of these systems (Fortunato (2010)). Even if graph theory has been used in

different areas, it comes with a set of basic common definitions. The following sections will provide the basic notions allowing to study and describe networks.

## 1.1 Basic definitions

### 1.1.1 Graphs

A graph **G** is defined giving a set of vertices and connections between them. Mathematically, $G = G(N, M)$ where $N$ is the total number of vertices and $M$ the total number of edges. Edge $(i, j)$ connects vertices $i$ and $j$ that are called **adjacent** (or **neighbors**).

Edges may have arrows or not, that is they can be traversed in one direction only. In this case the graph is an **oriented** graph. A further generalization is also possible: one can imagine that a value is assigned to every edge. In the case of transportation networks (a system of pipelines or the Internet cables) this could represent for example the maximum load allowed. Whenever this extra information is provided we deal with a **weighted** graph.

The number of nodes $N$ and the number of edges $M$ are not independent of each other. If we assume to have only one edge between two vertices there is a maximum number of edges we can draw. Consider that each vertex can establish an edge only with $(N - 1)$ different vertices (and not with itself). This holds for every one of the n vertices. This give a total number of $N(N - 1)$ possibilities counting every edge twice. The maximum number of edges is exactly one half of that $M_{max} = N(N - 1)/2$. If the starting and ending vertices make a difference (as in the case of oriented graph) then we

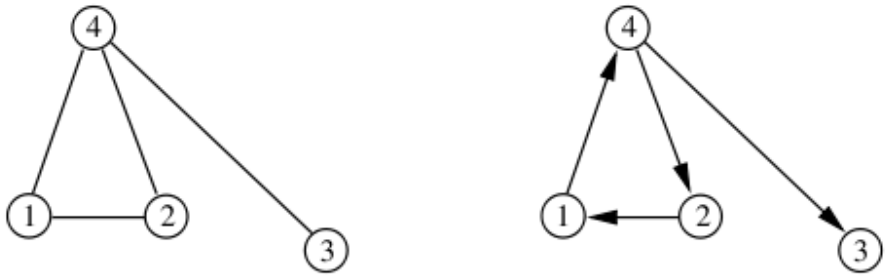
do not have to divide by two the above quantity. In this case the maximum number of edges is given simply by $N(N-1)$.

Two immediate limits are present. If no edge is drawn then the graph is empty and it is indicated by $E^N$ . If all the edges are drawn, the graph is **complete** and it is indicated by $K^N$.

### 1.1.2 Adjacency matrix

The structure of the graph $G(N, M)$ can be represented by means of a matrix. In the case of graphs we introduce the Adjacency Matrix $A(N, N)$ whose entries $A_{ij}$ are 0 if vertices $i$, $j$ are not connected and 1 otherwise.

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$




.

Figure 1.1: Two simple graphs and their adjacency matrices. Note that for oriented graph (right) the matrix is not symmetric (Caldarelli (2007))

The diagonal elements represent the presence of an edge between a vertex

and itself (**self-link**). Unless specified otherwise, we consider those entries equal to 0. If edges with itself are allowed (diagonal elements different from 0) then we have N more allowed edges. Therefore the maximum number of edges becomes $M_{max} = N(N-1)/2 + N = N(N+1)/2$.

Note that this matrix is symmetric (meaning $A_{ij} = A_{ji}$ ) only in the case of non-oriented graphs. For oriented graphs instead the elements $A_{ij}$ are generally different from the elements $A_{ji}$. This representation well extends to the case of the weighted graphs: instead of giving only 1 and 0, we assign a real number (the weight) to the entry $A_{ij}$. We obtain then an adjacency matrix composed by real numbers for the edges present and 0 otherwise. In the following we refer to this matrix with the symbol$A^W(N,N)$. Its elements will be then indicated by $a^w_{ij}$.

### 1.1.3 Basic quantities

**Order and size.** The **order** of a graph is the number of vertices $N$ while its **size** is defined as the number of edges $M$.

**Degree.** The **degree k** of a vertex is the number of its edges. As mentioned before, the sum of all the degrees in the graph is twice the number of the edges in the graph. This happens because any edge contributes to the degree of the vertex origin and to the degree of vertex destination.

A compact way to compute the degree consists in running on the different columns of a fixed row in the adjacency matrix $A(N,N)$ looking for all the

1's present. This means that the degree $k_i$ of a vertex $i$ can be computed as:

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{1.1}$$

In oriented graphs this quantity splits in **in-degree** $k_i^{in}$ and **out-degree** $k_i^{out}$ for edges pointing in and out respectively. Since in the adjacency matrix the $A_{ij}$'s are different from the $A_{ji}$ 's we have that $A_{ij} = 1$ if and only if an edge goes from $i$ to $j$. This means that:

$$k_i^{in} = \sum_{j=1}^{N} A_{ji} \tag{1.2}$$

$$k_i^{out} = \sum_{j=1}^{N} A_{ij} \tag{1.3}$$

For weighted graphs an extension of the degree is made by summing the weights of the edges rather than their number. In this case the degree it is called **strength**.

$$k_i^{w} = \sum_{j=1}^{N} A_{ij}^{w} \tag{1.4}$$

**Distance.** The **distance** $d_{ij}$ between two vertices $i$, $j$ is the shortest number of edges one needs to travel to get from $i$ to $j$. Therefore the neighbours of a vertex are all the vertices which are connected to that vertex by a single edge.

Using again the adjacency matrix properties, one can define distance as:

$$d_{ij} = min\{\sum_{k,l \in \mathcal{P}_{ij}} a_{kl}\} \tag{1.5}$$

where $\mathcal{P}_{ij}$ is a path connecting $i$ to $j$.

If the graph is oriented one has to follow the direction of the edges. Therefore the distances are generally larger than in the homologous non-oriented graphs. In the case of weighted graphs, instead of summing for every step a distance of 1 we can assume that the distance is related to the values of the weight.

Related to the distance is the **diameter D** of a graph that can be defined as the largest distance you can find between two vertices in the graph (Caldarelli (2007)). Some other definition (as the average distance) are possible.

### 1.1.4 Centrality measures

The importance of a node is usually defined as its *centrality*. There exist several measures to characterize the centrality of a node but the most used are the following:

**Degree centrality.** It's probably the most intuitive and immediate centrality measure. It is based on node degree and it says how well a node is connected to the others elements of the graph.

**Closeness Centrality.** In a graph, the **farness** of a node is defined as the sum of its distances to all other nodes, and its closeness is defined as the reciprocal of the farness as follows:

$$g_i = \frac{1}{\sum_{j \neq i} d_{ij}} \tag{1.6}$$

According to this measure, nodes having a smaller $g_i$ have a higher centrality value.

**Betweenness centrality.** While degree and closeness centrality take into account only the topological role the node plays in the graph, completely ignoring eventually crucial nodes that could be acting as bridges between different parts of the netwrk, betweenness centrality quantifies the number of times a node acts as a bridge along the *shortest path* between two other nodes. If $\sigma_{hj}$ is the number of shortest paths from $h$ to $j$ and $\sigma_{hj}(i)$ is the fraction of these shortes paths that pass through vertex $i$, the the betweenness centrality is defined as:

$$b_i = \sum_{h \neq j \neq i} \frac{\sigma_{hj}(i)}{\sigma_{hj}} \tag{1.7}$$

Betweenness centrality is often used in trasportation networks in order to extimate the traffic load each node can withstand and determine the most central ones. The nodes with the highest centrality are also the most important in the graph because they keep the graph connected.

**Eigenvector centrality.** It is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. The centrality score of vertex $i$ in graph G is:

$$x_i = \frac{1}{\lambda} \sum_{j \in G} A_{ij} x_j \tag{1.8}$$

where $\lambda$ is a constant and $A_{ij}$ is the adjacency matrix entry.

**Pagerank.** It is an algorithm used by Google Search to rank websites in their search engine results but it is now one of the most successful and widely used. The idea is that in directed graphs the nodes are considered important if they are pointed by other important nodes. PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. To do so it requires to be computed iteratively:

At t = 0, an initial probability distribution is assumed, usually $PR(i;0) = \frac{1}{N}$ where $N$ is the total number of nodes;

At each time step, the PageRank of node $i$ is computed as:

$$PR(i, t+1) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{PR(j,t)}{L(j)} \tag{1.9}$$

where $M(i)$ are the in-neighbors of node $i$ and $L(j)$ is the number of ougoing links from node $j$. Damping factor $d$, usually set to 0.85, ensures that the random walker doesn't get stuck in one node indefinitely.

### 1.1.5 Clustering coefficient

Even the neighbourhood of a node can determine its properties. That is the case of **clustering coefficient** $C_i$ which takes into acccount the number of edges near a vertex $i$. $C_i$ is given by the average fraction of pairs of neighbours (of the same vertex) that are also neighbours of each other (see figure 1.2).

The maximum value of $C_i = 1$ for every vertex $i$ is obtained for the completely connected clique. In general we may write the clustering coefficient as the fraction of actual edges over the possible ones between the vertices $i$,

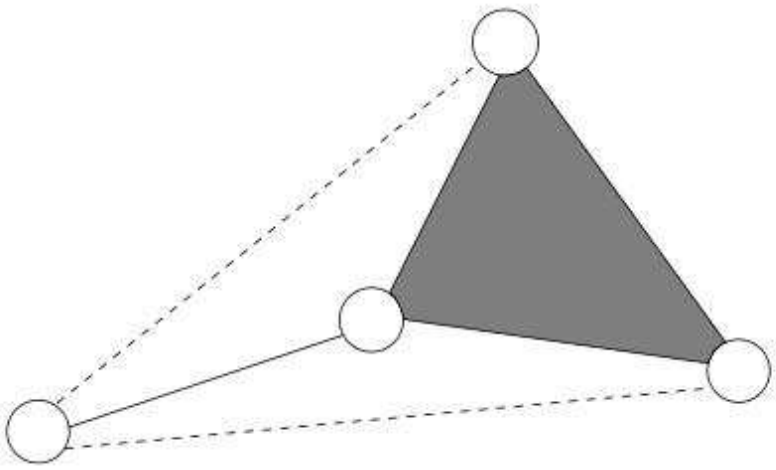


Figure 1.2: In this simple example the central vertex has 3 neighbors. These nodes can be connected in three different ways but, since only one of them is actually realized, the clustering coefficient of the central node is $C_i = 1/3$ (Caldarelli (2007)).

$j$, $k$. Using, again, the adjacency matrix formalism, clustering coefficient can be written as:

$$C_i = \frac{1}{k_i(k_i - 1)/2} \sum_{j,k} a_{ij} a_{ik} a_{jk} \tag{1.10}$$

The average clustering coefficient $C$ is defined as the average of the $C_i$'s over the vertices $i$ of the graph.

If the graph is oriented, the generalization of the clustering coefficient is not trivial. The more natural solution could be to separate in- and out-degree contributes but then the problem would be to decide which direction has to be considered. In general one tends to join the two possible direction such

that the clustering coefficient takes the form:

$$C_i^{in} = \frac{1}{k_i^{in}(k_i^{in}-1)/2} \sum_{j,k} a_{ji} a_{ki} \frac{(a_{jk}+a_{kj})}{2} \tag{1.11}$$

$$C_i^{out} = \frac{1}{k_i^{out}(k_i^{out}-1)/2} \sum_{j,k} a_{ji} a_{ki} \frac{(a_{jk}+a_{kj})}{2} \tag{1.12}$$

The weighted case is even less straightforward. Indeed, if it is relatively easy to convert the numerator in these expressions, it is not equally simple to do it with the denominator. In the oriented case one only needs to add one edge but in the weighted case one should also assign a weight to the added edge.

Several solutions have been proposed. One definition is the following:

$$C_i^w = \frac{1}{< a^w >^3 k_i(k_i-1)/2} \sum_{j,k} a_{ij}^w a_{ik}^w a_{jk}^w \tag{1.13}$$

where $< a^w >= \frac{1}{n} \sum_{ij} a_{ij}^w$ is the average weight of an edge in the graph. Another possible solution, useful in real situations where fluctuactions in edge weights are crucial, is:

$$C_i^w = \frac{1}{k_i^w(k_i-1)/2} \sum_{j,k} \frac{a_{ij}+a_{ik}}{2} \theta(a_{ij}^w)\theta(a_{ik}^w)\theta(a_{jk}^w) \tag{1.14}$$

$$C_i^w = \frac{1}{k_i^w(k_i-1)/2} \sum_{j,k} \frac{a_{ij}+a_{ik}}{2} \theta(a_{ij}^w)\theta(a_{ik}^w)\theta(a_{jk}^w) \tag{1.15}$$

where $\theta(x)$ is the step function equal to 1 when the argument is larger than 0 (Barrat, Barthèlemy, Pastor-Satorras and Vespignani, 2004a; Barrat, Barthèlemy, Pastor-Satorras and Vespignani, 2004b).

According to the situation, one definition or another can be more suitable (Caldarelli (2007)).

## 1.2 Trees

There is one general case in which the networks have a particular characteristic shape. In the case of a distribution network (as for example water supply, but in principle anything), the good is delivered to all clients trying to avoid to pass on the same vertex twice. The class of graphs for which this holds are called *trees*.

A **tree** is then a graph without cycles, where a cycle is defined as a closed path that visit each vertex only once. A set of disconnected trees is called a **forest**.

For the oriented trees the vertices with (out-)in-degree equal to one (the peripheral vertices of the tree) are called **leaves**. Sometimes it is useful to define a special vertex that is called **root**.

On a tree we can still use the quantities defined for graphs. Vertices have still a degree and we can measure also the distances. In a non-oriented tree there is always a path between any pair of vertices. For oriented trees instead, it is possible that some of the vertices are isolated from the others because the direction of the edges does not allow to join them. Therefore distances are generally larger in oriented graphs. The only exception is the clustering coefficient that is zero by construction.

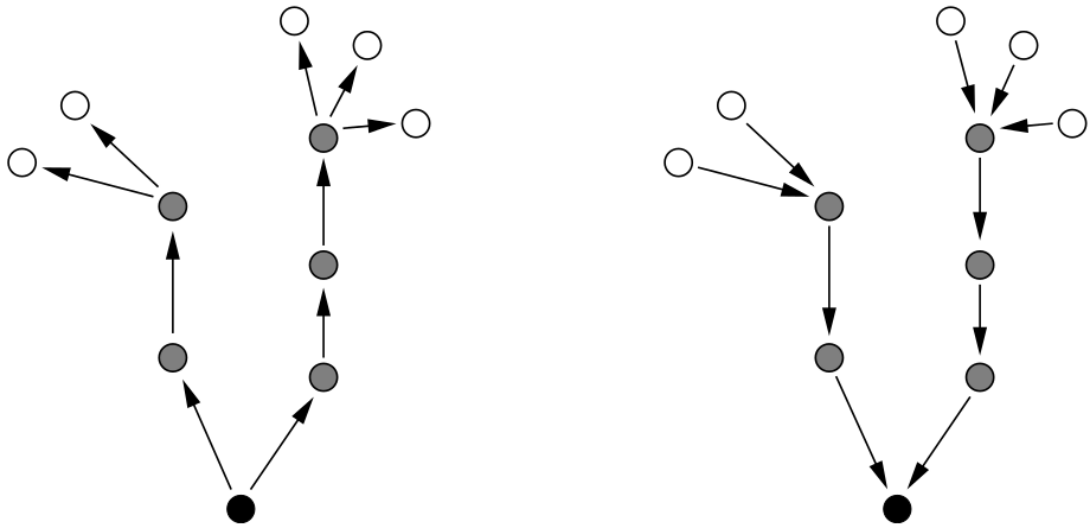


Figure 1.3: Two examples of special vertices in a tree. On the left (as in a real tree) nutrients flow from the root (dark) to reach the leaves (light). Root and leaves are defined through their in-degree (Caldarelli (2007)).

### 1.2.1 Classification of trees

Despite being so simple, trees can be very useful in describing complex objects. Trees can be *real*, i.e. the acyclical structure is intrinsic to the physical phenomenon (e.g. rivers network), or can be derived from a graph as *spanning trees*. Suppose that from a vertex in the graph you want to reach rapidly all the other vertices. Then, from a starting point (i.e. vertex $i$) one finds all the first neighbours and write them in a list; all of them are at a distance one from $i$. For every of these vertices one computes a second list made of their neighbours provided they are not already in the first list and they are not $i$. all the vertices in the second list are at distance two from $i$. Iterating the procedure one finds different shells of vertices around $i$, until all the vertices

are checked. This algorithm automatically produces a spanning tree (whose root is $i$) for the graph considered.

A **minimum spanning tree** can be defined as a spanning tree with weight less than or equal to the weight of every other spanning tree. More generally, any undirected graph (not necessarily connected) has a minimum spanning forest, which is a union of minimum spanning trees for its connected components.

The minimum spanning tree can be considered as a minimum-cost subgraph connecting all vertices, since subgraphs containing cycles necessarily have more total weight. Let's consider a telecommunication company laying new cables along the road; then there would be a graph representing which points are connected by those paths. Of course some of these paths are more expensive than others, so the minimum spanning tree would be the subset of these paths with the lowest total cost and would represent the least expensive path for laying the cable.

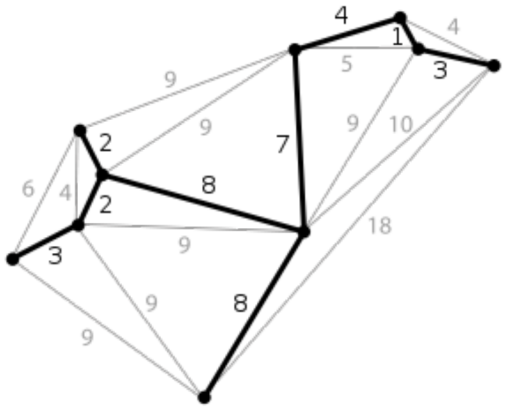



Figure 1.4: The only minimum spanning tree of a planar graph. Each edge is labeled with its weight, which here is roughly proportional to its length.

## 1.3 Statistical characterization of the network

When dealing with very large networks (large in some case means millions of vertices and many more edges.) it is not possible to study their elements or properties locally but it is necessary to adopt statical methods to be able to take into account their global behaviour. Some quantities play a crucial role in that sense: **degree distribution**, **distance distribution** e **assortativity**.

### 1.3.1 Degree distribution $P(k)$

Whichever the interest area, be it biology, technology, social sciences or physics, all these structures show the same statistical property: the Probability Distribution for the degree decays as a power law.

$$P(k) \propto x^{-\gamma} \tag{1.16}$$

This result indicates that large networks self-organize into a scale-free state, a feature unexpected by all existing random network models. This feature is found to be a consequence of the two generic mechanisms that networks expand continuously by the addition of new vertices (**growth**), and new vertices attach preferentially to already well connected sites (**preferential attachment**) (Barabasi and Albert (1999)).

This sort of "universality" means that the form of the distribution is similar in all these cases and that the system appears the same regardless the level at which one looks at it. This has profound implications, being the most important the **Scale invariance** (Barabasi and Albert (1999)).

### 1.3.2 Distance distribution $P(d)$

In most cases, even the distance distribution is the same, In particular its peak is around small values of $d$. This average value of the vertex-vertex distance is supposed to depend logarithmically on the number n of vertices in the network. This effect is known as **Small World** effect since in the social graphs where vertices represent individuals, a little number of relationships (edges) can connect two parts of the graph.

### 1.3.3 The Correlation between degrees: Assortativity

Another typical feature of the scale-free network is the tendency of vertices of a certain degrees to be connected with other vertices with similar (assortative) or dissimilar (disassortative) degree. Assortativity is often viewed as a correlation between two nodes.

The **assortativity coefficient r** is the Pearson correlation coefficient of the degree between pairs of linked nodes. Positive values of $r$ indicate a correlation between nodes of similar degree, while negative values indicate relationships between nodes of different degree. In general, $r$ lies between -1 and 1. When $r = 1$, the network is said to have perfect assortative mixing patterns, when $r = 0$ the network is non-assortative, while at $r$ = -1 the network is completely disassortative.

The assortativity coefficient is given by:

$$r = \frac{\sum_{k_1 k_2} k_1 k_2 (P(k_1, k_2) - P(k_1)P(k_2))}{\sigma^2} \tag{1.17}$$

where $P(k_1)$ and $P(k_2)$ are the degree distributions of nodes $k_1$ and $k_j$,

$P(k_1, k_2)$ is the joint probability distribution of the remaining degrees of the two vertices and $\sigma^2$ is the variance of the $P(k)$.

# Chapter 2

# Community Detection

In this chapter, we consider another property, which, as we will show, appears to be common to many networks, the property of *community structure*.

It is a matter of common experience that some networks seem to have subsets of vertices within which vertex-vertex connections are dense, but with few connections among these subsets (Girvan and Newman (2002)). Such different structures seem to suggest a natural subdivision of the network in **communities** or modules(2.1) [1].

Therefore communities are group of nodes sharing some common properties and/or play similar roles within the graph (Fortunato (2010)). Com-

[1]Communities are sometimes called *clusters* but they have a slightly different meaning: a cluster is a part of the graph where there are more internal links than external ones while a community is a set of vertices sharing the same topological properties. Nonetheless, if a set of vertices has common edges, it is not only a cluster but also a community since even edges can be cosidered a network property. In summary, a cluster is always associated with a community of some kind. Communities usually correspond to clustered subgraphs (Caldarelli (2007)). Therefore, in the following the two definitions will be considered equal.

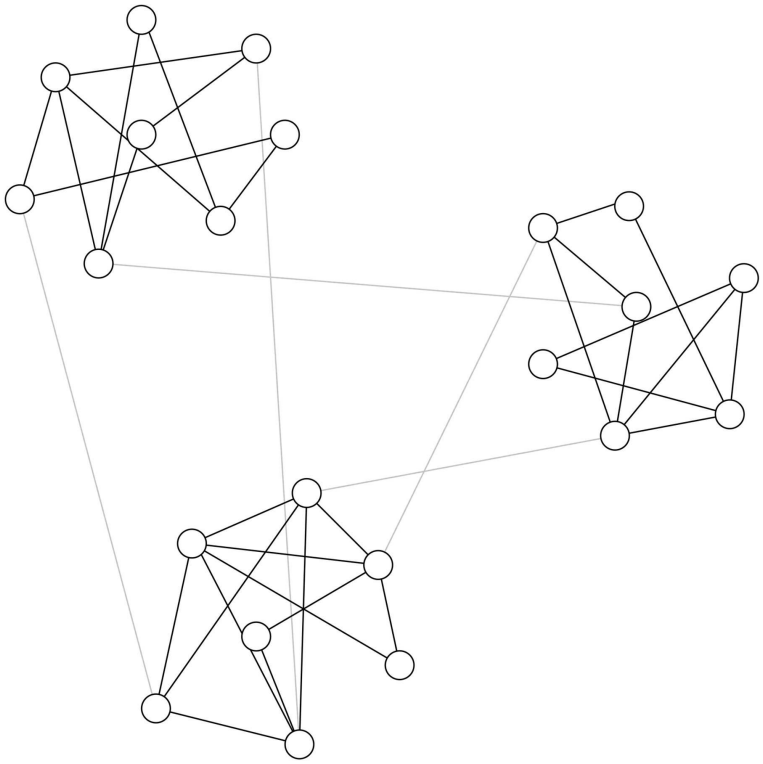


Figure 2.1: A schematic representation of a network with community structure. In this network there are three communities of densely connected vertices (circles with solid lines), with a much lower density of connections (gray lines) between them. (Girvan and Newman (2002)).

munities in a social network might represent real social groupings, perhaps by interest or background; communities in a citation network might represent related papers on a single topic; communities in a metabolic network might represent cycles and other functional groupings; communities on the web might represent pages on related topics.

Identifying modules and their boundaries allows for a classification of vertices, according to their structural position in the modules. So, vertices

with a central position in their clusters, i. e. sharing a large number of edges with the other group partners, may have an important function of control and stability within the group; vertices lying at the boundaries between modules play an important role of mediation and lead the relationships and exchanges between different communities. This is particularly true in metabolic (Lewis et al. (2010)) and social (Porter et al. (2009)) networks

Being able to identify these communities could help us to understand and exploit these networks more effectively (Girvan and Newman (2002)).

The aim of **community detection** in graphs is to identify the modules and, possibly, their hierarchical organization, by only using the information encoded in the graph topology.

## 2.1 Communities in real-world networks

Relations between nodes in real networks are not necessarily mutual; in most cases edges have a precise orientation and the networks are said to be directed. World Wide Web, for example, can be seen as a graph whose nodes are web pages and hyperlinks make the users surf form a page to another. Links are directed: usually, if a link from A to B exists, doesn't exist a link that gets back to A from B. Less than 10% of links are reciprocal.

Links direction obviously provides precious informations on the system; that is why taking it into account significantly improves the quality of the partitioning. Nonetheless, to develop community detection methods for directed neworks is quite difficult because not all the methods that are currently used can be extended to the directed case.

Real networks can also have other features that add difficulty to the picture. In many networks, in fact, some nodes don't belong exclusively to one community but can be shared with more than one. In such cases as these we talk about *overlapping communities*. Traditional community detection algorithms assign each node to a single community, but by doing so they neglect part of information since shared nodes are the ones that most probably act as the intermediary between the different compartments of the graph.

Community structure can be also *hyerarchical*. It happens when communities include other communities or are included by them. A hyerarchical behaviour is pretty common in both human and aniaml societies and are crucial for an efficient management of large organizations. In this case as well , traditional community detection algortihms tipically don't search for an inner hyerarchical organization but just for the better partition.

## 2.2 Elements of community-detection

The problem of graph clustering, intuitive at first sight, is actually not well defined. The main elements of the problem themselves, i. e. the concepts of community and partition, are not rigorously defined, and require some degree of arbitrariness and/or common sense.

### 2.2.1 Communities

Providing a quantitative definition of what a community is quite difficult since this definition often depends on the particular system considered. In most cases, also, communities are defined by the algorithm used and not *a*

*priori* (Fortunato (2010)).

Be $\mathcal{C}$ a subgraph of graph $\mathcal{G}$ with $|\mathcal{C}| = n_{\mathcal{C}}$ and $|\mathcal{G}| = n$ vertices, respectively. Let's define the *internal degree* $k_v^{int}$ and the *external degree* $k_v^{ext}$ of vertex $v \in \mathcal{C}$ as the number of edges connecting $v$ to the other vertices of $\mathcal{C}$ or to the rest of the graph, respectively.

If $k_v^{ext} = 0$, it means that the vertex has neighbours only within $\mathcal{C}$ which results then as a good cluster for $v$; on the other hand, if $k_v^{int} = 0$, it means that the vertex is completely untied from $\mathcal{C}$ and it probably belogs to another cluster.

For $\mathcal{C}$ to be a community, it has to be *connected*, i.e there must be a path between each pair of nodes in the community that passes only by nodes of $\mathcal{C}$.

With this basic knowledge, one can introduce various community definitions but we will stick to a global definition that sees the graph as a whole and a community as an essential part of it that cannot be taken apart without seriously affecting the funtioning of the graph itself.

Such a definition sits on the assumption that on the idea that a graph has community structure if it is different from a random graph.

A random graph is not expected to have community structure, as any two vertices have the same probability to be adjacent, so there should be no preferential linking involving special groups of vertices.

Therefore, one can define a null model, i. e. a graph which matches the original in some of its structural features, but which is otherwise a random graph. The null model is used as a term of comparison, to verify whether the graph at study displays community structure or not. The most popular *null*

*model* is that proposed by Newman and Girvan and consists of a randomized version of the original graph, where edges are rewired at random, under the constraint that the expected degree of each vertex matches the degree of the vertex in the original graph (Newman and Girvan (2004a)).

This null model is the basic concept behind the definition of *modularity*, a function which evaluates the goodness of partitions of a graph into clusters. Modularity will be discussed thoroughly in this chapter and widely used in the following, because it has the unique privilege of being at the same time a global criterion to define a community, a quality function and the key ingredient of the most popular method of graph clustering (Fortunato (2010)).

### 2.2.2 Quality functions: modularity

Reliable algorithms are supposed to identify good partitions, where a partition is a division of a graph in clusters, such that each vertex belongs to one cluster.

Many algorithms are able to identify a subset of meaningful partitions, ideally one or just a few, whereas some others, deliver a large number of partitions. That does not mean that the partitions found are equally good.

Therefore we need to have a quantitative criterion to assess the goodness of a graph partition. A *quality function* is a function that assigns a number to each partition of a graph. In this way one can rank partitions based on their score given by the quality function. Partitions with high scores are "good", so the one with the largest score is by definition the best (Fortunato (2010)).

The most popular quality function is the modularity of Newman and

Girvan (Newman and Girvan (2004a)). As previously said, it is based on the idea that a random graph is not expected to have a cluster structure, so the possible existence of clusters is revealed by the comparison between the actual density of edges in a subgraph and the density one would expect to have in the subgraph if the vertices of the graph were attached regardless of community structure. This expected edge density depends on the chosen null model, i. e. a copy of the original graph keeping some of its structural properties but without community structure.

The modularity function can be written as Newman and Girvan (2004a):

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij}) \delta(C_i, C_j) \tag{2.1}$$

where the sum runs over all the node pairs, $A$ is the adjacency matrix, $m$ is the total number of edges and is the expected number of edges between the vertices $i$ e $j$ for a given null model. The $\delta$ function will result in a null contribution for couples of vertices not belonging to the same community ($C_i \neq C_j$).

Even if the choice of the null model is, in principle, arbitrary, the most suitable choice to make would be to keep the degree sequence of the original graph, due to the important implications that broad degree distributions have for the structure and function of real networks (Albert et al. (2000)).

This choice is stricter than just requiring the matching of the degree distributions and is essentially equivalent to the *configuration model* (Molloy and Reed (1995)).

In this null model, a vertex could be attached to any other vertex of the graph and the probability that vertices $i$ and $j$, with degrees $k_i$ and $k_j$ , are

connected,can be calculated without problems.

In fact, in order to form an edge between i and j one needs to join two *stubs* (i. e. half-edges), incident with $i$ and $j$. The probability $p_i$ to pick at random a stub incident with $i$ is $k_i/2m$, as there are $k_i$ stubs incident with $i$ out of a total of $2m$. The probability of a connection between $i$ and $j$ is then given by the product $p_i p_j$ , since edges are placed independently of each other.

The result is $k_i k_j/4m^2$ , which yields an expected number $P_{ij} = 2m p_i p_j = k_i k_j/2m$ of edges between $i$ and $j$. So, the final expression of modularity is:

$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j) \tag{2.2}$$

Since the only contributions to the sum come from vertex pairs belonging to the same cluster, we can group these contributions together and rewrite the sum over the vertex pairs as a sum over the clusters:

$$Q = \sum_{c=1}^{n_c} \left[ \frac{l_c}{m} - \left( \frac{d_c}{2m} \right)^2 \right] \tag{2.3}$$

where $n_c$ is the number of clusters, $l_c$ the total number of edges joining vertices of module $c$ and $d_c$ the sum of the degrees of the vertices of $c$.

In other words, in eq. 2.3 the first term of each summand is the fraction of edges of the graph inside the module, whereas the second term represents the expected fraction of edges that would be there if the graph were a random graph with the same expected degree for each vertex.

Eq. 2.3 also suggests that a subgraph is a community only if it positively contributes to modularity. So, the more the number of internal edges

of the cluster exceeds the expected number, the better defined the community. Higher values of modularity indicates good partitions. Modularity grows up as the number of cluster or the size of the graph increases so it shouldn't be used to compare the quality of the community structure of graphs which are very different in size.

The modularity of the whole graph, taken as a single community, is zero, as the two terms of the only summand in this case are equal and opposite. Modularity is always smaller than one, and can be negative as well. If there are no partitions with positive modularity, the graph has no community structure.

Modularity function naturally extend to the weighted case: the adjacency matrix $A_{ij}$ will be replaced by the weighted adjacency matrix $W_{ij}$ (see chapter 1) and instead of degrees $k_i$ and $k_j$ we will have strengths $s_i$ and $s_j$.

Many algorithms use modularity as quality function and modularity optimization method itself is a popular method of community detection.

## 2.3 Modularity-based methods

By assumption, high values of modularity indicate good partitions. So, the partition corresponding to its maximum value on a given graph should be the best, or at least a very good one. This is the main motivation for modularity maximization, by far the most popular class of methods to detect communities in graphs. An exhaustive optimization of $Q$ is impossible, due to the huge number of ways in which it is possible to partition a graph, even when the latter is small. Thus, it is probably impossible to find the solution

in a time growing polynomially with the size of the graph. However, there are currently several algorithms able to find fairly good approximations of the modularity maximum in a reasonable time (Fortunato (2010)).

### 2.3.1 Louvain method

A new approach has been introduced by Blondel et al. (2008), for the general case of weighted graphs. This method, known as "Louvain"[2], is today one of the most widely used algorithms for community detection since it is very simple to implement but very powerful in identifying modules, even in very large graphs (for sizes up to 100 million nodes and billions of edges), in a reasonable time.

The method consists of two steps, repeated iteratively until convergence:

**Step 1:** Let's consider a weighted network of size N. Initially, all vertices of the graph are put in different communities, so there will be as many communities as the nodes of the network. Then, a sequential sweep is performed over all vertices. Given a vertex $i$, one computes the gain in weighted modularity coming from putting $i$ in the community of its neighbor $j$ and picks the community of the neighbor that yields the largest increase of $Q$, as long as it is positive. If a positive gain is not possible, the node stays in its own community. Modularity gain $\Delta Q$,

[2]It is called like that because, even though the co-authors now hold positions in other places, the method was devised when they all were at the Universitè catholique de Louvain.

obtained moving node $i$ in a community $C$ can be calculated as follows:

$$\Delta Q = \left[\frac{\sum_{in}+k_{i,in}}{2m} - \left(\frac{\sum_{tot}+k_i}{2m}\right)^2\right] - \left[\frac{\sum_{in}}{2m} - \left(\frac{\sum_{tot}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right] \tag{2.4}$$

where $\sum_{in}$ is the sum of the weights of the edges internal to $C$, $\sum_{tot}$ is the sum of the weights of the edges incident in $C$ , $k_i$ is the sum of the weights of the edges incident to node $i$, $k_{i,in}$ is the sum of the weights of the edges from $i$ to nodes in $C$ and $m$ is the sum of the weights of all the edges in the network. A similar espression can be used to compute the change in $Q$ when removing a node from its community.

**Step 2:** Communities are replaced by superver- tices, and two supervertices are connected if there is at least an edge between vertices of the corresponding com- munities. In this case, the weight of the edge between the supervertices is the sum of the weights of the edges between the represented communities at the lower level. Edges between nodes of the same community lead to self-loops for this community in the new network.

Once second phase is completed, the two steps are repeated, yielding new hierarchical levels and supergraphs (figure 2.2).

## 2.4 Limits of modularity

Modularity suffers from some known drawbacks, which are crucial to identify the domain of its applicability and the reliability of its measures.

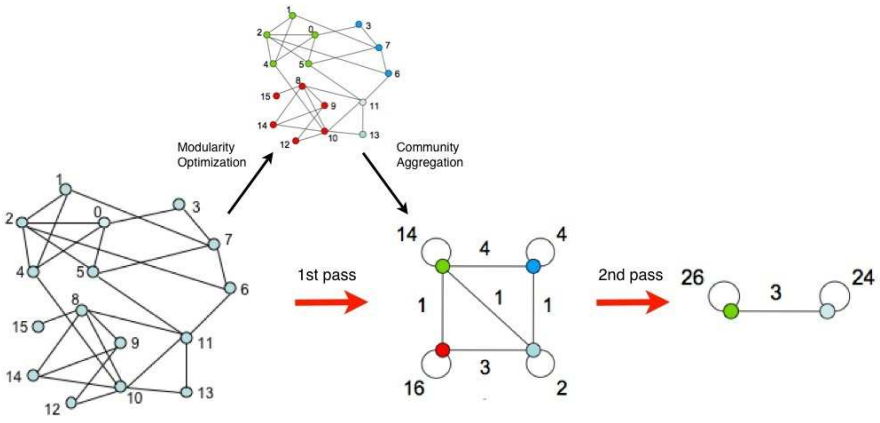



Figure 2.2: Visualization of the steps of Louvain algorithm. Each pass is made of two phases: one where modularity is optimized by allowing only local changes of communities; one where the found communities are aggregated in order to build a new network of communities. The passes are repeated iteratively until no increase of modularity is possible (Blondel et al. (2008)).

We said that a high value of modularity means that the given partition is a good one. However, a large value for the modularity maximum does not necessarily mean that a graph has community structure. Random graphs are supposed to have no community structure, as the linking probability between vertices is either constant or a function of the vertex degrees, so there is no bias a priori towards special groups of vertices. Still, random graphs may have partitions with large modularity values (Fortunato (2010)). This is due to fluctuations in the distribution of edges in the graph, in some cases they can concentrate in subsets of the network that can appear as communities.

A more fundamental issue, raised by Fortunato and Barthelemy (2007), is the so-called *resolution limit* and concerns the capability of modularity to detect communities which are comparatively small with respect to the graph

as a whole, even when they are well defined. So, if the partition with maximum modularity includes communities with total degree of the order of $\sqrt{m}$ or smaller, one cannot know *a priori* whether they are single communities or combinations of smaller weakly interconnected communities.

The resolution limit comes from the very definition of modularity, in particular from its null model. The weak point of the null model is the implicit assumption that each vertex can interact with every other vertex, which implies that each part of the graph knows about every- thing else (Fortunato (2010)).

The resolution limit problem seems to be circumvented in Blondel et al. (2008) thanks to the intrinsic multi-level nature of the algorithm. Since the first phase of the method involves the displacement of single nodes from one community to another, the probability that two distinct communities can be merged by moving nodes one by one is very low. These communities may possibly be merged in the following steps, after blocks of nodes have been aggregated. However, the algorithm provides a decomposition of the network into communities for different levels of organization so that one can observe its structure with the desired resolution.

# Chapter 3

# Data

Since 1973, when currencies began to be traded in financial markets and their values determined by the foreign exchange market, the volume of foreign exchange trading has been growing at an impressive rate. The transaction volume in 1995 was 80 times what it was in 1973. In the 1980s electronic trading, already a part of the environment of the major stock exchanges, was adapted to the foreign exchange market.

Physicists have generally investigated economic systems and problems only occasionally. Recently, however, a growing number of physicists is becoming involved in the analysis of economic systems.

Financial markets are, indeed, remarkably well-defined complex systems, which are continuously monitored - down to time scales of seconds. Further, virtually every economic transaction is recorded, and an increasing fraction of the total number of recorded economic data is becoming accessible to interested researchers, making financial markets extremely attractive for researchers interested in developing a deeper understanding of modeling of

complex systems (Mantegna et al. (2000)).

Also space is central to the work of economic institutions, providing the framework for survey design, sample selection, data collection, tabulation, and dissemination. Geography provides meaning and context to statistical data.

Given the diversity of population, economic activities, and geographic areas considered when dealing with economic or financial datasets, a spatial framework is then critical to provide real insight on data. Therefore, geographic area concepts, information, and statistical data must keep pace with the needs of the researchers and analysts who work to understand the changing distribution and characteristics of people, places and economy.

Given that, the following sections will describe in detail the data used in the papers reviewed in this thesis.

## 3.1 Sardinian Inter-municipal Commuting Network (SMCN)

Sardinia is the second largest Mediterranean island with an area of approximately 24.000 square kilometers and 1.600,000 inhabitants. In 1991, when the census was carried out, the island was partitioned in 375 municipalities, the second simplest body in the Italian public administration, each one of those generally corresponding to a major urban centre (in Figure 3.1 we report the geographical distribution of the municipalities).

For the whole set of municipalities the Italian National Institute of Statistics IST (1991) has issued the origin-destination table (OD) corresponding

Figure 3.1: Geographical representation of the the Sardinian inter-municipal commuting network (SMCN).

to the commuting traffic at the inter-city level. The OD is constructed on the output of a survey about commuting behaviors of Sardinian citizens. This survey refers to the daily movement from the habitual residence (the origin) to the most frequent place of employment (the destination): the data comprise both the transportation means used and the time usually spent for displacement. Hence, OD data give access to the flows of people regularly commuting among the Sardinian municipalities.

## 3.2 Atlanta Regional Commission (ARC)

The Atlanta Regional Commission maintains a network model for land use purposes of the metropolitan area of the city of Atlanta, in the State of Georgia, USA. The ARC travel demand model is designed to represent the

state of the practice in travel demand modeling and to meet all modeling requirements in the US EPA Transportation Conformity Rule.

The main data source for the calibration of the travel demand models was a household travel survey of eight thousand households conducted for the ARC from April 2001 through April 2002. The household survey data was the main source of data for developing the trip generation and distribution model. The trip generation model is a fairly unique trip based model in that it estimated the frequency a person will make trips, by the purpose of the trip, and then applies this frequency to individual persons to determine the total amount of travel made by the residents of the territorial unit (TAZ).

Therefore, just like in the case of the SMCN network, the trips reported in the ARC model are produced by a trip generation model, which is calibrated according to the result of a survey (further details are available in ARC (2008)). The calibration is achieved by matching the trip length, frequency and by evaluating geographic area biases (e.g., natural features, political or service delivery boundaries, etc).

## 3.3 Patent data

The OECD Patent Database was set up to develop patent indicators that are suitable for statistical analysis and that can help address S& T policy issues. The Patent Database covers data on patent applications to the European Patent Office (EPO), the US Patent and Trademark Office (USPTO), patent applications filed under the Patent Co-operation Treaty (PCT) that designate the EPO, as well as Triadic patent families. Data mainly derives from

the latest version of the EPO's Worldwide Patent Statistical Database (PATSTAT) (`http://www.oecd.org/science/inno/oecdpatentdatabases.htm`).

The OECD REGPAT database used here presents patent data that have been linked to regions according to the addresses of the applicants and inventors. The data have been regionalised at a very detailed level so that more than 2000 regions are covered across OECD countries. REGPAT allows patent data to be used in connection with other regional data such as GDP or labour force statistics, and other patent-based information such as citations, technical fields and patent holders characteristics (industry, university, etc.), thus providing researchers with the means to develop a rich set of new indicators and undertake a broad range of analyses to address issues relating to the regional dimension of innovation (Maraut et al. (2008),Webb et al. (2005)).

### 3.3.1 The OECD Regional Database

The OECD Regional Database (RDB) provides quantitative information on socio-economic issues in 2014 regions within 30 OECD member countries. The database includes regional statistics on four major topics (demographics, regional accounts, labour market, and social issues). The database contains annual data from 1990 to the present.

The RDB has been established to provide an internationally comparable database for the analysis of economic, institutional and environmental issues at the sub-national level. In any analytical study conducted at sub-national levels, the choice of the territorial unit is of prime importance.

To address this issue, regions within each member country have been clas-

sified in two territorial levels (TLs). The higher, more aggregate, Territorial Level 2 (TL2) consists of about 335 macro-regions while the lower, more detailed Territorial Level 3 (TL3) is composed of 1679 micro-regions. (Maraut et al. (2008))

While for European countries, this classification is largely consistent with the Eurostat classification (`http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction`) in NUTS levels 2 and 3, Canada and Australia are not part of the OECD official territorial grids so, for the sake of simplicity, they are labelled as Non Official Grids (3.2).

## 3.4 BACI data

BACI is the World trade database developed by the CEPII, providing bilateral values and quantities of exports of goods and services at the HS 6-digit product disaggregation, for more than 200 countries since 1995. It is updated every year (`http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=1`)..

Original data are provided by the United Nations Statistical Division (COMTRADE database). BACI is constructed using an original procedure that reconciles the declarations of the exporter and the importer. This harmonization procedure enables to extend considerably the number of countries for which trade data are available, as compared to the original dataset.

The dataset gives information about the value of trade (v, in thousands of US dollars) and the quantity (q, in tons). Individual trade flows are identified by the exporter (i), the importer (j), the product category (hs6) and the year

| Country | Large Regions (TL2) | Small Regions (TL3) | Non-Official Grids (NOGs) |
|---|---|---|---|
| Australia | 8 States/Territories | 58 Statistical Divisions | 30 LFS Dissemination Regions |
| Austria | 9 Bundesländer | 35 Gruppen von Politischen Bezirken | - |
| Belgium | 3 Régions | 11 Provinces | - |
| Canada | 12 Provinces and Territories | 288 Census Divisions | 71 LFS Economic Areas |
| Czech Republic | 8 Groups of Kraje | 14 Kraje | - |
| Denmark | 3 Regions | 15 Amter | - |
| Finland | 5 Suuralueet | 20 Maakunnat | - |
| France (without DOM-TOM) | 22 Régions | 96 Départements | - |
| Germany | 16 Länder | 97 Spatial planning regions (groups of Kreise) | - |
| Greece | 4 Groups of Development regions | 13 Development regions | - |
| Hungary | 7 Tervezesi-statisztikai regio | 20 Megyek (+Budapest) | - |
| Iceland | 2 regions | 8 Landsvaedi | - |
| Ireland | 2 Groups Regional Authority Regions | 8 Regional Authority Regions | - |
| Italy | 21 Regioni | 103 Province | - |
| Japan | 10 Groups of prefectures | 47 Prefectures | - |
| Korea | 7 Regions | 16 Special city, Metropolitan area and Province | - |
| Luxembourg | 1 State | 1 State | - |
| Mexico | 32 Estados | 209 Grupos de Municipios | - |
| Netherlands | 4 Landsdelen | 12 Provinces | - |
| New Zealand | 2 Groups of regional Councils | 14 Regional Councils | - |
| Norway | 7 Landsdeler | 19 Fylker | - |
| Poland | 16 Voïvodships | 45 Subregions | - |
| Portugal | 5 Comissaoes de coordenaçao regional + 2 Regioes autonomas | 30 Grupos de Concelhos | - |
| Slovak Republic | 4 Zoskupenia Karajov | 8 Kraj | - |
| Spain | 19 Comunidades autonomas | 52 Provincias | - |
| Sweden | 8 Riksomraden | 21 Län | - |
| Switzerland | 7 Grandes régions | 26 Cantons | - |
| Turkey | 26 Regions | 81 Provinces | - |
| United Kingdom | 12 Government Office Regions + Countries | 133 Upper tier authorities or groups of lower tier authorities or groups of unitary authorities or LECs or groups of districts | - |
| United States | 51 States | 179 BEA Economic Areas | - |

Figure 3.2: Territorial grids by country (Maraut et al. (2008)).

(t). BACI is available with versions 1992 (HS1), 1996 (HS2) and 2002 (HS3) of the Harmonized System (HS).

### 3.4.1 Harmonized Commodity Description and Coding Systems (HS)

The Harmonized System is an international nomenclature for the classification of products. It allows participating countries to classify traded goods on a common basis for customs purposes. At the international level, the Harmonized System (HS) for classifying goods is a six-digit code system.

The HS comprises approximately 5300 article/product descriptions that appear as headings and subheadings, arranged in 99 chapters, grouped in 21 sections. The six digits can be broken down into three parts. The first two digits (HS-2) identify the chapter the goods are classified in, e.g. 09 = Coffee, Tea, Maté and Spices. The next two digits (HS-4) identify groupings within that chapter, e.g. 09.02 = Tea, whether or not flavoured. The next two digits (HS-6) are even more specific, e.g. 09.02.10 Green tea (not fermented)... Up to the HS-6 digit level, all countries classify products in the same way (a few exceptions exist where some countries apply old versions of the HS).

The Harmonized System was introduced in 1988 and has been adopted by most of the countries worldwide. It has undergone several changes in the classification of products. These changes are called revisions and happened in 1996, 2002 and 2007.

## 3.5 World Input-Output Database

The World Input-Output Database has been developed to analyse the effects of globalization on trade patterns, environmental pressures and socio-economic development across a wide set of countries. At the time of writ-

ing, the WIOD input-output tables cover 35 industries for each of the 40 economies (27 EU countries and 13 major economies in other regions) plus the rest of the world (RoW) and the years from 1995 to 2011 (Timmer et al. (2012)). It is downloadable at `http://www.wiod.org/database/index.htm`.

Tables 3.1 and 3.2 have the lists of countries and industries covered in the WIOD. For each year, there is a harmonized global level input-output table recording the input-output relationships between any pair of industries in any pair of economies [1]. The numbers in the WIOD are in current basic (producers') prices and are expressed in millions of US dollars.

[1] The relationship can also be an industry to itself and within the same economy.

| Euro-Zone | | Non-Euro EU | | NAFTA | | East Asia | | BRIIAT | |
|---|---|---|---|---|---|---|---|---|---|
| **Economy** | **3L Code** | **Economy** | **3L Code** | **Economy** | **3L Code** | **Economy** | **3L Code** | **Economy** | **3L Code** |
| Austria | AUT | Bulgaria | BGR | Canada | CAN | China | CHN | Australia | AUS |
| Belgium | BEL | Czech Rep. | CZE | Mexico | MEX | Japan | JPN | Brazil | BRA |
| Cyprus | CYP | Denmark | DNK | USA | USA | South Korea | KOR | India | IND |
| Estonia | EST | Hungary | HUN | | | Taiwan | TWN | Indonesia | IDN |
| Finland | FIN | Latvia | LVA | | | | | Russia | RUS |
| France | FRA | Lithuania | LTU | | | | | Turkey | TUR |
| Germany | DEU | Poland | POL | | | | | | |
| Greece | GRC | Romania | ROM | | | | | | |
| Ireland | IRL | Sweden | SWE | | | | | | |
| Italy | ITA | UK | GBR | | | | | | |
| Luxembourg | LUX | | | | | | | | |
| Malta | MLT | | | | | | | | |
| Netherlands | NLD | | | | | | | | |
| Portugal | PRT | | | | | | | | |
| Slovakia | SVK | | | | | | | | |
| Slovenia | SVN | | | | | | | | |
| Spain | ESP | | | | | | | | |

Table 3.1: **List of WIOD Economies.**

### 3.5.1 Concept of a world input-output table (WIOT)

To outline the basic concept of a world input-output tables (WIOT), let's start discussing a national input-output table (IOT). Figure 3.3 shows the schematic outline for a traditional industry by industry IOT.

As in Timmer et al. (2012), we assume that each industry produces only one (unique) product. The rows in the upper parts indicate the use of products, being for intermediate or final use. Each product can be an intermediate

| Full Name | ISIC Rev. 3 Code | WIOD Code | 3-Letter Code |
|---|---|---|---|
| Agriculture, Hunting, Forestry and Fishing | AtB | c1 | Agr |
| Mining and Quarrying | C | c2 | Min |
| Food, Beverages and Tobacco | 15t16 | c3 | Fod |
| Textiles and Textile Products | 17t18 | c4 | Tex |
| Leather, Leather and Footwear | 19 | c5 | Lth |
| Wood and Products of Wood and Cork | 20 | c6 | Wod |
| Pulp, Paper, Paper , Printing and Publishing | 21t22 | c7 | Pup |
| Coke, Refined Petroleum and Nuclear Fuel | 23 | c8 | Cok |
| Chemicals and Chemical Products | 24 | c9 | Chm |
| Rubber and Plastics | 25 | c10 | Rub |
| Other Non-Metallic Mineral | 26 | c11 | Omn |
| Basic Metals and Fabricated Metal | 27t28 | c12 | Met |
| Machinery, Nec | 29 | c13 | Mch |
| Electrical and Optical Equipment | 30t33 | c14 | Elc |
| Transport Equipment | 34t35 | c15 | Tpt |
| Manufacturing, Nec; Recycling | 36t37 | c16 | Mnf |
| Electricity, Gas and Water Supply | E | c17 | Ele |
| Construction | F | c18 | Cst |
| Sale, Maintenance and Repair of Motor Vehicles and Motorcycles; Retail Sale of Fuel | 50 | c19 | Sal |
| Wholesale Trade and Commission Trade, Except of Motor Vehicles and Motorcycles | 51 | c20 | Whl |
| Retail Trade, Except of Motor Vehicles and Motorcycles; Repair of Household Goods | 52 | c21 | Rtl |
| Hotels and Restaurants | H | c22 | Htl |
| Inland Transport | 60 | c23 | Ldt |
| Water Transport | 61 | c24 | Wtt |
| Air Transport | 62 | c25 | Ait |
| Other Supporting and Auxiliary Transport Activities; Activities of Travel Agencies | 63 | c26 | Otr |
| Post and Telecommunications | 64 | c27 | Pst |
| Financial Intermediation | J | c28 | Fin |
| Real Estate Activities | 70 | c29 | Est |
| Renting of M&Eq and Other Business Activities | 71t74 | c30 | Obs |
| Public Admin and Defence; Compulsory Social Security | L | c31 | Pub |
| Education | M | c32 | Edu |
| Health and Social Work | N | c33 | Hth |
| Other Community, Social and Personal Services | O | c34 | Ocm |
| Private Households with Employed Persons | P | c35 | Pvt |

Table 3.2: **List of WIOD Industries.**

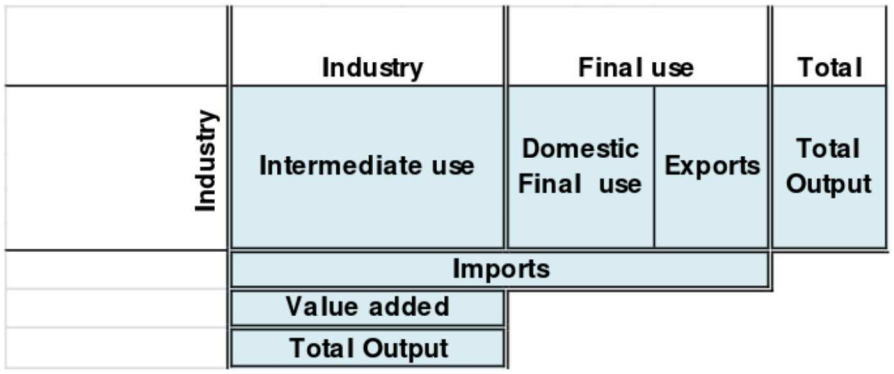



Figure 3.3: Schematic outline of a national input-output table (Timmer et al. (2012)).

in the production of other products (intermediate use). Final use includes domestic use (private or government consumption and investment) and exports. The final element in each row indicates the total use of each product. The industry columns in the IOT contain information on the supply of each product. A product can be imported or domestically produced. The column indicates the values of all intermediate, labour and capital inputs used in production. The vector of input shares in output is often referred to as the technology for domestic production. The compensation for labour and capital services together make up value added which indicates the value added by the use of domestic labour and capital services to the value of the intermediate inputs. Total supply of the product in the economy is determined by domestic output plus imports. An important accounting identity in the IOT is that total output by the domestic industry is equal to the use of output from the domestic industry such that all flows in the economic system are accounted for.

A world input-output table (WIOT) is an extension of the same concept.

The difference with the national tables is that the use of products is broken down according to their origin. Each product is produced either by a domestic industry or by a foreign industry. In contrast to the national IOT, this information is made explicit in the WIOT. For a country A, flows of products both for intermediate and final use are split into domestically produced or imported. In addition, the WIOT shows in which foreign industry the product was produced. This is illustrated by the schematic outline for a WIOT in Figure 3.4

| | | Country A Intermediate *Industry* | Country B Intermediate *Industry* | Rest of World Intermediate *Industry* | Country A Final domestic | Country B Final domestic | Rest of World Final domestic | Total |
|---|---|---|---|---|---|---|---|---|
| Country A | *Industry* | Intermediate use of domestic output | Intermediate use by B of exports from A | Intermediate use by RoW of exports from A | Final use of domestic output | Final use by B of exports from A | Final use by RoW of exports from A | Output in A |
| Country B | *Industry* | Intermediate use by A of exports from B | Intermediate use of domestic output | Intermediate use by RoW of exports from B | Final use by A of exports from B | Final use of domestic output | Final use by RoW of exports from B | Output in B |
| Rest of World (RoW) | *Industry* | Intermediate use by A of exports from RoW | Intermediate use by B of exports from RoW | Intermediate use of domestic output | Final use by A of exports from RoW | Final use by B of exports from RoW | Final use of domestic output | Output in RoW |
| | | Value added | Value added | Value added | | | | |
| | | Output in A | Output in B | Output in RoW | | | | |

Figure 3.4: Schematic outline of a world input-output table (3 regions) (Timmer et al. (2012)).

This combination of national and international flows provides a powerful tool for analysis of global production chains and their effects on employment, value added and investment patterns and on shifts in environmental pressures (Timmer et al., 2012).

# Chapter 4

# Applications

The research presented here has been carried out focusing on community detection; in particular were considered cases where the spatial component was relevant or intrinsic. It is indeed true that, nowadays, many systems, represented as complex networks, are affected, more or less naturally, by the geographical distance, location and organization.

This holds true even for economic events: it has been proved that trade and exchanges between countries are necessarily suffocated by the geographical proximity or impeded by natural obstacles. We all saw how the 2007-2008 financial crisis spread following preferential routes that could be ascribed to the relational patterns between states. Relations that are influenced, among the other things, by geographical distance.

Thus, if we want to efficiently study phenomena happening in the physical world, space cannot be overlooked.

According to this view we developed a way to enforce standard community detection methodology with a set of space-oriented tools, such as spatial

modularity, outreach index and geo-localization methods, and representation.

Nonetheless, community detection alone is not sufficient to describe the whole picture, since it gives no information about the internal structure of a community. Therefore we developed the novel *core detection* method, natural counterpart of the community detection algorithm and meant to be performed alongside it, which is, at the same time, simple and powerful.

Thanks to both community and core detection we are now able to have a deeper insight on the inner workings of community formation, we can identify the leading members in a group and reveal influence basins, unknown otherwise .

In sections 4.1 and 4.2 we will explain the spatial modularity and core detection methods and in sections 4.3, 4.4, 4.5 thoroughly show results and applications (each section is titled as the relative published - or still submitted - work).

## 4.1 Spatial Correlations in Attribute Communities (Cerina et al. (2012))

In spatial networks, each node is described by its coordinates (usually in a 2d space) but has in general other attributes. For individuals, it can be any cultural or socio-economical parameter. For infrastructure networks such as power grids, it can be the voltage at the electric substations. In general, this attribute depends on space and the resulting network displays entangled layers of parameters. An important goal in the analysis of these networks is to disentangle these different levels and to extract some mesoscopic information

from the spatial network structure. If one is interested in studying effects beyond space (Expert et al., 2011), one should have a straightforward way to 'subtract' it from the network, or in other words, to disentangle space and the other attributes.

A natural tool for such a task is community detection which was used for the characterization at a mesoscopic scale of the properties of complex networks. Community detection can have several purposes in spatial networks Guimerà et al. (2005), Kaluza et al. (2010), Hu et al. (2011b), Gregory (2011), but probably the main one is to disentangle these various aspects, including spatial correlations of any type. In most cases Guimerà et al. (2005), Kaluza et al. (2010) communities are determined by the geography only, which results from the simple fact that the most important flows are among nodes in the same geographical regions. In this sense, community detection in spatial networks offers a visual representation of large exchange zones. This even suggests that community detection might be an important tool in geography and in the determination of new administrative or economical boundaries De Montis et al. (2011).

In the general case, for a given network we don't know to what extent the existence of a link between a pair of nodes is due to a specific factor or to space only. The link could exist because of a strong attribute affinity between the nodes, or in the other extreme case, because they are close neighbors. In general, one could expect a combination of these two effects. If we are interested in recovering communities defined by an attribute (such as language for example) from the network structure, we then have to consider various assumptions such as the correlation between link formation, attribute values

and space. In order to understand the effect of the underlying correlations, we can consider two extreme cases. When the links are purely spatial and independent from the attributes, if we remove the spatial component, we will observe random communities (obtained for a random graph) which contain a random number of nodes with random attributes. In this situation, community detection is unapplicable and there is no way to recover attribute communities from the network structure. The other extreme case is when the formation of a link depends on the attributes only. In this case, space is irrelevant and any standard community detection method should give sensible results, ie. communities made of nodes with the same attribute.

The important problem we wanted to focus on here is thus the intermediate case when the probability to have a link depends both on attributes and on space. In this case we have to eliminate spatial effects in order to recover the attribute structure. An important point in the discussion is then the existence of correlation between space and attributes. The nature and existence of these correlations will govern the way we will have to do community detection.

In this work, we constructed a simple artificial network model allowing us to investigate the effect of these correlations on the results of the community detection procedure and test various methods on it.

### 4.1.1 A benchmark for spatial networks with attributes

In order to test these ideas and how community detection acts on spatial networks, we define a simple model of spatial networks where nodes and their attributes are randomly distributed in space. The attributes could be

anything and we will restrict - without loss of generality - to the simple binary case where the attributes can have two possible values at each node. In general, according to the various parameters of the model, the attributes can be delocalized in space or, on the contrary, be localized in some well-defined region. In some cases, some attribute community could emerge in space, but our target community structure will always be the partition of the network in the two subgraphs composed of nodes with the same attribute and we will test how various methods can recover these two communities. In this respect the main focus of our work will be the disentanglement of the sole attribute network features beyond the spatial node arrangements.

We construct the test (benchmark) network defining the vertex and edge properties in the following way.

**Vertex properties**:

1. We generate points/nodes in the $2d$ space $(x - z)$ in two spatial communities, say the North and the South, around the two centers $(x, z) = (0, +L)$ and $(x, z) = (0, -L)$ (see Fig. 4.1). A simple way to do that is to generate points $i$ around the two centers according to the probability

$$p(x_i, z_i) \propto e^{-d_{ci}/\ell} \tag{4.1}$$

where $d_{ci}$ is the euclidean distance between one of the centers $c$ and the node $i$ of coordinates $(x_i, z_i)$.

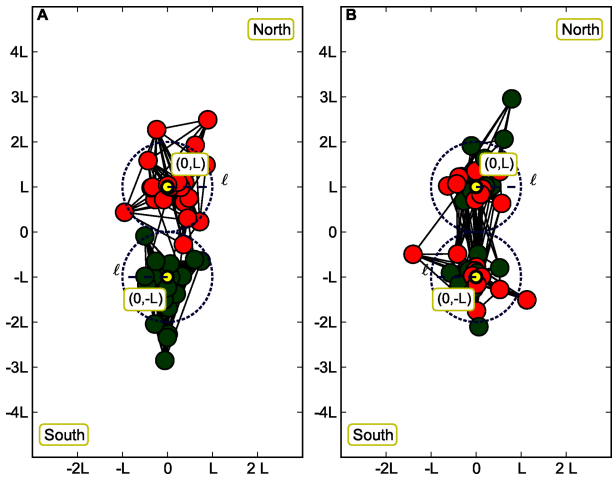



Figure 4.1: The two spatial communities North and South are well separated having their average size $\ell = L$. In the A panel we present the case $\epsilon = 0$ where there is a perfect correlation between the space and the attributes (green and red colors). In the B panel, the uncorrelated case $\epsilon = 0.5$ is presented where the attribute colors are randomly distributed between the two segregated spatial communities (for the sake of clarity, only 40 out of the 100 nodes used in our simulations are shown here, and $\beta = 1.0$).

2. We assign an attribute $S_i$ to each node $i$. In the following we will focus on the simplest case where this attribute can take only two values $S_i = \pm 1$ (which in this work are the red and green colors). A simple way to control correlations between attribute and space is to choose $S_i = +1$ with probability $q$ for $z > 0$ and $S_i = -1$ with probability $1-q$. In order to tune the various cases we introduce the parameter $\epsilon$, with $q = 1 - \epsilon$, that determines the mixing between space and attributes, ranging from 0.0 to 0.5. In the case $\epsilon = 0.0$ space and attributes are strongly correlated, while for $\epsilon = 0.5$ space and attribute are totally uncorrelated.

   So the relevant parameters for the generation of network nodes are $\ell$

and $\epsilon$.

**Edge properties**:

3. We then construct the network: for each pair of nodes, we create a link between nodes $i$ and $j$ with probability $p_{link}(i,j) \propto e^{\beta S_i S_j - d_{ij}/\ell_0}$ where $\ell_0$ plays the role of the typical size of the spatial community (and where $d_{ij}$ is the euclidian distance between $i$ and $j$). It is worth observing that the parameter $l_0$ is the typical length of links when space dominates while $\ell$ is the typical spatial size of the northern and southern communities. Here the relevant edge parameters are $\beta$ and $\ell_0$, but in order to simplify the model and to focus on the efficiency of community detection methods, we choose $\ell = \ell_0$. This choice implies that when space dominates the link formation, the links cannot be much larger than the community size. In this case, the only spatial relevant parameter will be $\ell/L$ and we can fix $L$ to be equal to 1.0 so that the spatial variability will be governed by $\ell$. We can rewrite the probability $p_{link}(i,j)$ as

$$p_{link}(i,j) = \frac{1}{\mathcal{N}} e^{\beta(S_i S_j - d_{ij}/\ell\beta)} \tag{4.2}$$

where $\mathcal{N} = \sum_{i<j} \exp\left(\beta S_i S_j - d_{ij}/\ell\right)$ is the normalization constant. As in the Erdos-Renyi random graph, the number of edges is a random variable with small fluctuations around its average. The number of nodes is thus fixed in each network but not the number of edges or the average degree, and this implies that we will have to average our observables over different realizations of the network.

When $\beta\ell$ is large, links are essentially between nodes with the same attribute (irrespective of their distance) and if $\beta\ell$ is small then space is the governing factor and links are essentially between neighboring nodes.

In this way the probability associated to a link depends on both space and attribute, and the correlation between attributed and space can be controlled. If the attribute is the same between two nodes the probability to have a link will be reinforced, otherwise it will be weakened, the interplay being controlled by the parameter $\beta$ . Concerning the spatial factor, the closer the nodes and the larger the probability associated to this link.

The generation of attributes is an important point. We have two values of the attribute only so that we need to generate attributes for only half $(N/2)$ of the nodes. So in the following we will study the specific case of an attribute community structure of equal size communities: half of the nodes has attribute $S_i = +1$ and the other half has $S_i = -1$. We will investigate here two extreme situations:

- Attributes and space uncorrelated: this case is recovered by choosing $\epsilon = 1/2$.
- Attributes and space are strongly correlated. For this, we choose $\epsilon$ small. In this case, the spatial communities are also attribute communities.

Furthermore we can distinguish two different spatial arrangements for the northern and southern communities. The first case corresponds to a situation where the two communities are well separated with their average size $\ell \leq L$ and the spatial effects dominate the community structure (see Fig. 4.1). The

second situation corresponds to a larger value of the average community size $\ell$ where the two communities start mixing up while $\ell$ approaches $L$ (see Fig. 4.2).

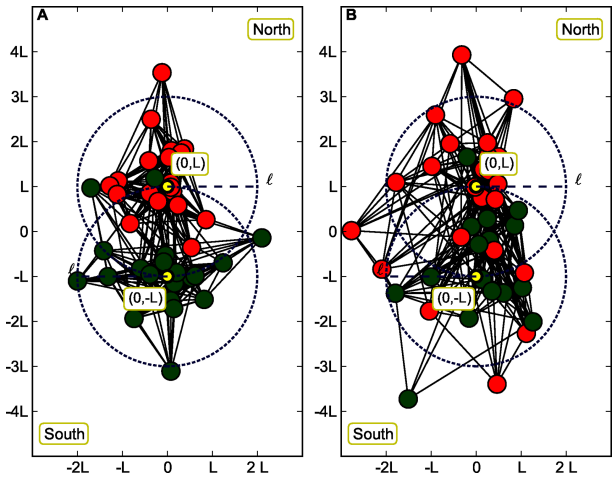



Figure 4.2: The two communities North and South are mixing up each other with their average size $\ell$ approaching the value of $L$ (in this case $\ell = 2L$). In the A panel, we display the case $\epsilon = 0.0$. Even if the spatial correlation is fading away the space-attribute correlation is still strong enough to display an attribute community. In the B panel, we show the extreme case $\epsilon = 0.5$ where the attributes are not correlated with space. In this case spatial mixing destroys the attribute community structure (for the sake of clarity, only 40 out of the 100 nodes used in our simulations are shown here, and $\beta = 1.0$).

There are many proposal in the literature for networks benchmarking (see for example Lancichinetti and Fortunato (2009)), but this is -up to our knowledge- the first one which takes into account the correlation between space and node attributes.

### 4.1.2 Methods

The interplay between space and attributes can lead to various situations that need to be understood within the framework of community detection. Indeed we have two main regimes $\beta\ell \gg 1$ and $\beta\ell \ll 1$ (see also Figure 4.3):

| **Spatial correlation** $\epsilon$ | $\beta\ell \ll 1$: **Space is the governing factor** | $\beta\ell \gg 1$: **The spatial component of the links is irrelevant** |
|---|---|---|
| Spatially correlated: ($\epsilon = 0.0$) | • Links are between neighboring nodes but spatial communities correspond to the attribute ones.<br>• Any regular community detection will work. | • Links are between nodes with the same attribute.<br>• Any community detection method should work. |
| Spatially uncorrelated: ($\epsilon = 0.5$) | • Links are between neighboring nodes but the attributes are anywhere in space.<br>• It is necessary to 'remove' space in order to uncover the attribute communities. | • Links are between nodes with the same attribute.<br>• Any community detection method should work. |

Figure 4.3: The table gives an account of the behaviour of the model in the regimes $\beta\ell \ll 1$ and $\beta\ell \ll 1$ both in the correlated ($\epsilon = 0.0$) and uncorrelated ($\epsilon = 0.5$) case.

$\beta\ell \gg 1$ . In this case, the spatial component of the links becomes irrelevant (see Eq. 4.2) and for a given value of $\beta$ the community structure due to the node attributes will emerge, independently from the correlation between space and attributes. In this regime any community detection method should work.

$\beta\ell \ll 1$ . Here we have two subcases depending con the correlation between space and attributes:

- ($\epsilon = 0.0$) Space and attributes are correlated: any regular community detection will work and moreover if you carefully remove the spatial effect the attribute community structure will be recovered.
- ($\epsilon = 0.5$) Space and attributes are uncorrelated: in this case the links are between neighboring nodes but the attributes are anywhere in space. Standard community detection methods won't

work and it is then necessary to 'remove' space in order to uncover the attribute communities.

The general assumption of our model is to what extent it is possible to detect communities even if there is a spatial influence. Without space the initial situation is clear: we have two communities by construction and the probability of two nodes to be connected is related to the attribute similarities. Nodes with S=+1 tend mainly to connect to each other and the same for the S=-1 nodes. If we then put nodes in space and enhance the connection probability due to the proximity of nodes, it is not clear if a regular community detection method is able to detect the original two communities structure. We thus see that correlations between space and attributes can be misleading and any community detection method for spatial networks should take into account this problem. There are now many community detection methods Fortunato (2010) and in the following we will use modularity optimization introduced by Newman and Girvan Newman and Girvan (2004a). This method suffers from various problems, the most important being the existence of a resolution limit Fortunato and Barthelemy (2007) which prevent it to detect smaller modules, but it is simple enough to implement. In addition, our point here is to understand the effect of space-attributes correlations on community detection and not to compare various methods. In the following we will thus essentially probe the Newman-Girvan method and variants proposed here and in Expert et al. (2011) for cases where the space and attribute have different degrees of correlation.

The modularity function which needs to be optimized is Newman and Girvan's modularity, as defined in section 2.2.2. In order to introduce ex-

plicitly space, the idea is to change the null model defined by $P_{ij}$ and to compare the actual network with this null model. Recently, such a proposal was made in Expert et al. (2011) where the quantity $P_{ij}$ is directly obtained from the data describing the network. More precisely, Expert et al. Expert et al. (2011) used the following form

$$P_{ij}^{Data} = N_i N_j f(d_{ij}) \tag{4.3}$$

where $N_i$ is related to the importance of the node $i$ (such as the population for example). This form is reminiscent of the gravitional model for traffic flows (see for example Erlander and Stewart (1990)) where flows are proportional to the product of populations and decrease with distance. In Expert et al. (2011), the authors proposed to estimate the unknown function $f$ directly from the empirical data by

$$f(d) = \frac{\sum_{i,j|d_{ij}=d} A_{ij}}{\sum_{i,j|d_{ij}=d} N_i N_j} \tag{4.4}$$

which can be seen as the probability to have two nodes connected at a distance $d$. Note that there is a binning procedure hidden in Eq. (4.4). The usual way to proceed in these cases consists in introducing a discretization of the space in bins that capture classes of distances. Following Expert et al. (2011), we performed a binning of distances selecting the best value for the number of bins after a detailed stability study of the distributions obtained from the data.

Expert et al. Expert et al. (2011) applied this method to the specific case of the phone network in Belgium, and try to reconstruct linguistic communities (Flemish and French) beyond individuals spatial location. This choice is probably the best one if there are no correlations between the attribute

under study (in their case the linguistic membership of the people calling each other) and space. In this specific case, extracting the node spatial dependencies from the actual link distribution present in the network data is the most effective way to subtract the spatial component. Otherwise if there are any correlations between space and node attributes, the data contain in an unknown proportion the two informations (space and attribute) and their method needs to be reformulated. One possible way to do this is to explicitly guess a spatial dependency of the link distribution and to put it as an independent factor in the optimization function definition. In order to be able to deal with the correlated case and to remove spatial effect only, we thus propose the following explicit function of space for $P_{ij}$

$$P_{ij}^{Spatial} = \frac{1}{Z} k_i k_j g(d_{ij}) \tag{4.5}$$

where $Z$ is the normalization constant, $k_i$ the degree of the node $i$, $d_{ij}$ the euclidean distance between node $i$ and node $j$. The function $g(d)$ is a decreasing function of distance and its role is to remove the spatial effect. A simple choice is

$$g(d) = e^{-d/\overline{\ell}} \tag{4.6}$$

where $\overline{\ell}$ is the average distance between nodes in the network. Of course $\overline{\ell}$ is a rough approximation of the real $\ell$ value, but we will see in the following that it is enough to capture the essence of the spatial signature of the network.

We now need a method to compare the community structure obtained with the modularity optimization and the expected one for the attribute membership. Many proposals have been introduced Danon et al. (2005), Campello (2007), Karrer et al. (2007), and we decided to use here the *Jaccard Index* Jain and Dubes (1988), Halkidi et al. (2001). This index is an

extension of the Rand index Rand (1971), and is considered to be one of the most robust measure for the clustering and classification assessment of graphs Denoeud et al. (2006). If $C$ is the partition to be evaluated and $C'$ the reference one the definition is as follows

$$J_I = \frac{a}{a+b+c} \tag{4.7}$$

where $a$ is the number of vertices pairs that are in the same community for both $C$ and $C'$, $b$ is the number of pairs that are in different communities in $C$ but in the same one in $C'$ and finally $c$ is the number of vertices pairs that are in the same community in $C$ but not in $C'$ (or conversely). This quantity $J_I$ is in the interval $[0, 1]$ and the closer to one, the better the agreement between the two partitions. For $J_I = 1$ there is a perfect match between the two community structures. In our case, it would mean that the attribute communities are exactly detected. For values of $J_I$ less than 1 the discrepancy can depend both on the size of the partitions in the community structure and/or the number of them and in this respect the *Jaccard Index* is a good method to compare a very heterogeneous range of community structures.

In order to get a more intuitive picture of the Jaccard index, we show three different cases in Fig. 4.4 for the same value $\beta\ell = 0.2$ (and in the case $\epsilon = 0.0$, $\ell = 1.0$ and $L = 1.0$) but with different values of $J_I$. The first case corresponds to a relatively small value $J_I = 0.232$ (obtained with the 'Data' method of Expert et al. (2011), where the binning is done as in their work, which shows a partition in four communities (instead of the two associated with the attributes in red and green colors). For intermediate values such as $J_I = 0.579$ (obtained with our 'Spatial' method) the communities reduce to three with a prevalence of circles in the nothern part and triangles in the

southern (see B panel in Fig. 4.4). The last case (obtained with the original Newman-Girvan formulation) corresponds to a value $J_I = 0.903$, that almost recovers the attribute community structure.

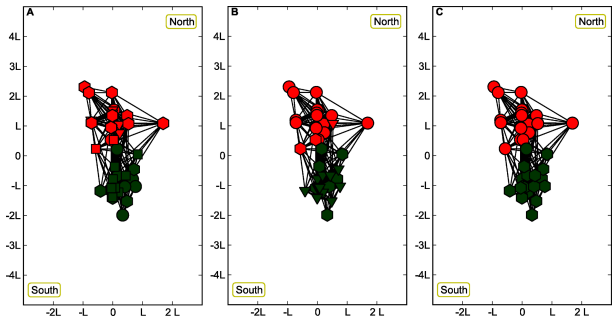



Figure 4.4: Three spatial network configurations are presented for the constant value $\beta\ell = 0.2$ and the correlated case $\epsilon = 0.0$ with $\ell = 1.0$ and $L = 1.0$. The color (red and green) are the attributes, while the geometrical shapes represent the community memberships found with the various community detection procedure discussed in this section. In the A panel, we present the case $J_I = 0.232$, obtained with the Data method. Due to the low $J_I$ value four communities are present (instead of the two associated with the attributes in red and green colors) and they are also mixed up between the south and the north spatial regions. In the B panel we show the $J_I = 0.579$ case obtained with the Spatial method. Three communities are present and in the northern part there is a prevalence of circles while in the southern of triangles. The C panel displays the case $J_I = 0.903$ obtained with the Newman-Girvan formulation and the attribute community structure is almost completely recovered.

Finally, in order to have a baseline value we also computed the average Jaccard for a completely random partition for $N = 100$ nodes and we obtain the value $J_I = 0.08 \pm 0.05$.

### 4.1.3 Results

The goal of this spatial community detection is to substract the spatial component and to recover the (two) attribute communities. We thus have three community detection methods: the original Newman-Girvan method, the 'Data' method proposed in Expert et al. (2011), and our 'Spatial' method defined by the null model of Eq. (4.5) and, in order to understand their limits , we will test them against the benchmark network introduced above.

We will now see how these three different methods perform in the two extreme cases of attribute correlated ($\epsilon = 0$) and uncorrelated ($\epsilon = 0.5$) with space, both varying the size of the spatial communities $\ell$ and the attribute linkage strength $\beta$. The size of the test network is $N = 100$ nodes and the number of links depends on the probability previously defined (Eq. 4.2). We generated 100 network realizations for each set of parameters ($\beta$, $\ell$, $\epsilon$ and $L = 1$). For each point of the simulation curve the error bars are the standard deviation for 100 modularity measures. To optimize the modularity we used the Louvain method Blondel et al. (2008).

The behavior of the model depends on both parameters $\beta$ and $\ell$ and we will first show the case with fixed attribute strength $\beta$. We show on the A panel of figure 4.5 the correlated case ($\epsilon = 0$) with a fixed $\beta = 1.0$.

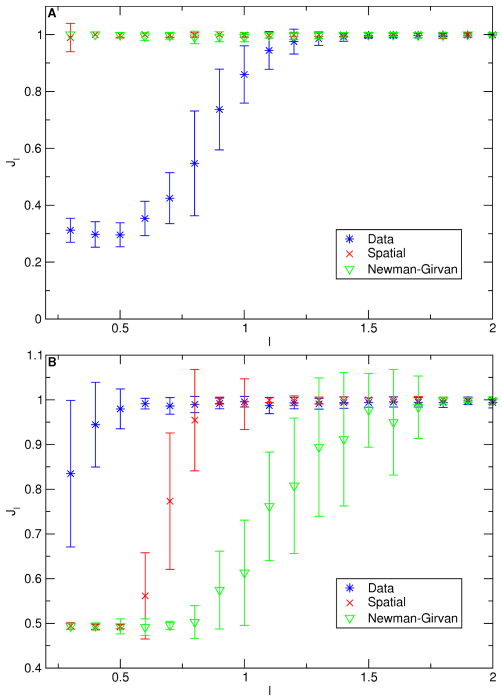



Figure 4.5: The community structure obtained for various values of $\ell$ with fixed $\beta = 1.0$. Each point represents the average Jaccard index for 100 network community detection and the error bar is its standard deviation. The correlated case $\epsilon = 0$ is shown on the A panel, and on the B panel we show the uncorrelated case $\epsilon = 0.5$. In A for the regime $\beta\ell \ll 1$ both the Newman-Girvan and the 'Spatial' method formulations give the right attribute community structure corresponding to the Jaccard index $J_I = 1.0$. For the regime $\beta\ell \gg 1$ all the three formulations work well since the links due to the attribute similarity are strong enough to preserve the community structure irrespectively from the node's location. In the uncorrelated case (B panel), the Data based formulation performs better respect to the Spatial formulation, since it extracts correctly the spatial information, directly from the data. In any case both spatial methods reach the right attribute community structure at almost the same value for $\ell \simeq 1.0$. The Newman-Girvan standard formulation instead fails to detect the correct result up to values of $\ell \simeq 1.8$. Note that in the x-axis we considered only values equal or above 0.3 since we verified that below this value the model generates disconnected networks.

In this case, for $\beta\ell \gg 1$, all the three methods work well, as expected and we obtain a perfect match ($J_I = 1$) between the community structure resulting from the modularity optimization and the attribute communities. Space is not relevant in this regime and links exist essentially among nodes with the same attribute. For $\beta\ell \ll 1$ both the Newman-Girvan modularity and the 'Spatial' method give the correct result. The latter actually subtract only the spatial dependency while the the 'Data' method mixes the space effect with the correlated attribute feature, resulting in a wrong community detection. The 'Data' method, for a sufficiently large value of $\ell$ will approach anyway the correct $J_I = 1.0$ value.

In the uncorrelated case (Fig. 4.5, B panel) and for a low values of $\beta\ell$, the Newman-Girvan modularity is not able to detect the right attribute communities, since the attribute correlation is not strong enough to group together the nodes of similar type. Instead the other two methods perform better in getting the attribute communities since they are able to correctly eliminate the effect of space and recover the attribute community structure, even for a small attribute correlation. The formulation based on Data performs even better since it eliminates the effect of space almost pointwise, but in any case the correct result of $J_I = 1$ is reached almost at the same value $\ell \simeq 1.0$ for both spatial methods.

In Figure 4.6 we show the results for the case of a fixed community size ($\ell = 1.0$) but where we vary the attribute strength $\beta$. In the A panel the correlated case is presented ($\epsilon = 0$). As expected the 'Data' method for low values of $\beta$ has problems in detecting the attribute community structure and only for high attribute strengths ($\beta$) it starts to correctly detect the target

communities. In the uncorrelated case, where the space is irrelevant, the standard Newman-Girvan formulation fails, while the two spatial methods performs similarly better (Fig. 4.6).

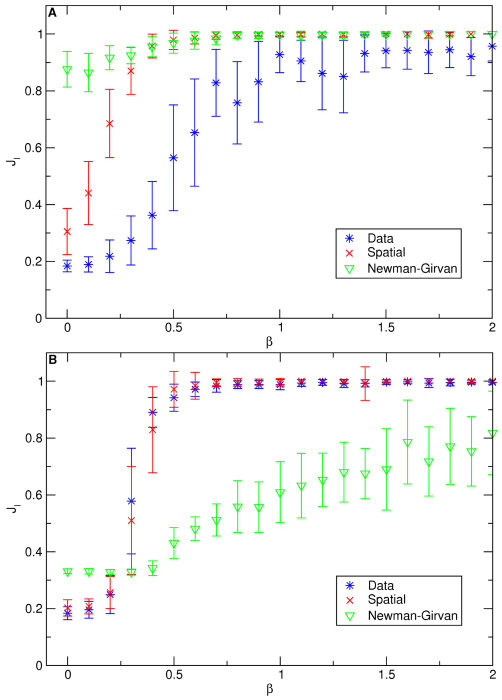



Figure 4.6: The community structure obtained for various values of $\beta$ with fixed community size $\ell = 1.0$. Each point represents the average Jaccard index for 100 network community detection and the error bar is its standard deviation. The correlated case $\epsilon = 0$ is shown on the A panel, and on the B panel we show the uncorrelated case $\epsilon = 0.5$. In the uncorrelated case the 'Data' method fails in detecting the attribute community structure for all the $\beta\ell$ regimes present in the figure, while the other two methods start working at $\beta = 0.8$. In the uncorrelated case the Newman-Girvan method is not able to detect the attribute community structure, while the spatial methods perform similarly better approaching the correct $J_I = 1.0$ value around $\beta = 0.8$.

In order to summarize these results we show in Table 4.1 the only relevant regime (b) previously defined, $\beta\ell \ll 1$ (the (a) regime $\beta\ell \gg 1$ is trivial as we

can verify in Figs 4.5 and 4.6) for all the parameters of interest ($\epsilon$, $\ell$ and $\beta$) and for the three community detection methods. From this Table, it clearly emerges that the Spatial method is a very good interplay in all situations, while to get the best performances one has to choose the suitable method for any specific case.

| Spatial correlation $\epsilon$ | | Newman-Girvan | Data | Spatial |
|---|---|---|---|---|
| 0.0 (correlated) | $\ell$ | VG | B | VG |
| | $\beta$ | VG | B | G |
| 0.5 (uncorrelated) | $\ell$ | B | VG | G |
| | $\beta$ | B | G | G |

Table 4.1: The table summarizes the performances, as can be extracted from Figs 4.5 and 4.6, of the three methods (Newman-Girvan, Data and Spatial) in the only non trivial regime $\beta\ell \ll 1$, both in the correlated ($\epsilon = 0.0$) and uncorrelated ($\epsilon = 0.5$) case. Since in the plots we vary both $\ell$ and $\beta$, we distinguish here these two cases. In order to be able to compare this results we classified them according to the following criteria: **B**, **G** and **VG** that stand for **Bad**, **Good** and **Very Good**. We assign VG when there is a very good agreement with the target attribute community structure ($J_I$ very close to 1), G when the behavior is rapidly approaching the correct result even for low/medium values of the parameters $\ell$ and $\beta$, and finally B when it completely fails to recover the right community structure.

We note that the behavior of the error bar sizes in these figures 4.5, is

interesting. For $\beta\ell \ll 1$ and $\beta\ell \gg 1$, the error in the modularity estimate is relatively small. The error bar -or equivalently the fluctuations of the Jaccard index- are the largest for $\beta\ell \simeq 1$. In this region, the community detection methods are thus more sensitive to small fluctuations of the network which implies a peak in the 'susceptibility' of the system. This behavior is reminiscent of the phase transition between detectability and non-detectability presented in Hu et al. (2011a), Decelle et al. (2011). Indeed, in figure 4.7 we show the limiting case of $l \gg L$ (here we choose numerically $l = 4$ and $L = 1$) for which the effect of space is irrelevant. In this limit, our model becomes equivalent to the stochastic block model of Decelle et al. (2011) with $q = 2$ possible values of the attribute. In our case the control parameter ($c_{out}/c_{in}$ in Decelle et al. (2011)) is $\exp(-2\beta)$, while the order parameter is the Jaccard index. It is clear from Fig. 4.7 that the same effect is present (see figure 2 in Decelle et al. (2011)) even if the critical point is shifted due to a different community detection method and another definition of the order parameter. Moreover, respect to the result in Decelle et al. (2011), in the undetactable regime ($\beta = 0$), the value of the order parameter is not zero. As mentioned above, for a completely random partition the $J_I$ is $J_I = 0.08 \pm 0.05$. We observe that in our case we are a little bit above because it is known that even for a random network the modularity can be positive Guimerà et al. (2004) and in this way the maximization of the modularity extracts a subset of the ensemble of all the possible partitions that increases the average modularity and consequently the average Jaccard index.

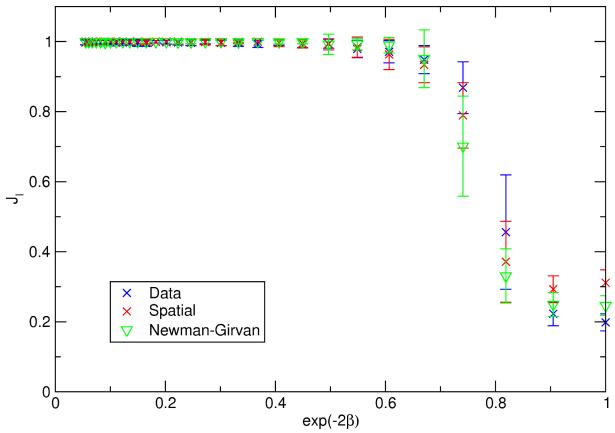



Figure 4.7: Transition obtained in the case $\ell \gg L$ from the detectable to the undetectable community structure regions. This transition was described in Decelle et al. (2011) for the stochastic block model which corresponds to our model with $q = 2$ attributes when the effect of space is absent, i.e. $\ell$ large ($\ell = 4.0$ in the actual simulation). The control parameter is then $\exp(-2\beta)$ and the Jaccard index is our order parameter. All the three community detection methods discussed in this work display the same behavior adding evidence to the universality of the transition presented in Decelle et al. (2011).

We thus recover the results of Decelle et al. (2011) and in addition our result seems to point to the existence of a spatial phase transition actually independent of the community detection method used.

Finally, we checked the performances of the Data and Spatial formulations looking at the $J_I$ values when varying the $\epsilon$ parameter for a fixed $\beta\ell$ value (see Fig. 4.8). For each value of $\epsilon$ an higher $J_I$ value signals a better behavior since it is closer to the maximum value $J_I = 1$. We choose first the value $\beta\ell = 0.8$ (we also tested $\beta\ell = 1.0$ which gives similar results). There is a crossover in the performances around $\epsilon \simeq 0.25$. Below this value, the Spatial method performs better while above that point the Data method does slightly better. This result thus shows that there can be a non-negligible range of correlations

(measured here by $\epsilon$) for which the spatial community detection results can be incorrect.

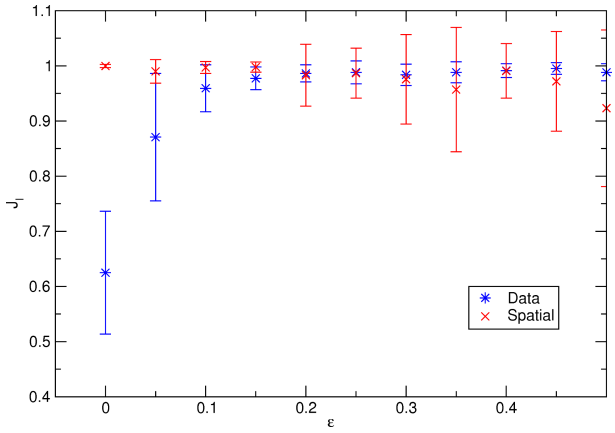



Figure 4.8: Performances of the Spatial and Data modularity formulations. We show here the case $\beta\ell = 0.8$ where there is a crossover in the performances around $\epsilon \simeq 0.25$. Below this value $\epsilon = 0.25$ the Spatial method performs better and above the Data method is slightly better.

We have shown here that community detection in spatial networks should be taken with great care because it could lead to incorrect results, depending on the correlations between space and attributes.

## 4.2 Community core detection in transportation networks (De Leo et al. (2013))

Still, spatial community detection methodology, just like the traditional one, does not provide any information about the importance of nodes in their own community. As was pointed out by the authors of Fortunato (2010), communities are algorithmically defined, i.e., they are the final product of

the algorithm, without a precise *a priori* definition.

In this work we aim to understand the nature of these communities and find ways to determine the importance of the vertices inside each community, revealing its inner hierarchy by means of a novel method for core detection (De Leo et al. (2013)).

We used two different structures, Sardinian Inter-Municipal and Atlanta metropolitan area commuters data, as testing ground for this new method. The field of transportation is a natural choice for the definition of a community structure, though the field itself has some inherent limitations. On a practical matter, the measurement of important traffic variables is lengthy and expensive. For one, different methods to count traffic volumes return different answers, especially in the identification of commercial vehicles fhw. Additionally, the development of a region-wide origin-destination (OD) matrix at the zone level is a long and costly procedure; in particular, the matrix of the metropolitan area used in this study has been derived after a year-long survey process, and the final OD matrix was assembled by weighting a matrix of survey responses according to the population of the areas where the participants lived. A second calibration stage is generally done to test whether the OD matrix obtained assigns traffic compatibly with the traffic on the major highways of the study area; as a result of this process, the trip distribution and assignment may work well *globally*, but larger discrepancies may persist *locally*. Finally, during the time occurred to carry out this process, conditions on the ground may have already changed, since the land-use of an area is constantly changing, therefore creating discrepancies in the final OD matrix.

Notwithstanding these inherent difficulties, the identification of communities within a metropolitan area network still holds great importance. First, the formation of communities in a network is a byproduct of land-use development. Land-use development occurs for a number of reasons (service maximization, profit, etc), and the location for development is chosen according to the optimization in terms of different variables, like the price of land, proximity to transit, and regulation, which are, however, variables related to each zone or vertex of the system. For example, the demand for transport between two vertices may lead to the opening of a new edge (e.g., a new bus route, a new road), which in turn may lead to more demand for transport (in the form of "induced demand", Mishan (1972), Pashigian (1995)). The community structure is not solely a function of the attributes of each zone or vertex, but also of the network arrangement; hence it forms a more comprehensive measure of the importance of a group of zones as a subsection of the zone system.

It is important to know which vertices are the most relevant from the point of view of the internal stability of a community and the overall partition structure. We will see in this section that this idea is at the cornerstone of the community stability. In other fields the problem has been studied in terms of network breakdown, which has found applications in the accessibility of a transportation network for flood damage. Knowledge of community structure can serve planners in the situation of natural disasters to predict the onset of network breakdown, as was studied by the authors of Sohn (2006). In other fields, it has been applied to the identification of crucial edges in a web network under cybernetic attack Albert et al. (2000), Solé et al. (2008),

Schneider et al. (2011).

This work analyses methods for the identification and the stability of a community structure using two networks from the field of transportation. The first network is a regionwide network of commuting trips in the insular region of Sardinia, in Italy, the Sardinian Inter-municipal Commuting Network (SMCN). The second network is a network of daily commuting trips in the metropolitan area of Atlanta, GA, USA, part of the Atlanta Regional Commission (ARC) model. In both cases, we have studied the distribution of commuting trips, i.e., home-to-work trips and viceversa (see section 3.1 and section 3.2). The choice was determined by the fact that trips of these types are clearly defined to planners, because their correlation to the land-use is well understood, necessarily tied to the population of the origin zone and the employment of the destination zone.

### 4.2.1 $dQ$ analysis for cores detection in a partition

The starting point of this method is the modularity optimization introduced by Newman and Girvan Newman and Girvan (2004b). By definition, if the modularity associated to a network has been optimized, every perturbation in the partition leads to a negative variation in the modularity ($dQ$).

As shown in Fig. 4.9, if we move a node from its community we have $M - 1$ possible choices (with $M$ the number of communities) as possible targets for the new host community of this node. We decided to define the $dQ$ associated to each node as the smallest variation in absolute value (or the closest to 0 since $dQ$ is always a negative number) for all the possible choices and this is in our view a measure of how that node is internal in its

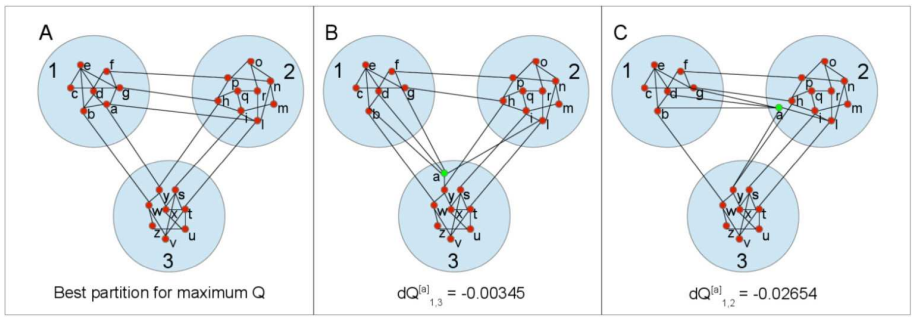



Figure 4.9: This picture (A) describes the situation in which the modularity of a network has been maximized. Starting from this state it is possible to determine the list of negative $dQ$ values associated to each node assuming to move it in every other community. As a matter of fact, if a node (in this picture we consider as example the node "a") would change its belonging to the community in which has been placed during the modularity optimization, the modularity of the network would obviously decrease, as shown in (B) and (C). This negative variation is related to the fact that, for each change in the partition, like the ones depicted in (B) and (C), the total number of links internal to the communities is always smaller with respect to the one associated to (A).

community.

Fig. 4.10 shows the typical $dQ$ frequency distribution of nodes inside a community; the data points were fitted using a decaying exponential form $\exp(-x/\ell)$ with typical length $\ell$. The typical lenght $\ell$ and defines a starting point to discriminate the core nodes. For practical purposes, the threshold value $d_{\text{thr}} = 2\ell$ is an appropriate boundary value to differentiate between core nodes (the ones below the threshold) and the border nodes (the peripheral nodes). With this choice we found that, for what it concerns the networks described in this work, the percentage of core nodes is, for every community of every network, always equal to the 8% of the total amount of nodes in that particular community.

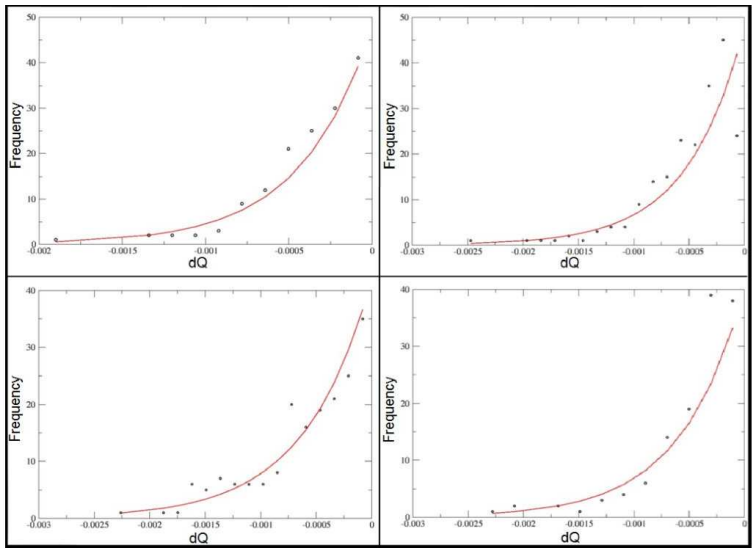



Figure 4.10: *dQ* frequency plots relative to four communities detected for the city of Atlanta, GA. The correlation coefficients of the exponential fits are (from top right to bottom left, respectively) 0.956, 0.946, 0.937 and 0.933. In general, these distributions are the typical *dQ* frequency distribution inside a community (provided there are enough nodes to perform an exponential fit).

Figure 4.11 shows the cores detected for the city of Atlanta, GA, using the method described above. The nodes of the network correspond to the Traffic Analysis Zone (TAZ) of the city and the links' weight have been computed summing, for each couple of TAZ, the corresponding traffic flow in both directions, as described in more detail later.

### 4.2.2 Sardinian Inter-municipal Commuting Network

To test this new methodology we focused on the flows of individuals (workers and students) commuting throughout the set of Sardinian municipalities by all means of transportation, described in section 3.1. This data source

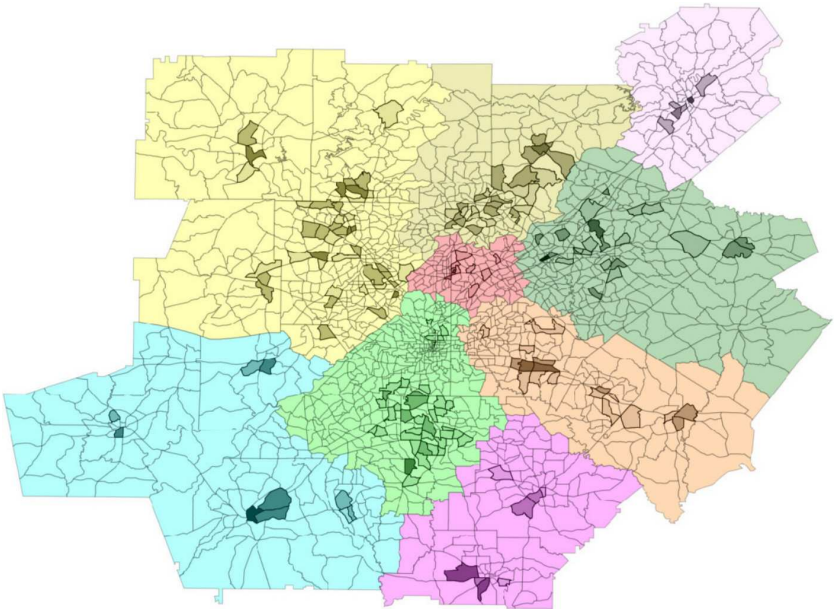


Figure 4.11: Cores detected for the city of Atlanta, GA, using a threshold equal to double the typical length of the exponential distribution of the $dQ$ frequencies.

allows the construction of the SMCN in which each node corresponds to a given municipality and the links represent the presence of a non-zero flow of commuters among the corresponding municipalities.

We are able then to construct a symmetric weighted adjacency matrix $W$ in which the elements $w_{ij}$ are computed as the sum of the $i \to j$ and $j \to i$ flows between the corresponding municipalities (per day). The elements $w_{ij}$ are null in the case of municipalities $i$ and $j$ which do not exchange commuting traffic and by definition the diagonal elements are set to zero . According to the assumption of regular bi-directional movements along the links, the weight matrix is symmetric and the network is described as an undirected weighted graph. The weighted graph provides a richer description since it considers the topology along with the quantitative information on

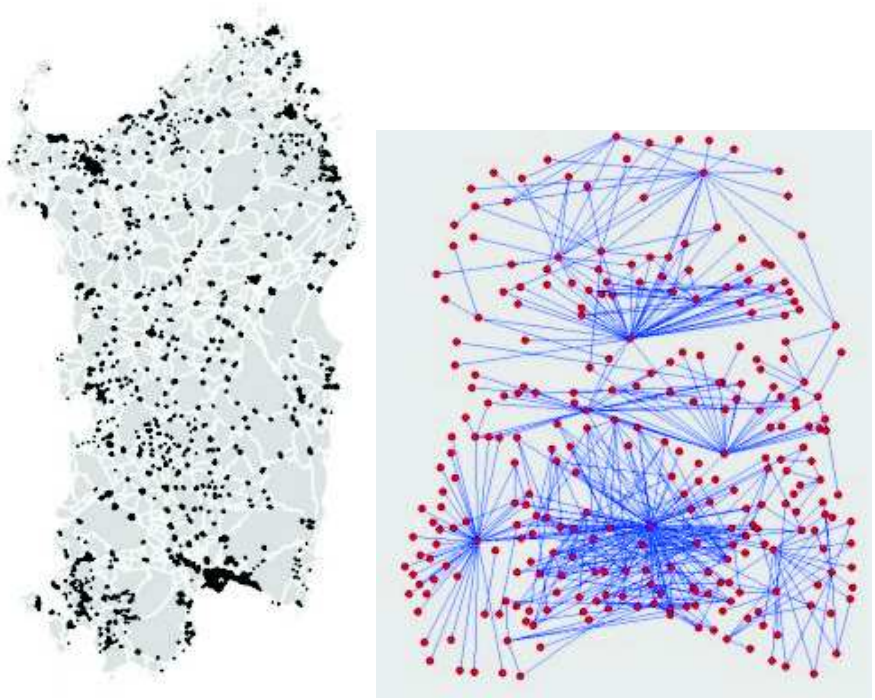


Figure 4.12: Geographical versus topologic representation of the the Sardinian inter-municipal commuting network (SMCN): the nodes (red points) correspond to the towns, while the links to a flow value larger than 50 commuters between two towns.

the dynamics occurring in the whole network.

### 4.2.3 ARC Network

The present work is centered on the activity of commuters, shown as blue lines in Figure 4.13, which in the ARC model are described as "Home Based Work" (HBW) trips (see section 3.2. It is commonplace to describe such trips as trips made for the purpose of work and which either begin or end at the traveler's home. This is a typical trip purpose that is related to the employment at the destination zone and population and household income of the traveler or

the household at the origin zone. The nature of the relationship between the demand for travel and land use are further explored in the modeling review works by Wilson (1998) and Batty (1976).

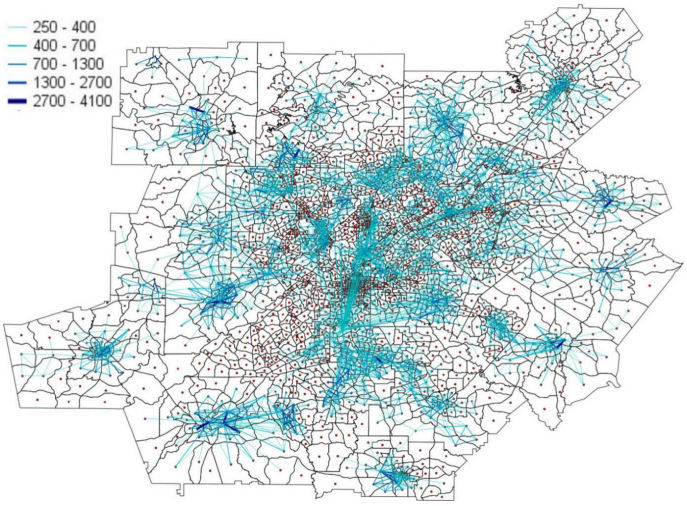



Figure 4.13: Extension of the zone system in the ARC model. Only the links with a weight greater than 250 have been shown. Each point is a centroid of a TAZ.

A number of socioeconomic variables are recorded in the ARC model, which are of importance for planning purposes and as inputs to the trip generation and demand growth algorithms. The figures below show, in order, the gradient plots of population and employment per zone, as recorded in the nationwide Census 2010. Darker zones indicate a higher value for the corresponding variable.

Figure 4.14 shows the gradient plot of the zone population. Population is seen in this figure as being scattered around the center that forms the core of the downtown area.

Figure 4.15 shows the gradient plot for the zone employment, measured as the number of jobs located in the zone the variable refers to. Employment

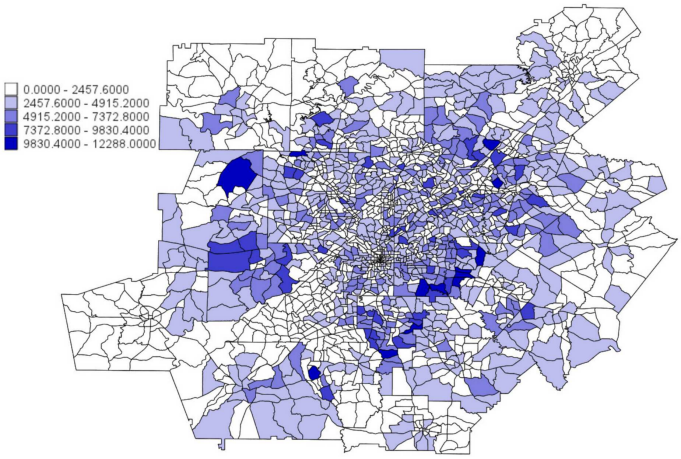



Figure 4.14: Gradient plot for population in the ARC model.

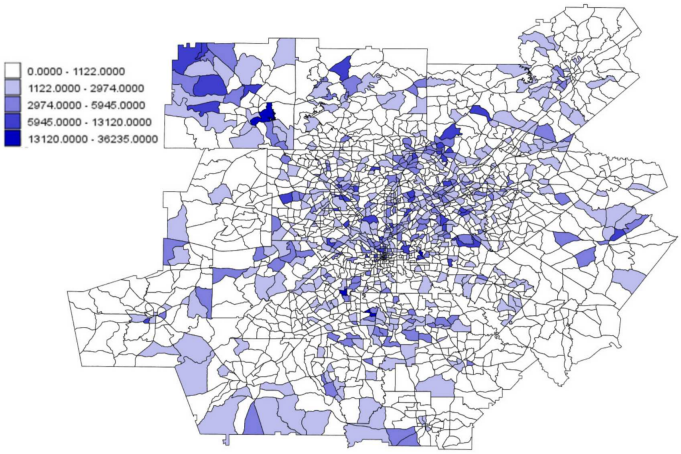



Figure 4.15: Gradient plot for employment in the ARC model.

is seen in this figure as primarily located in the downtown zones (which are quite small in size) plus other job centers in the suburban metropolitan areas.

### 4.2.4 Results

The sequence of charts that follow describes the correlation of the quantity $dQ$ and the various socioeconomic variables that are available for analysis.

| Network | In-strength | Employment |
|---|---|---|
| SMCN | 0.984 | 0.984 |
| ARC | 0.782 | 0.520 |

Table 4.2: Results of correlation analysis between $dQ$ and the in-strength and employment.

The table below shows the result of the correlation analysis between the computed $dQ$ and the in-strength of the various zones in the SMCN network. For the sake of clarity, the Sardinian and ARC networks are directed, as previously described, and the in-strength has been computed starting from these original networks. However, the community detection has been performed using undirected networks obtained from the directed ones by summing up the weights of incoming and outgoing links. The correlation results shown in Table 4.2 only give an overall picture of the quality of correlation between the traffic and community structure. Figures 4.16 and 4.17 show the geographic distribution of the gradients of $dQ$ values across the zone system. Figure 4.16 shows the values of $dQ$ arranged by color (darker color indicates higher value). Higher $dQ$ indicates that the zone under investigation is more to the center of a community than the zones with lighter color. The data in Figure 4.16 shows that the two likeliest centers of a community (the two darkest zones in the figure) are not both centers of population and/or employment, nor are all large centers of population and/or employment necessarily key zones to the definition (and for its definition, stability) of a community. In other words, community and socioeconomic activity are not on a one-to-one

relationship, and it is not always possible to imply a ranking of one of these quantities with respect to the other and viceversa.

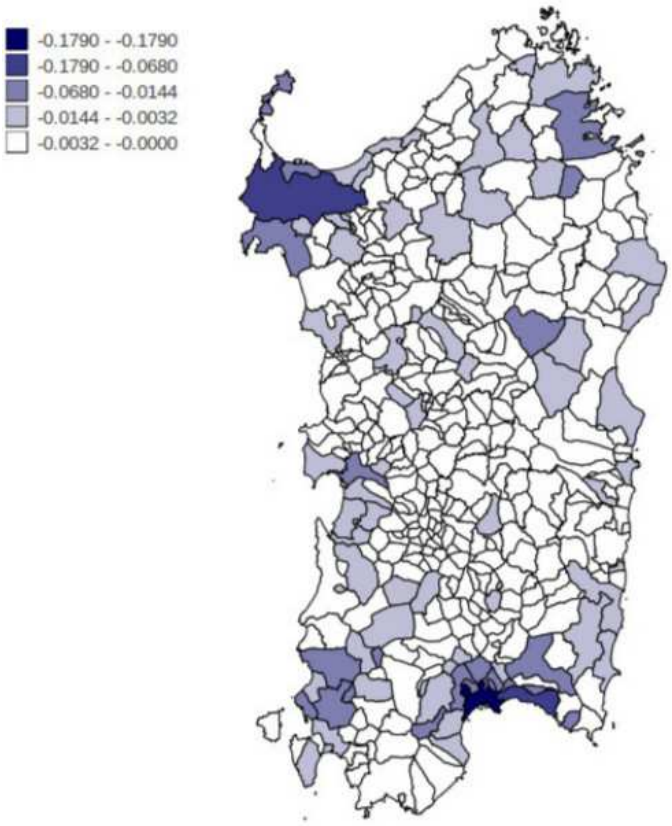



Figure 4.16: $dQ$ plot for the network related to employment in the SMCN network.

Figure 4.17 (right) below shows what the communities identified look like with respect to the political subdivisions of the island of Sardinia, the provinces that corresponds to the NUT3 regions in the international classifications (left). To put this result in context, it is important to note that the present political subdivision in eight provinces took effect in 2005 after a law passed in 2001 raised the number of provinces from the original number of four. Therefore, at the time the ISTAT data was collected (2001), Sar-

dinia was subdivided politically in four provinces, hence the results of the modularity analysis showed that at least seven communities existed, subdivided geographically roughly along the lines of the boundary of the new (and present time) provinces. The two subdivisions, "topological" the first, political the second, are remarkably alike, suggesting that either the political subdivision was designed to accommodate the arrangement of commuting movements, or the topological subdivision is a result of ease of movement within a (not yet established) political subdivision.

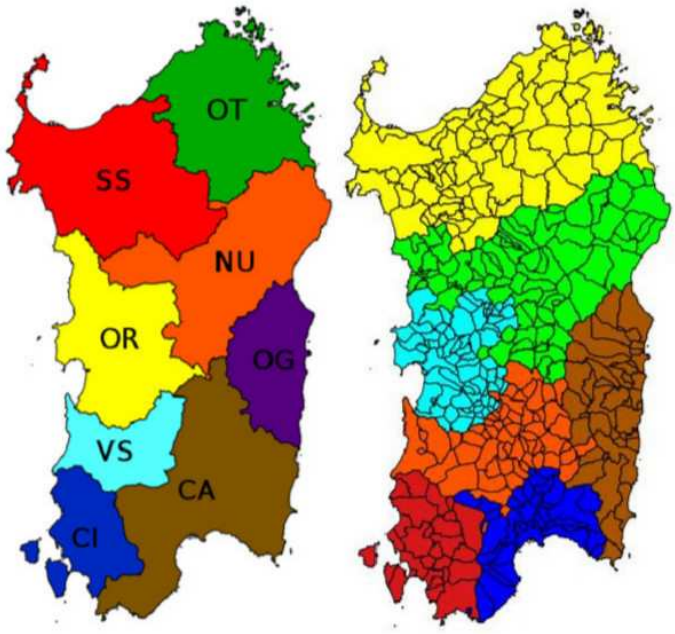



Figure 4.17: A comparison between the current provincial division (CA = Cagliari, CI = Carbonia-Iglesias, VS = Medio Campidano, OR = Oristano, OG = Ogliastra, NU = Nuoro, SS = Sassari and OT = Olbia-Tempio) of the Sardinia region, Italy, and the result of the community detection.

Finally, it is worth noting that, according to the results of a regional referendum in May 2012, the four new provinces established in according to

the 2001 law were abolished in March 2013.

Table 4.2 shows also the result of the correlation between in-strength, $dQ$, and employment for the ARC network. The correlation with employment is poorer, while as in the case of the SMCN network, the correlation with the in-strength is quite good. It is instructive then to see the geographic arrangement of the communities and other features of the network. Figure 4.18 shows the $dQ$ distribution for the ARC network. Darker zones indicate zones with higher $dQ$, and the darkest zones can be considered as the center of a community. Figure 4.19 show (color-coded) the community boundaries. The correlation between $dQ$ and in-strength is explored by means of Fig.

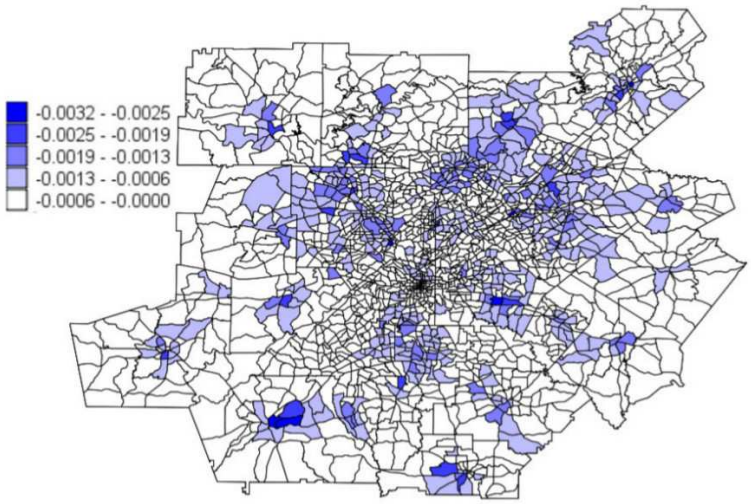



Figure 4.18: $dQ$ plot for the ARC network.

4.20, which shows a correlation of almost 0.8.

The novel core detection method has here been applied to a territorial network but its definition is quite general and can be naturally extended to other networked systems.

The following sections will show three different applications. in section 4.3

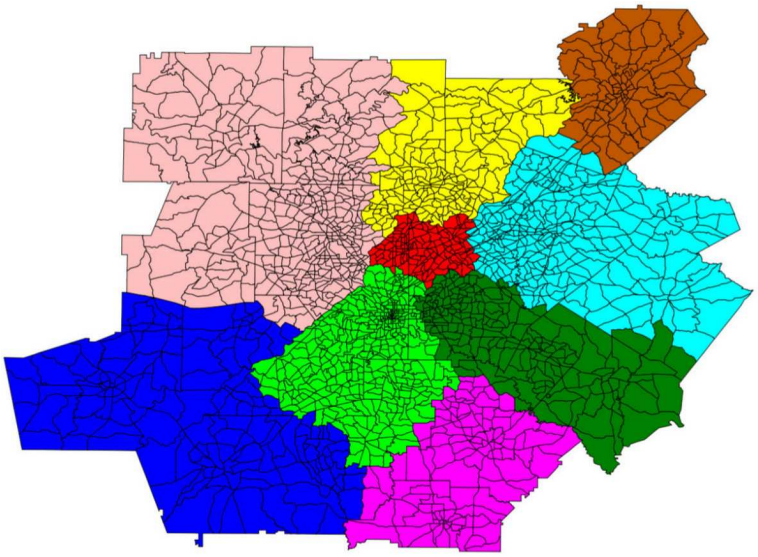


Figure 4.19: $dQ$ and community boundary plot for the ARC network.

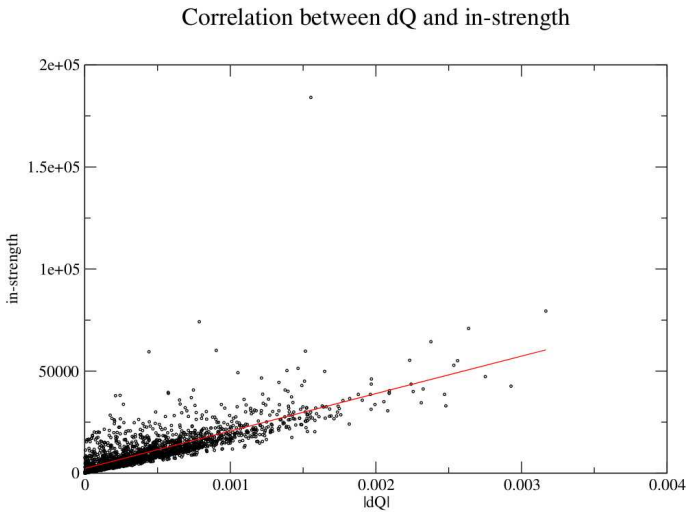



Figure 4.20: The correlation between $dQ$ and in-strength is equal to 0.78.

core detection is used as a criterion to select the center of a community in order to calculate the novel outreach index; in section 4.4 it helps reveal a leadership change within the same community, impossible to detect with

community detection alone; in section 4.5 it is used as an alternative method to assess the most important industries in a country.

## 4.3 Network communities within and across borders (Cerina et al. (2014))

Over recent decades, political changes and new transportation and information technologies have enhanced international openness and cross-border integration. Globalization has made social networks more international and human communities more integrated across cultural and political borders. This is witnessed by the increasing number of long-range connections in multiple networks, such as trade, human mobility, communications, financial investments and scientific collaborations Arunachalam and Doss (2000), Scherngell and Lata (2012), Hoekman et al. (2013), Thiemann et al. (2010).

Enabled by modern technology, people from all over the world are offered a myriad of opportunities for social interactions and group assembly with increasingly larger geographic ranges Onnela et al. (2011). Nonetheless, this does not mean that networks can stretch across a borderless world indefinitely: as for climate networks one can detect geographical regions with the same climate variability Tsonis and Roebber (2004), Daqing et al. (2011), Berezin et al. (2012). As individual nodes in socio-economic networks occupy a given region in space, it is reasonable to assume that geographical proximity also plays a crucial role in social link formation Watts and Strogatz (1998). Indeed, a power-law decay in link probability with distance acting as a spatial constraint has been observed Onnela et al. (2011), Daqing et al. (2011), Lam-

biotte et al. (2008), Goldenberg and Levy (2009). In a recent meta-analysis estimating the role of distance in international trade, it has been shown that $t \propto d^{-\gamma}$, where $t$ is trade, $d$ is the distance and $\gamma \approx 1$ over more than a century of data Disdier and Head (2008). Arguably, distance is not the only spatial constraint on link formation, since natural and artificial borders also have the power to hamper connectivity. Communication and transportation routes, on the contrary, facilitate long-distance interactions Brockmann and Helbing (2013). Geographical and institutional borders are relevant in all networks where distance matters, such as power grid networks, transportation and communication networks as well as collaboration networks. Natural, artificial and administrative borders can substantially reduce the probability of link formation by introducing a major physical constraint in terms of cost, service, capacity and reliability of global networks. Two of the most widely accepted results in international economics are that trade is impeded by distance and that the crossing of national borders also sharply reduces trade. It has been shown, for instance, that national borders are responsible for a five-fold decrease in world trade when compared to a borderless world Eaton and Kortum (2004). Well-known global networks are transportation and communication networks, such as the airline network and the World Wide Web, for which the role of borders has been recently documented Halavais (2000), Guimerà et al. (2005).

Traditionally, international openness has been proxied by the share of cross-border links over the total number of connections. More recently, various network-based measures of cross-border integration have been introduced Kali and Reyes (2007), Arribas et al. (2009), Duernecker et al. (2012). Sim-

ilarly, a multinational corporation consists in a group of geographically dispersed organizations that include its headquarters and the various national subsidiaries. Such an entity can be conceptualized as an international spatially embedded network to develop network measures of firm internationalization Ghoshal and Bartlett (1990), Rauch (2001). Network-based measures take "who connects with whom" into consideration, rather than just looking at the degree of openness. On a different level, the effective borders between spatially embedded networks only partially overlap with existing administrative borders Thiemann et al. (2010). To properly measure the extent of the international span of networks and cross-national communities it is of paramount importance to assess the effectiveness of policies devoted to international collaboration, such as the ones implemented by the European Union to favor the free movement of people, goods, investments and ideas across European borders. As part of this effort, the European Research Area has been recently deemed equivalent in terms of research and innovation with respect to the European common market for goods and services. In this work (Cerina et al. (2014)) we explore the effect of borders on the European and US co-inventorship networks as a way to assess the progress toward the effective cross country integration of scientific and technological communities Chessa et al. (2013).

### 4.3.1 Data

The data analyzed in this study are drawn from the June 2012 release of the OECD REGPAT database Webb et al. (2005), Maraut et al. (2008), described in section 3.3.1, which contains $2.4 \times 10^6$ patent applications filed

with the European Patent Office (EPO) from 1960 to the present. In this database the geographical location of each patent inventor and applicant has been matched to one of the appropriate 5,552 regions in one of the 50 OECD or OECD-partner countries. This allows us to construct the geographical networks of patent co-inventorship. (More details in chapter section 3.3).

Starting from these data we define $w_{ij}$ as the number of links between regions $i$ and $j$. In our network $w_{ij}$ will be equal to the number of patents jointly invented by the two regions. We use a full-counting approach so that a patent with $N(> 1)$ inventors accounts for $\sum_{i=1}^{N-1}(N-i)$ regional links (hence, patents with only one inventor do not appear in this network by construction). Therefore, we analyze a weighted undirected network of scientific and technological collaborations across regions. In the co-inventor network the intensity of a link between two regions is equal to the number of patents jointly invented by inventors located in those regions.

Patent data has long been analyzed to measure innovation outcome, just as patent co-inventorship has been used to study the network of innovators within and across national borders. Recently, it has been found that scientific collaborations in Europe are much more constrained by spatial interaction than in the US Crescenzi et al. (2007), Andersson and Gråsjö (2009), Chessa et al. (2013). The European Union clearly represents a real case of transnational network since borders in this case are not only geographical but also political, administrative and cultural (states in the European Union differ by government, legislation, language and even religion). Conversely, state borders in the United States are of a different nature: despite being under the federal system the United States still share the same central government, the

same language and more or less the same culture. Thus we use the US innovation system as a benchmark to estimate the impact of national borders on European network formation.

In order to unveil these differences we compare the European and US co-inventorship networks. Nodes are the NUT3 regions for Europe and the FIPS (Federal Information Processing Standard) geographical units for the USA, which corresponds to counties. (The Nomenclature of Units for Territorial Statistics (NUTS) is a geo-code standard for referencing the subdivisions of countries for statistical purposes. The nomenclature has been introduced by the European Union, for its member states. The OECD provides an extended version of NUTS3 for its non-EU member and partner states). Since we are interested in long-range connectivity across borders, only interactions that took place between different NUTS3 (or FIPS) are taken into consideration. That is to say, we do not consider self-loops in the following analysis. Nevertheless, our approach still naturally extended to the case of directed weighted networks with self-loops.

### 4.3.2 Community detection and core regions

Beyond the local topological features, many networks have groups of nodes marked by the high density of their internal links with respect to the outgoing links that connect the groups with each other. This is especially true if the nodes are embedded in space and subject to geographical constraints that tend to segregate them into spatial communities. This kind of segregation can be even more pronounced if administrative and political boundaries are present; a proper method for detecting possible communities in the network

could be a way to assess the role of external geographical constraints.

Indeed, if geography has such a strong role in link formation, after performing a community detection analysis we would expect to find well-defined communities of spatially connected nodes. However, it has been already shown that geographical clusters and network communities do not perfectly overlap Thiemann et al. (2010). In the following passage we will use modularity optimization introduced by Newman and Girvan Newman and Girvan (2004a) through the "Louvain" algorithm Blondel et al. (2008) and core detection as in De Leo et al. (2013).

## Geographical span and community outreach

The geographical dispersion of a community $s$, or *geographical span* $D_s$, can be measured as Onnela et al. (2011):

$$D_s = \frac{1}{n_s} \sum_{i \in C_s} \sqrt{(X_s - x_i)^2 + (Y_s - y_i)^2} \tag{4.8}$$

where $n_s$ is the number of nodes in the community $C_s$ and $(X_s, Y_s)$ are the coordinates of the geographical center of the community, with $X_s = (1/n_s) \sum_{i \in C_s} x_i$ and $Y_s = (1/n_s) \sum_{i \in C_s} y_i$.

The geographical dispersion is a pure spatial index and does not contain any information about the network structure of the possible links connecting the nodes embedded in space. Since this index is neither normalized nor weighted, it is inadequate for the comparison of different structures, like Europe and the US, where the distances are considerably different. Moreover, the geographic span does not measure how communities reach out. To do this,

we introduce a new index, the *outreach index*, defined as follows:

$$O_s(\mathcal{N}_s) = 1 - \frac{\sum_{i,j\in\mathcal{N}_s} d_{ij}w_{ij}}{\sum_{i,j\in\mathcal{C}_s} d_{ij}w_{ij}},\ O_s(\mathcal{N}_s) \in [0,1] \tag{4.9}$$

where $\mathcal{N}_s$ is the home base of the community $s$ , $d_{ij}$ and $w_{ij}$ are the distance and the weight of the link between nodes $i$ and $j$ and $\mathcal{C}_s$ is the community $s$ as before. The outreach index is defined as the ratio of all the weighted links except for the ones between the nodes $i$ and $j$ which are internal to $\mathcal{N}_s$, but still belonging to $\mathcal{C}_s$, with respect to the same quantity calculated for all pairs of nodes $i$ and $j$ belonging to $\mathcal{C}_s$. Figure 4.21 provides a schematic representation of the way in which the outreach index is obtained.

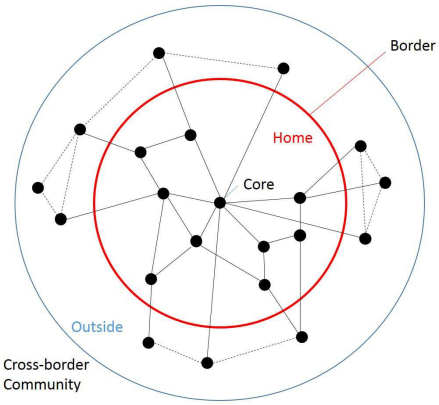



Figure 4.21: **Outreach index.** The outreach index $O_s(\mathcal{N}_s)$ of a cross-border community measures the fraction of cross border and external ties (dashed lines in the plot), weighted by distance and relational intensity. The home base of the community is defined according to three criteria: : simple $|dQ|$, $|dQ| * S$, where $S$ is the node strengh, and internal link density $W_{int}$. When boundaries constrain the span of network communities – as for European R&D collaborations – the outreach index lean towards zero. Conversely, if borders do not affect the shape of network communities – like in the US – $O_s(\mathcal{N}_s) \approx 1$, the topology of the network is conditioned by the presence of borders, which significantly reduces the probability of cross-border connectivity.

Multiple criteria can be used to select the home base $\mathcal{N}_s$:

1. the home base is located in the region with the highest $|dQ|$. The regions with the highest $|dQ|$ can be defined as the core of the community, based on the intensity of intra-community ties.

2. the center of the community can be chosen as the one with the highest $|dQ| * S$, where $S$ is the sum of the weights of all outgoing and incoming links of a node. This index accounts for both the role the node plays in the intra-community connectivity ($|dQ|$) and the overall centrality of the region, as measured by the node strength ($S$).

3. the area that scores the highest internal link density is selected as the home base. Intuitively, this criterion identifies the region with the highest share of inner linkages in the community. In such a case the selected regions will be the biggest ones, regardless of where the core region is located.

In our analysis we find that the above listed criteria tend to provide similar results. Therefore the choice between them depends on the selected community detection method and data availability.

## Results

Table 4.4 reports the value of the *geographical span* on our data. Larger values of $D$ means that the members of the community are geographically spread out; at first glance one could conclude that US community members are more spread out than the European ones on average.

### Distance distribution

It is well known that social interactions negatively depend on distance. More precisely, it has been shown that the probability of a tie between any pair of nodes decays with distance as a power-law $\sim d^{-\alpha}$ where $1 \leq \alpha \leq 2$ Onnela et al. (2011), Lambiotte et al. (2008). Figure 4.22 compares the distance distribution of links in the European and US networks from 1986 to 2009. The two distributions clearly depict different behaviors: the US distribution is well approximated by a power-law as reported in the literature with an exponent $\alpha \approx 1$, whereas the European one shows an exponential behavior.

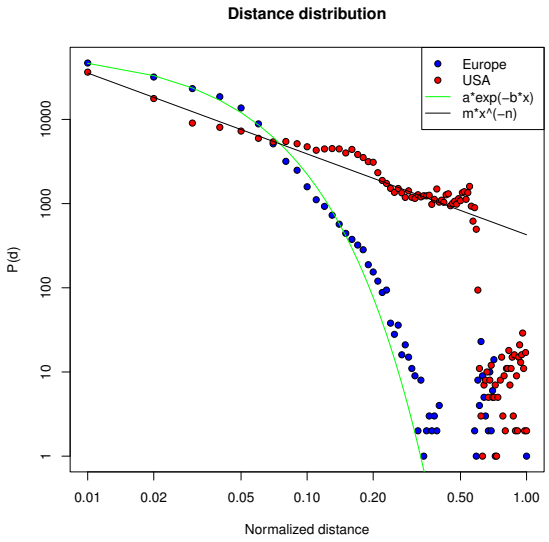



Figure 4.22: Figure shows the distance distribution of the links both for Europe (blue dots) and USA (red dots) in log-log scale and their best fits: a power law $y = \beta d^{-\alpha}$ for the US (black), and exponential distribution $m \exp(-nx)$ for Europe (green), with $\beta = 426.583 \pm 29.474$, eps $\alpha = 0.960 \pm 0.017$, $m = 65030 \pm 570.5$, $n = 33.7 \pm .35$.

Figure 4.23 shows the results of the analysis performed using the modularity method. For the sake of clarity, every node that corresponds to a geographical region, which is geo-referenced and displayed on a map, is given the same color as the community it belongs to. This results in nodes in different communities having different colors as well. Figure 4.23 also shows the results of the core analysis performed on the partition obtained using the Newman-Girvan Modularity. In this representation each community has been given a different color. In the European case, the community structure almost perfectly matches the national boundaries of the Member States of the European Union. The only significant difference seems to be Germany, which

is sectioned off into multiple communities with an average size in the order of a Land Region (NUTS2 level). The US community structure reveals, on the other hand, a practically opposite behavior: communities are stretched out over more than one state, at great distances from the alleged geographical core. For each community, colors are graduated according to the $dQ$ of the node: darker colored nodes have a higher $|dQ|$ and are therefore "more central" while lighter colored ones are less central since they have a lower $|dQ|$. We chose to define the nodes with the highest $|dQ|$ as the community core regions. Combined with our previous results regarding the different decay of connectivity at a distance (power-law in the US, exponential in Europe), we clearly show that the presence of national borders in Europe has a strong role in shaping the topology of the network, both reducing connectivity at a distance and constraining networks community in space.

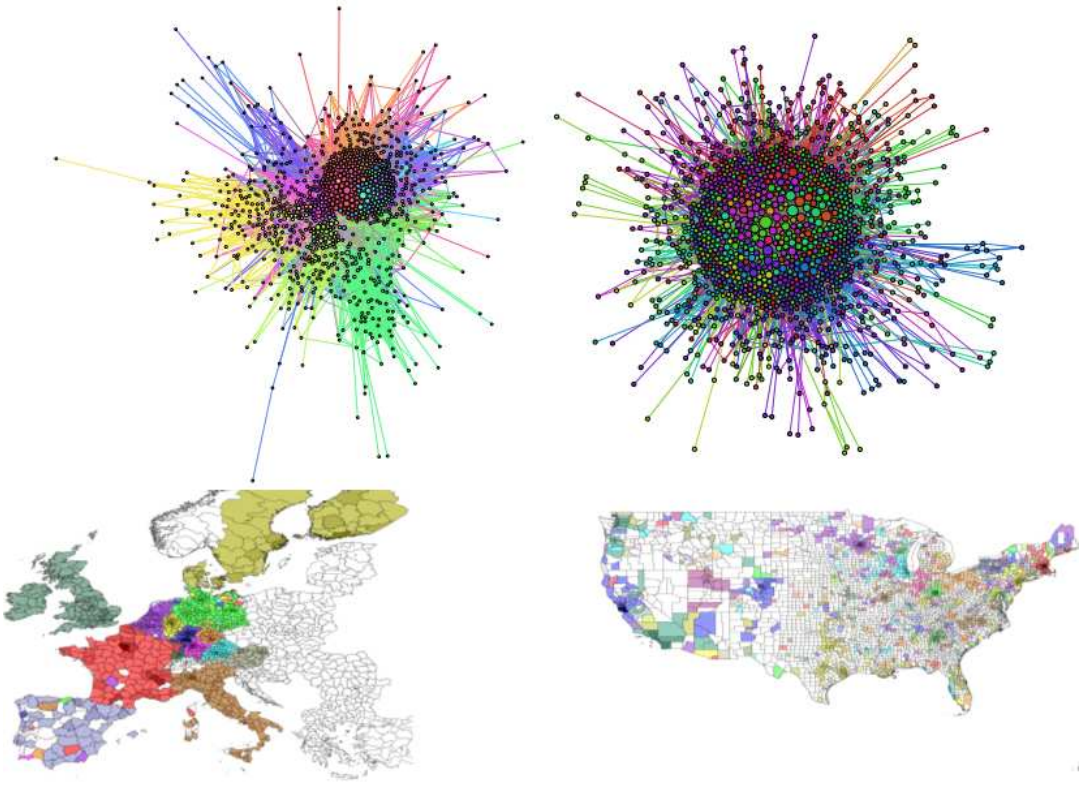


Figure 4.23: Figure shows the results obtained performing a community detection analysis both on the a) European and the b) US networks using the Newman-Girvan method. In this representation each community has been given a different color. It can be seen that, in the European case, the community structure almost perfectly matches the national boundaries of the 15 member states of the European Union. The only significant difference seems to be Germany, which is sectioned off into more than one community. The US community structure reveals almost an opposite behavior: communities are stretched out over more than one state and at great distances. Each community color is graduated according to the $dQ$ of the node: darker colored nodes have a higher $|dQ|$ so they are "more central", while lighter colored ones are less central since they have a lower $|dQ|$. The nodes with the highest $|dQ|$ are considered community core regions. We use the same community color coding for networks (top) and maps (bottom). Maps and networks were generated using the open source software Gephi and QGIS, respectively.

As previously said, Europe is a genuine transnational network whereas the US system is not. Accordingly, their different behaviors do not come as a surprise. On the one hand, since the US innovation system is rather homogeneous, the probability for coast-to-coast interaction (up to a cut off distance of $\sim 10^3$ kilometers) is high. On the other hand, Europe is a collection of almost independent national systems of innovation (see Figure 4.22). Namely, in the European case there is a cut off in the distribution due to strong country border effects, that eventually results in the exponential decay behavior with a characteristic length that is roughly of the size of the average country diameter (about 363 Kilometers). The European network thus differs sharply from the US case, where the state border effect is almost negligible. In the US, scientific and technological communities span throughout the country without any characteristic scale. Moreover we find a power law decay of connectivity at a distance; even when we focus on a single European nation such as Germany, we still note some interesting differences with borders that play a stronger role in reducing connectivity between German Länders when compared to their US counterparts.

Next we proceed to consider the outreach index of the communities. Before doing that, we must determine which one of the three criteria reported in the previous section is the most appropriate. In Table 4.3 we report the home base country of the communities we identify. As one can see, the outcome is the same in all the cases except for the ninth European community, for which we obtained Denmark as the home base according to the first criterion, and Finland for the other two. All in all, it turns out that the final result does not crucially depend on the method we use to identify the home country. There-

fore, the outreach index has an high degree of universality. In the following analysis we will opt for the sensible solution of using a balanced method which takes both the topological centrality inside the community ($|dQ|$) and the total weight ($S$) attached to that node (second criterion) into account. In the cases in which the community detection does not come from a modularity optimization and the $|dQ|$ value is not available, the third criterion can also be considered as a viable alternative.

Table 4.4 reports the value of the outreach index by choosing the home base according to the second criterion. Given that the *outreach index* always lies between 0 and 1, we are allowed to compare the outreach of European communities with US counterparts (see Table 4.4). As expected, the outreach value is about 0 for almost every community in Europe, while this value is always close to 1 in the US. This means that the communities in the United States undertake more outreach than in Europe. However, we should remember that the United States are not a truly transnational network, and accordingly it makes sense to compare the US with Germany as we did before. Indeed, community detection showed that Germany behaves differently and splits into several sub clusters. Then, if we take the NUTS2 level Länd (the German equivalent of a US state) as the reference nation $\mathcal{N}_s$ instead of Germany as a whole (which is NUTS1), the outreach values are sensibly different. They become comparable to the US values ranging from .49 for the region centered around Munich (Oberbayern) to .95 for the region of Mannheim (Karlsruhe).

| Europe | | | | United States | | | |
|---|---|---|---|---|---|---|---|
| Community | $\mathcal{N}_s(dQ)$ | $\mathcal{N}_s(dQ*S)$ | $\mathcal{N}_s(W_{int})$ | Community | $\mathcal{N}_s(dQ)$ | $\mathcal{N}_s(dQ*S)$ | $\mathcal{N}_s(W_{int})$ |
| I | DE | DE | DE | I | CA | CA | CA |
| II | DE | DE | DE | II | NJ | NJ | NJ |
| III | FR | FR | FR | III | MA | MA | MA |
| IV | DE | DE | DE | IV | OH | OH | OH |
| V | DE | DE | DE | V | PE | PE | PE |
| VI | NL | NL | NL | VI | MN | MN | MN |
| VII | DE | DE | DE | VII | IL | IL | IL |
| VIII | UK | UK | UK | VIII | CA | CA | CA |
| IX | DK | FI | FI | IX | TX | TX | TX |
| X | DE | DE | DE | X | OH | OH | OH |
| XI | IT | IT | IT | XI | NC | NC | NC |
| XII | AT | AT | AT | XII | CT | CT | CT |
| XIII | ES | ES | ES | XIII | NY | NY | NY |
| XIV | DE | DE | DE | XIV | GA | GA | GA |

Table 4.3: The table compares the home base of countries found for each community using three different criteria: simple $|dQ|$, $|dQ| * S$, where $S$ is the node strengh, and internal link density $W_{int}$. Results do not vary significantly with the only exception of the Nordic cluster that, when we use $|dQ|$, has its center in Denmark instead of Finland ($|dQ| * S$ and $W_{int}$).

| Europe | | | | | United States | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Community | Core | Country | $D_s$ | $O_s(\mathcal{N}_s)$ | Community | Core | State | $D_s$ | $O_s(\mathcal{N}_s)$ |
| I | Mannheim | DE | 1.9090 | 0.0509 | I | San Jose | CA | 13.8236 | 0.8775 |
| II | Düsseldorf | DE | 1.1016 | 0.0038 | II | New Brunswick | NJ | 7.8733 | 0.9379 |
| III | Paris | FR | 6.0087 | 0.0181 | III | Cambridge | MA | 11.2038 | 0.8911 |
| IV | Berlin | DE | 2.2098 | 0.0477 | IV | Cincinnati | OH | 3.9768 | 0.7525 |
| V | Stuttgart | DE | 1.4729 | 0.0087 | V | Philadelphia | PE | 9.6278 | 0.9792 |
| VI | Eindhoven | NL | 1.5392 | 0.5440 | VI | Minneapolis | MN | 7.9104 | 0.9803 |
| VII | Munich | DE | 1.1793 | 0.0000 | VII | Chicago | IL | 7.9630 | 0.9909 |
| VIII | Cambridge | UK | 2.5560 | 0.1335 | VIII | San Diego | CA | 17.8420 | 0.9288 |
| IX | Helsinki | FI | 6.3203 | 0.6942 | IX | Houston | TX | 6.9423 | 0.9803 |
| X | Nuremberg | DE | 1.5577 | 0.0223 | X | Cleveland | OH | 6.7129 | 0.9734 |
| XI | Milan | IT | 3.4392 | 0.0309 | XI | Raleigh | NC | 6.4801 | 0.9876 |
| XII | Wien | AT | 1.9614 | 0.0537 | XII | New Haven | CT | 11.3220 | 0.9641 |
| XIII | Barcelona | ES | 5.2386 | 0.5065 | XIII | Schenectady | NY | 6.7066 | 0.9718 |
| XIV | Lörrach | DE | 0.2957 | 0.6356 | XIV | Atlanta | GA | 5.7034 | 0.9364 |
| Average Europe | | | 2.6278 | 0.1964 | Average US | | | 8.8634 | 0.9394 |

Table 4.4: The table compares cross-border communities in the US and Europe. Communities are identified based on the main city of the NUTS3 region with the highest $|dQ|$ (core). We report the values of geographical span $D_s$ and the outreach index $O_s(\mathcal{N}_s)$ with respect to home country for European regions and states in the US. Home country has been selected base on the maximum $|dQ| * S$.

### Simulations

The inspection of the outreach index for the US and Europe reveals the existence of four main community types: the French case, which exemplifies European national communities, the Benelux and Nordic clusters, which are two cases of regional integration, and the California case, which is representative of the general behavior of long-range out reaching communities in the United States. In order to uncover the internal mechanics involved in the creation of these cases we reproduced the relevant patterns that emerged from real data by simulating the internal structure of each community.

The artificial model is defined as follows. Out of a total number $N$ of nodes we decide what fraction of them, say $N_i$, to place into the central/home region and what fraction $N_e$ to place into the the external one(s). For the sake of simplicity we choose to shape the regions as circles whose radii are proportional to the number of nodes, so that the more the nodes the bigger the region.

As the regions belonging to the same community can either be adjacent to each other or not, so we introduce the parameter $d$ to regulate the spatial separation between them.

Once the nodes have been placed into communities, we randomly place links between them until the maximum number of links is reached. In general, if $M$ is the total number of possible links $M = N(N-1)/2$ for an undirected network, we determine the density of the network, $\gamma$, as a number between 0 and 1, so that the total number of links will be $P = \gamma M$.

We fine tuned different network densities according to the number of regions that belong to the community and the number of nodes that each

| | Outreach index, simulated | | | | | Parameters | | |
|---|---|---|---|---|---|---|---|---|
| $\mathcal{N}_s$ | $d = 0$ | $d = 10^1$ | $d = 10^2$ | $d = 10^3$ | $d = 10^4$ | $\gamma_{in}$ | $\gamma_{out}$ | $\gamma_{across}$ |
| FR | 0.019 ± .002 | 0.022 ± .003 | 0.034 ± .005 | 0.121 ± .007 | 0.285 ± .019 | 1 | 0 | 0.08 |
| NL | 0.533 ± .021 | 0.542 ± .021 | 0.538 ± .025 | 0.544 ± .022 | 0.534 ± .017 | 1 | 0.1 | 0.10 |
| FI | 0.711 ± .018 | 0.721 ± .015 | 0.722 ± .018 | 0.721 ± .025 | 0.725 ± .011 | 1 | 0.5 | 0.20 |
| CA | 0.887 ± .011 | 0.889 ± .009 | 0.884 ± .009 | 0.889 ± .008 | 0.886 ± .008 | 1 | 0.1 | 0.20 |

Table 4.5: The table reports the values of the simulated outreach index $O$ obtained for the 4 cases FR, NL, FI, CA and for different values of the separating distance $d$. We use here different values of $\gamma$ to differentiate the four cases of the simulation: $\gamma_{in}$ is the network density within borders, $\gamma_{out}$ is the network density outside borders and $\gamma_{across}$ regulates cross-border links between the inner and the outer part of the community.

one of them contains. Thus we have different $P$'s regulating internal links in the central and the external regions, cross-border links between the external regions and the central one and, finally, links between external regions (if more than one).

We then calculated the outreach index for different values of $d$ (see Table 4.5) for these networks. Even when $d = 0$, the case for which there is no separation among the regions (see Figure 4.24), in addition to a set of parameters extracted from the data, the model closely reproduces the spatial organization of the four real cases. As we can also observe in Table 4.5, the simulated outreach indexes are similar to the empirically observed values.

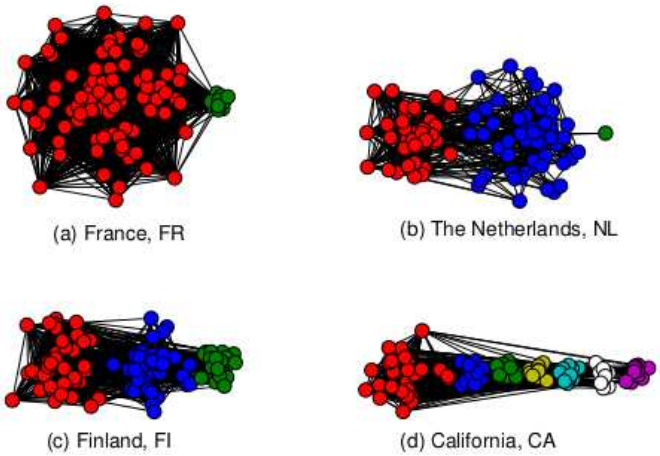



Figure 4.24: Figure shows the simulations reproducing the 4 different cases that we identified as France (FR), the Netherlands (NL), Finland (FI) and California (CA). In the FR case there is a main community, which is very well connected in the inside ($\gamma \sim 1$), with few links that reach out to external regions. The NL and FI cases are intermediate with well-structured external regions that still interact with the central one. In the last case, CA, there is a strong central region with many, and progressively distant, small regions. The FI and CA cases present similar outreach indices due to the fact that, by definition, a huge mass of links in the immediately external regions is equivalent to having just a few interconnected nodes at great distance.

## 4.4 The Rise of China in the International Trade Network: A Community Core Detection Approach (Zhu et al. (2014))

A fast-growing literature has been built in recent years by viewing the international trade system as an interdependent complex network, where countries

are represented by nodes and trade relationships are represented by edges (Serrano and Boguná (2003),Garlaschelli and Loffredo (2005),Fagiolo et al. (2009),De Benedictis and Tajoli (2011),Riccaboni and Schiavo (2010),Riccaboni et al. (2013),Riccaboni M (2014)). As a result, many topics in international economics have been re-investigated through the lens of networks, and globalization and regionalization are certainly no exception. However, even with the networks approach, the question of whether we have a more globalized or regionalized world is still answered with mixed results (Kim and Shin (2002),Tzekina et al. (2008),Piccardi and Tajoli (2012),Reyes et al. (2014)). Moreover, the contribution of network analysis to our understanding of international trade has been questioned, since there is still little evidence about the importance of indirect or network effects on the performances of individual countries (nodes) and trade relationships (edges).

In this work (Zhu et al. (2014)), we re-examine the relationship between globalization and regionalization from a different angle. Instead of assuming that the two are contradictory to each other and attempting to figure out which is dominating the other, we take into account the dynamics in the ITN at both regional level and global level and investigate the interaction between the two. Besides that, we will take advantage of a unique "natural experiment" that is the opening of China to the world trade and the entry of China in the World Trade Organization in 2001, to analyze the reverberations of a huge country-specific shock on the structure of the ITN.

We make use of the CEPII BACI Database (Gaulier and Zignago (2010)) from 1995 to 2011 to build up the ITN: we set countries as nodes and the total bilateral trade flow between countries i and j as the edge weight $A_{ij}$

(see chapter 3 for further information).

We use the modularity optimization method (Newman and Girvan (2004a)) to detect both communities and community cores (De Leo et al. (2013)) in the ITN during the years 1995-2011. The global dynamics are defined as the disappearance or emergence of the communities over time and the regional dynamics are defined as the leadership (community core) change between community members.

We find that the Asia-Oceania community displayed an interesting interaction between the two, which can be roughly summarized in the following three stages:

1. During 1995-2001, the Asia-Oceania community was present (Only with a brief interruption in 1998, when the Asia-Oceania community was integrated with the America community. Also, during 1999-2001, while China was always a member of the Asia-Oceania community, Japan, Oceania, part of the Southeast Asia, and some other Asian economies were integrated with the America community) in the ITN and was led by Japan (During 1999-2001, when Japan was integrated with America, the Asia-Oceania community was led by Hong Kong instead.);

2. During 2002-2004, the Asia-Oceania community disappeared and was integrated with the American community, which was led by the United States;

3. During 2005-2011, the Asia-Oceania community reemerged and was led by China.

Our simulation results show that the disappearance and reemergence of the communities can be generated by a dynamic-edge-weight mechanism for

both inter- and intra-communities. In a network with a fixed number of nodes and a preset initial community structure, each period a node will be selected and by chance it may increase its edge weight with an inter-community node (if the edge already exists; otherwise a new edge will be established). It will then increase its edge weight with an intra-community neighbor. Those neighbors with more inter-community strength will be preferred. In light of the dynamic-edge-weight mechanism, the rise of China in the Asia-Oceania community can be explained by its dramatic increase of inter-community trade since 2002. The intuition is that, the Asia-Oceania community collapsed after China entered the WTO and built strong trade relationships with other communities, especially with the external cores, i.e., the United States and Germany. China then became regionally *attractive* and restored the Asia-Oceania community as the community leader after it gained a significant portion of trade globally.

Our contribution to the analysis of the ITN is twofold. First, we provide some evidence of a deviation from the Barabási-Albert preferential attachment rule (Albert and Barabási (2002)), (Barabasi and Albert (1999)) and the law of gravity (Bergstrand (1985),Baldwin and Taglioni (2006),Carrere (2006)) in the world trade. Second, we identify a mechanism that can account for this deviation and validate it via simulations and empirical analysis. We show that by increasing its global export China is also increasing the chance to import more goods from regional trading partners. In other words, part of the Chinese export growth shock gets transmitted to other economies in the same region by means of a corresponding increase in Chinese imports of intermediate goods and partial delocalization of production. The transmission

mechanism we identify provides further support for a network approach to the analysis of world trade, since we show how local changes in the intensity of trade diffuse to other nodes in the network. We argue that a reductionist approach, which relies exclusively on node and link specific information, misses some important network effects in the world trade structure.

### 4.4.1 Global Dynamics versus Regional Dynamics

During the years 1995-2011 we have examined, the ITN was mainly characterized by three communities, namely, the America community, the Europe community, and the Asia-Oceania community. According to the United Nations definitions of macro geographical regions (See the website of the United Nations Statistics Division https://unstats.un.org/unsd/methods/m49/m49regin.htm), the America community is more or less comprised of Americas. The Europe community is more or less comprised of Europe and Central Asia. The Asia-Oceania community is more or less comprised of Eastern Asia, Southern Asia, South-Eastern Asia, and Oceania. (Countries in Africa and Western Asia don't have consistent community memberships over time. Therefore, they are not classified in any of the three communities.)

However, among the three main communities, the America community and the Europe community were more stable than the Asia-Oceania community. First, over the 17 years, the America community and the Europe community were always present while the Asia-Oceania community experienced disappearance and reemergence. Second, the intra-community structure was more stable in the America community and the Europe community in a sense that the community leaders (cores) over time were always the United States

and Germany, respectively. The Asia-Oceania community on the other hand experienced a leadership change from Japan to China.

Since the Asia-Oceania community has shown rich dynamics both internally and externally, we will focus our attention on it.

### 4.4.2 The Asia-Oceania Community

The dynamics of the Asia-Oceania community can be roughly divided into three stages, namely, its presence with Japan's leadership during 1995-2001, its disappearance and integration with the America community during 2002-2004, and finally its reemergence with China's leadership during 2005-2011.

The same pattern is shown in Figure 4.25, where three years, 1995, 2002, and 2011, are selected to represent the three stages respectively. (See Figure 4.26 for the community detection results for all years and Figure 4.27 for the community core detection results for all years.) The first row shows the community maps in the three years. The America community is colored yellow, the Europe community is colored red, and the Asia-Oceania community is colored blue. Notice that in 2002 the blue community was by and large merged with the yellow community. (Another interesting change in the world trade community structure is the emergence of the Arab community after 2001. This interesting phenomenon deserves further scrutiny in future research.) The second row shows the community core detection results for the three years. The redder the more important the country is in reserving its community. Equivalently, the yellower the less important the country is in reserving its community. This can be used to identify the leaders in the communities. Notice that in 1995 the reddest country in the Asia-Oceania

community was Japan while China became in 2011. Finally, the third row provides a topological view of the community structure in the three years. Again, Japan was central in the Asia-Oceania community in 1995 and it was replaced by China in 2011.

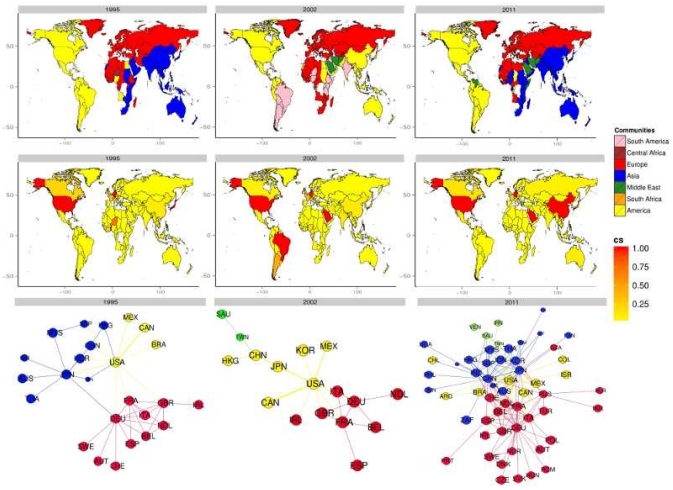



Figure 4.25: From left to right, the three columns are corresponding to the years 1995, 2002, and 2011, respectively. The first row shows the Newman-Girvan community detection results. The America community is colored yellow, the Europe community is colored red, and the Asia-Oceania community is colored blue. Asia-Oceania and America were separated from each other in 1995 and 2011 but was integrated in 2002. The second row shows the community core detection results by normalizing $|dQ| * s$ for each community. The redness of each country is proportional to its relative magnitude of $|dQ| * s$ within its community (CS). The reddest country in the Asia-Oceania community was Japan back in 1995 but became China in 2011. Finally, the third row provides a topological view of the community structure in the three years. Only the edges with no less than 10 million US dollars are shown. Again, Japan was central in the Asia-Oceania community in 1995 and it was replaced by China in 2011.

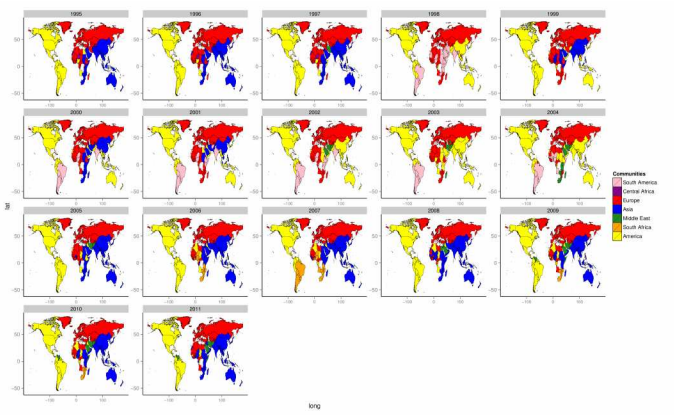



Figure 4.26: Here we show the Newman-Girvan community detection results for the ITN during 1995-2011. The America community is colored yellow, the Europe community is colored red, and the Asia-Oceania community is colored blue. During 1995-2001, the Asia-Oceania community was present (only with a brief interruption in 1998, when the Asia-Oceania community was integrated with the America community). During 2002-2004, the Asia-Oceania community disappeared and was integrated with the American community. Finally, during 2005-2011, the Asia-Oceania community reemerged.

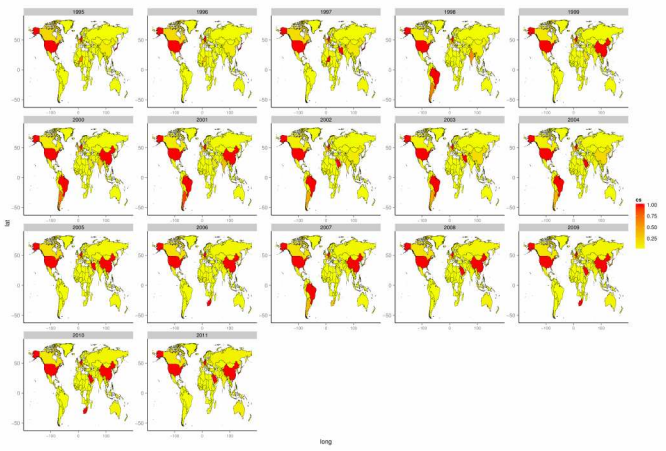


Figure 4.27: Here we show the community core detection results during 1995-2011 by normalizing $|dQ| * s$ for each community. The redness of each country is proportional to its relative magnitude of $|dQ| * s$ within its community (CS). During 1995-2001, the Asia-Oceania was mostly led by Japan (except for 1999-2001, when Japan was integrated with America, the Asia-Oceania community was led by Hong Kong instead). During 2002-2004, the Asia-Oceania community disappeared and was integrated with the American community, which was led by the United States. During 2005-2011, the Asia-Oceania community reemerged and was led by China.

### 4.4.3 The Linkage Between Global and Regional Network Dynamics

Given its breathtaking economic growth during 1995-2011, it is not surprising to see China's rise in the regional trade community. The rationale behind is the long-established gravity model of trade (Bergstrand (1985),Baldwin and Taglioni (2006),Carrere (2006)). That is, the increased economic mass of China tends to attract more trade flows with other economies. What remains unexplained, however, is the fact that the leadership change from Japan to China is correlated with the disappearance and reemergence of the Asia-

Oceania community.

The dynamics observed in the Asia-Oceania community also differ from the prediction of the Barabási-Albert preferential attachment model ((Albert et al., 2000),(Barabasi and Albert, 1999)). According to the preferential attachment mechanism, when choosing another community member with whom the edge weight is to be increased, the given node will prefer the candidates with higher strength. If this is the case, the leadership of Japan in the Asia-Oceania community should be reinforced given that its strength was well ahead of China before 2000. However, Japan was later replaced by China as the community leader. Therefore, we conjecture that not only the magnitude of strength matters but the attributes of nodes such as size and distance also matter in the process of network growth. Moreover, instead of mechanically following an attachment rule, any economic agent plays strategically in choosing its partner to interact with (Riccaboni M (2009),Grossman and Helpman (1997)). Finally, unlike the assumption of the preferential attachment model, countries often have limited resources and competences and cannot freely choose trading partners.

To account for the linkage between the global dynamics and the regional dynamics, we propose a simple dynamic-edge-weight mechanism for both inter- and intra-communities.

### A Simple Mechanism for Both Inter- and Intra-Communities

Since the number of countries in the ITN is constant over time and the evolution of the ITN is only concerned with the trade flows between countries, our model is therefore based on a fixed number of nodes and a dynamic-edge-

weight mechanism for both inter- and intra-communities. (There exists some related literature to our model. For example, (Barrat et al., 2004) and Riccaboni M (2014) examine the network evolution with dynamic edge weights. Li and Maini (2005) investigate the network properties with a preferential attachment mechanism for both inter- and intra-communities. However, to the best of our knowledge, our model is the first attempt to bring the dynamics both inter- and intra-communities to the context of a weighted network with a fixed number of nodes.) Additionally, our model is based on an undirected network because the ITN is constructed by total bilateral trade flows.

The initial status of the network is characterized by $M$ arbitrarily imposed local communities. (In the context of ITN, the communities can be formed, for instance, by continents.) For simplicity, each community has the same number of nodes, $m_0$. As a subgraph, each community is completely connected with a equal edge weight, i.e., every node is connected with every node by the same edge weight in the community. Between any two communities, there is only one edge connecting two randomly selected nodes in the two communities respectively. Again for simplicity, the inter-community edge weight is set to equal the initial intra-community edge weight. After the initial set-up, each period the dynamic-edge-weight mechanism is comprised of the following steps:

1. One node, $i$, is randomly selected based on a uniform distribution across all the nodes in the network;

2. Suppose that $i$ belongs to community $j$, by chance, $i$ can increase its edge weight with a node outside community $j$. And the reach-out probability is $\frac{1}{\alpha}$, where $\alpha \geq 1$ and a big $\alpha$ (In the context of the ITN, a

high value of $\alpha$ can be interpreted as trade barriers such as tariffs, transportation costs, and language difference.) means that any node will have low probability to reach out to other communities;

3. There are $(M-1)m_0$ nodes outside community $j$. They are equally likely to be chosen by $i$ to increase the mutual edge weight. After the inter-community node is identified, the mutual edge weight will be increased by $\beta^{inter}$;

4. The next step for $i$ is to choose a neighbor in the same community $j$ to increase the edge weight. The neighbor is selected by the following probability mass function:

$$P_{-i,j}^{intra} = \frac{(1-\gamma)s_{-i,j}^{intra} + \gamma \sum_{-j} s_{-i,-j}^{inter}}{(1-\gamma)\sum_{-i} s_{-i,j}^{intra} + \gamma \sum_{-j}\sum_{-i} s_{-i,-j}^{inter}} \tag{4.10}$$

where $-i$ is a neighbor to $i$ in the community $j$ and $-j$ is a community other than community $j$. $0 \leq \gamma \geq 1$ and when gets close to 1, although $i$ prefers to increase the edge weight with the neighbors with more intra-community strength, it prefers even more the ones with more inter-community strength. After the neighbor is identified, the mutual edge weight will be increased by $\beta^{intra}$;

5. Finally, the modularity optimization method is used to detect the community structure, which may deviate from the original set-up.

**Simulation Results**

The initial status of our simulation is a network with 3 preset communities. Each community has 5 nodes and, as mentioned above, each community is

completely connected and there is a single edge between any two communities. Other model parameters are $\alpha = 40$, $\beta^{intra} = 0.05$, and $\beta^{inter} = 2$, respectively. Setting $\alpha$ to 40 and having a relatively big $\beta^{inter}$ compared to $\beta^{intra}$ are to make it difficult for a node to reach out to other communities so that the preset community structure can be restored over time. However, when a node does reach out, it is enough to introduce a perturbation to the community structure. Finally, we vary the value of $\gamma$ from 0.1 to 0.9 with the step size of 0.05.

We define a trial of simulation as running the above dynamic-edge-weight mechanism for 5000 periods. We also calculate the percentage of the number of the periods with exactly the same community structure as the initial status out of the 5000 periods as an indicator of the community structure stability of the network. For each value of , we collect a sample size of 100 trials to compute the 95% confidence interval of the estimated original community structure percentage. The result is reported in Figure 4.28. As expected, putting more weight on the neighbors with more inter-community strength (i.e., bigger ) tends to make the community structure more stable (i.e., bigger original community percentage). The intuition is that the reaching-out nodes will be dragged back to their original communities by the preference for their growing inter-community strength.

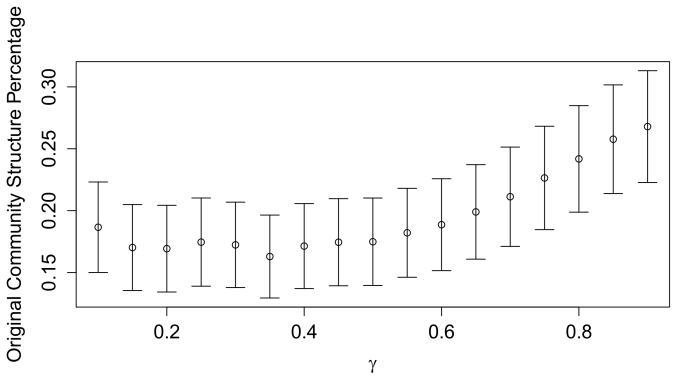



Figure 4.28: The 95% confidence interval is calculated for each value of from 0.1 to 0.9 with the step size of 0.05 (the x-axis). The simulation is based on a dynamic-edge-weight mechanism for both inter- and intra-communities. Other model parameters are $\alpha = 40$, $\beta^{intra} = 0.05$, and $\beta^{inter} = 2$, respectively. We define a trial of simulation as running the dynamic-edge-weight mechanism for 5000 periods. As an indicator of the community structure stability of the network, the y-axis is the percentage of the number of the periods with exactly the same community structure as the initial status out of the 5000 periods. Finally, for each value of , we estimate the confidence interval of the original community structure percentage by collecting a 100-trial sample.

As a detailed example of the simulation, Figure 4.29 selects 4 periods of a single trial. The 3 preset communities are X1-X5, X6-X10, and X11-X15, respectively. Different colors represent different communities detected by the modularity optimization method. The red edges are inter-community ones while the black ones are intra-community. Like what we observe from the ITN, the disappearance and reemergence of the communities can be generated by the dynamic-edge-weight mechanism for both inter- and intra-communities.

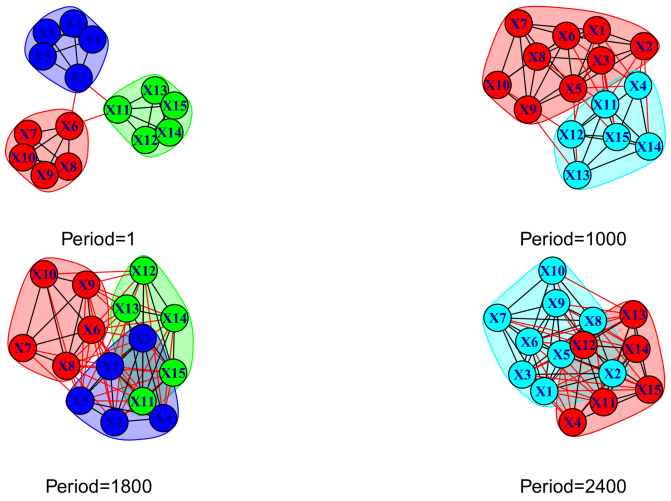



Figure 4.29: The figure is based on a single trial of simulation. Different colors represent different communities detected by the Newman-Girvan method. The inter-community edges are colored red while the intra-community ones are colored black. Although the community detection takes into account the edge weights, all the edges in the figure have the same width. In period 1, three predetermined communities, X1-X5, X6-X10, and X11-X15, are imposed in the network. The number of communities detected in this 15-node network bounces back and forth between 3 and 2 during the simulated periods. That is, like what we observe from the ITN, the disappearance and reemergence of the communities can be generated by the dynamic-edge-weight mechanism for both inter- and intra-communities.

### Empirical Evidence

We now turn back to the ITN and present some empirical evidence for the dynamic-edge-weight mechanism for both inter- and intra-communities.

First, for the inter-community dynamics, we calculate the ratio of the inter-community trade to the intra-community trade between the Asia-Oceania community and the America community. As shown in Figure 4.30, the ratio first went up and then went down and formed a hump shape over time.

This finding coincides with the disappearance and reemergence of the Asia-Oceania community observed in Figure 4.25. In 1995, when the Asia-Oceania community was present, the inter-community trade between Asia-Oceania and America was about 44% of the intra-community trade within the two communities. In 2002, when the Asia-Oceania community disappeared, the ratio went up to about 51%. Finally, the ratio went back to about 43% in 2011, when the Asia-Oceania community was present again.

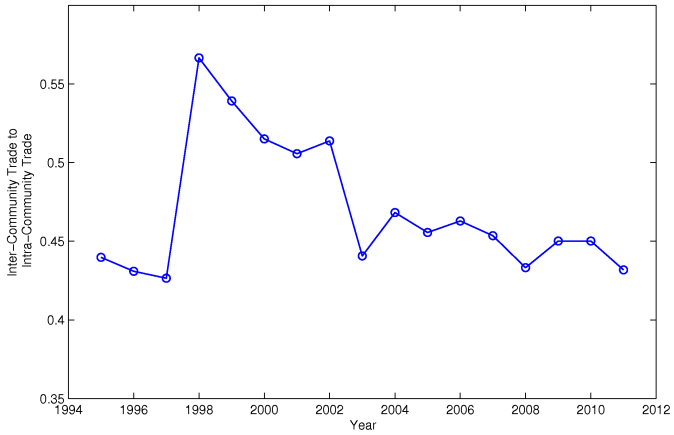



Figure 4.30: We calculate the ratio of the inter-community trade to the intra-community trade between the Asia-Oceania community and the America community. The ratio first went up and then went down and formed a hump shape over time. This finding coincides with the disappearance and reemergence of the Asia-Oceania community observed in Figure 4.25.

Second, for the intra-community dynamics, we compare the intra-community strength and the inter-community strength between Japan and China. As shown in Figure 4.31, before 2003, Japan always had more inter-community trade than China and had more intra-community trade in the beginning and slightly less later. After 2003, China surpassed Japan in terms of both inter-

and intra-community trade. This finding coincides with the leadership change from Japan to China observed in Figure 4.25. Also notice that, for both countries, the intra-community trade follows closely to the inter-community trade, which can be considered as evidence of the intra-community dynamic-edge-weight mechanism.

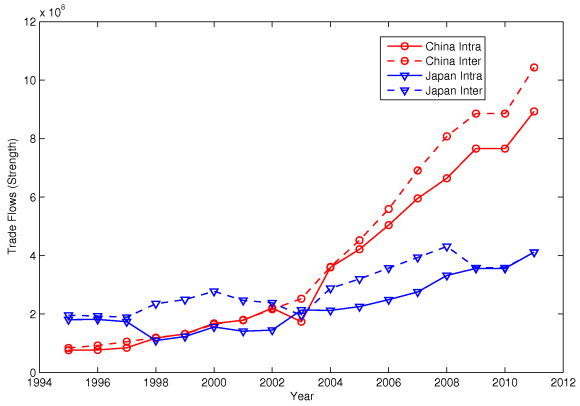



Figure 4.31: We calculate both the inter- and intra-community trade volumes for Japan and China. Japan had more inter-community trade than China before 2003. However, after 2003, China surpassed Japan in terms of both inter- and intra-community trade. This finding coincides with the leadership change from Japan to China observed in Figure 4.25. Furthermore, for both countries, the intra-community trade follows closely to the inter-community trade, which can be viewed as evidence of the intra-community dynamic-edge-weight mechanism.

We also check the regional trade agreements (RTAs) for the intra-community dynamics. Table 4.32 summarizes the effective RTAs signed with China during 1995-2011. Only after its accession to WTO in the end of 2001, China started to form RTAs in 2003 and with countries almost exclusively in the Asia-Oceania community.

| RTA Name | Date of Entry into Force |
|---|---|
| China - Hong Kong, China | 29-Jun-2003 |
| China - Macao, China | 17-Oct-2003 |
| ASEAN - China | 01-Jan-2005(G); 01-Jul-2007(S) |
| Chile - China | 01-Oct-2006(G); 01-Aug-2010(S) |
| Pakistan - China | 01-Jul-2007(G); 10-Oct-2009(S) |
| China - New Zealand | 01-Oct-2008 |
| China - Singapore | 01-Jan-2009 |
| Peru - China | 01-Mar-2010 |
| China - Costa Rica | 01-Aug-2011 |

This table has all the effective RTAs involving China during 1995–2011. (G) stands for Goods and (S) for Services. The data is extracted from the WTO website, http://rtais.wto.org/UI/PublicAllRTAList.aspx.
doi:10.1371/journal.pone.0105496.t001

Figure 4.32: China's Effective RTAs.

Last but not least, it is a well observed fact that the Asia-Oceania community is an active participant of the global production chain (or global value chain) (Athukorala (2005),Athukorala and Yamashita (2006),Baldwin (2008)). Therefore, the intra-community preference over the nodes with more inter-community strength can be understood as the incentive to have better market access through the regional big player in the global production chains.

## 4.5 World Input-Output Network (Cerina et al.)

As the global economy becomes increasingly integrated, an isolated view based on the national input-output table is no longer sufficient to assess an individual economy's strength and weakness, not to mention finding solutions to global challenges such as climate change and financial crises. Hence, a multi-regional input-output (MRIO) framework is needed to draw a high-

resolution representation of the global economy (Wiedmann et al. (2011)).

In practice, however, due to the expensive process of collecting data and the variety of classifications used by different agencies, for a long time, the input-output tables have only been available for a limited number of countries and for discontinuous years. Fortunately, the fully-fledged MRIO databases started to become available in recent years (Tukker and Dietzenbacher (2013)). Unlike the national input-output table where exports and imports are aggregated and appended to final demand and country-specific value added respectively, for each individual economy, the MRIO table splits its exports into intermediate use and final use in every foreign economy and also traces its imports back to the industry origins in every foreign economy. As a result, the inter-industrial relationships in the MRIO table are recorded not only within the same economy but also across economies.

In this work we move forward by considering the global MRIO system as a world input-output complex network (WION), where the nodes are the individual industries in different economies and the edges are the monetary goods flows 3 between industries, similarly to what have been done recently by Acemoglu, Carvalho, Ozdaglar and Tahbaz-Salehi for the US economy only (Acemoglu et al. (2012)). However, to the best of our knowledge, our work is the first attempt to explore the MRIO tables from a networks perspective, even though there have been some networks studies of the input-output tables at the national level and for selected countries (Blöchl et al. (2011),McNerney et al. (2013),Contreras and Fagiolo (2014)).

Different from many network systems observed in reality, the WION has the following features: 1) It is directed and weighted, i.e., an industry can act

as both a seller and a buyer at the same time and the monetary goods flows between industries vary a lot; 2) It is much denser within the same economy than across economies, i.e., despite the continuously integrated global economy, most economic transactions still happen within the country border; 3) It is with strong self-loops, i.e., an industry can acquire a significant amount of inputs from itself. Taking into account the features above, we explore the WION by quantifying not only some global network properties such as assortativity but also some local network properties such as PageRank centrality. Furthermore, we apply community detection and community core detection techniques to examine the structure of the WION over time.

This work makes some significant contributions to the literature of input-output economics. First, it is the first attempt to quantify the network properties of the WION by taking into account its edge weights and directedness. By doing that, we detect a marked increase in cross-country connectivity, apart from a sharp drop in 2009 due to the financial crisis. Second, the community detection results reveal growing input-output international communities. Among them, we notice in particular the emergence of a large European community led by Germany. Third, we use the network-based PageRank centrality and community coreness measure to identify the key industries and economies in the WION and the results are different from the one obtained by the traditional final-demand-weighted backward linkage measure.

In the following we will quantify some global network properties of the WION and its subgraph structure and dynamics by using community detection techniques. Moreover, we use the network-based PageRank centrality and community coreness measure to identify the key industries in the WION.

### 4.5.1 From WIOD to WION

We use the World Input-Output Database (WIOD) Timmer et al. (2012) to map out the WION (see section 4.5 for the lists of countries and industries covered in the WIOD) from 1995 to 2011. For each year, there is a harmonized global level input-output table recording the input-output relationships between any pair of industries in any pair of economies. Table 4.6 shows an example of a MRIO table with two economies and two industries. The $4 \times 4$ inter-industry table is called the transactions matrix and is often denoted by $\mathbf{Z}$. The rows of $\mathbf{Z}$ record the distributions of the industry outputs throughout the two economies while the columns of $\mathbf{Z}$ record the composition of inputs required by each industry. Notice that in this example all the industries buy inputs from themselves, which is often observed in real data. Besides intermediate industry use, the remaining outputs are absorbed by the additional columns of final demand, which includes household consumption, government expenditure, and so forth [1]. Similarly, production necessitates not only inter-industry transactions but also labor, management, depreciation of capital, and taxes, which are summarized as the additional row of value added. The final demand matrix is often denoted by $\mathbf{F}$ and the value added vector is often denoted by $\mathbf{v}$. Finally, the last row and the last column record the total industry outputs and its vector is denoted by $\mathbf{x}$.

Aa complex networks approach has been widely used in economics and finance in recent years Kitsak et al. (2010), Pammolli and Riccaboni (2002), Riccaboni and Schiavo (2010), Riccaboni et al. (2013), Chessa et al. (2013), Caldarelli et al. (2013). Designed to keep track of the inter-industrial rela-

[1] Here we only show the aggregated final demand for the two economies

| | | Buyer Industry | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Economy 1 | | Economy 2 | | Final Demand | | |
| Seller Industry | | Industry 1 | Industry 2 | Industry 1 | Industry 2 | Economy 1 | Economy 2 | Total Output |
| Economy 1 | Industry 1 | 5 | 10 | 20 | 10 | 45 | 10 | 100 |
| | Industry 2 | 10 | 5 | 10 | 20 | 50 | 5 | 100 |
| Economy 2 | Industry 1 | 30 | 15 | 800 | 500 | 5 | 8650 | 10000 |
| | Industry 2 | 35 | 30 | 1000 | 1000 | 25 | 7910 | 10000 |
| Value Added | | 20 | 40 | 8170 | 8470 | | | |
| Total Output | | 100 | 100 | 10000 | 10000 | | | |

Table 4.6: Table shows a hypothetical two-economy-two-industry MRIO table. The $4 \times 4$ inter-industry transactions matrix records outputs selling in its rows and inputs buying in its columns. The additional columns are the final demand and the additional row is the value added. Finally, the last column and the last row record the total industry outputs. The numbers are made up in such a way that Economy 2 is a lot larger than Economy 1 in terms of industry outputs. However, as shown below, an unweighted backward linkage measure will consider industries in Economy 1 more important than the ones in Economy 2. Hence, we adopt a final-demand-weighted backward linkage measure to identify the key industries in the WIOD.

tionships, the input-output system is an ideal test bed for network science. Particularly the MRIO system can be viewed as an interdependent complex network, i.e., the WION, where the nodes are the individual industries in different economies and the edges are the monetary goods flows between industries.

Figure 4.33 provides a topological view of Table 4.6. The blue nodes are the individual industries. The red nodes are the value added sources from the two economies, whereas the green nodes are the final demand destinations in the two economies. The edges are with arrows indicating the directions of the monetary goods flows and with varying widths indicating the magnitudes of the flows. The color of the edge is set the same as the source node's. Finally, because we are only concerned with the inter-industrial input-output relationships, when formulating the WION, we focus our attention on the network among the blue nodes.

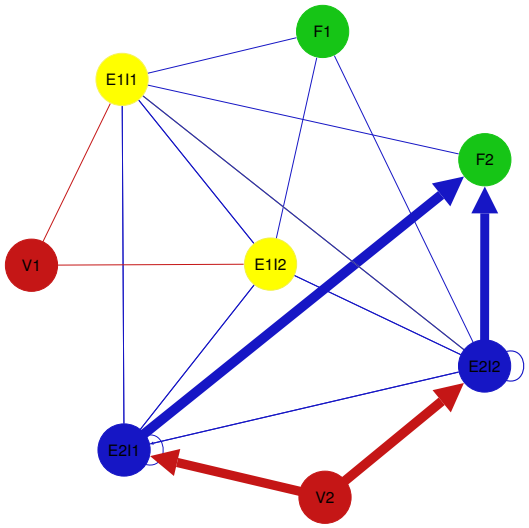
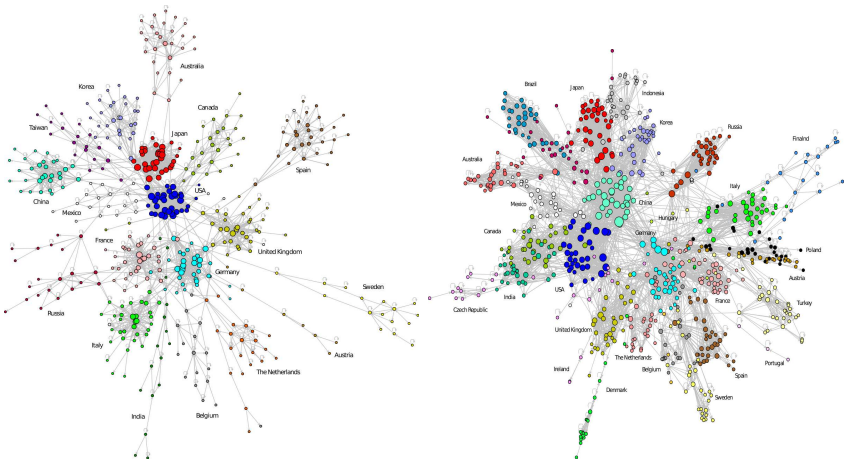



Figure 4.33: A hypothetical two-economy-two-industry WION. This is a topological view of Table 4.6. The blue nodes are the individual industries. The label "ExIy" should read "Industry y in Econoomy x". The red nodes are the value added sources from the two economies, whereas the green nodes are the final demand destinations in the two economies. The label "Vx" should read "Value Added from Economy x", whereas the label "Fx" should read "Final Demand in Economy x". The edges are with arrows indicating the directions of the monetary goods flows and with varying widths indicating the magnitudes of the flows. The color of the edge is set the same as the source node's. Finally, because we are only concerned with the inter-industrial input-output relationships, when formulating the WION, we focus our attention on the network among the blue nodes.

The visualization of the WION in 1995 and in 2011 are shown in Figure 4.34. Each node represents a certain industry in a certain economy. The size of the node is proportional to its total degree. The edges are directed and only

those with strength greater than one billion US dollars are present. Finally, different colors represent different economies. Clearly the WION has become denser over time and some countries like China have moved to the core of the network, thus confirming the results in Table 4.7.




(a) 1995 (b) 2011

Figure 4.34: Figure shows the WION in 1995 and in 2011. Each node represents a certain industry in a certain economy. The size of the node is proportional to its total degree (number of edges). The edges are directed and only those with strength greater than 1000 millions of US dollars are present. Finally, different colors represent different economies.

Table 4.7 has identified the top 20 industries for the years 1995, 2003, and 2011, respectively. The first column of each year is produced by the final-demand-weighted backward linkage measure (see appendix A), i.e., $\mathbf{w}$. For the selected years, only four large economies, China, Germany, Japan, and USA, ever qualified for the top 20. Another noticeable change over time is the rise of China, which topped the list in 2011 with its industry of construction.

| | 1995 | | | 2003 | | | 2011 | | |
|---|---|---|---|---|---|---|---|---|---|
| **Rank** | **w** | $PR$ | $\lvert dQ \rvert$ | **w** | $PR$ | $\lvert dQ \rvert$ | **w** | $PR$ | $\lvert dQ \rvert$ |
| **1** | USA-Pub | USA-Pub | USA-Pub | USA-Pub | USA-Hth | USA-Obs | CHN-Cst | GBR-Hth | CHN-Cst |
| **2** | JPN-Cst | USA-Tpt | JPN-Cst | USA-Hth | DEU-Tpt | USA-Est | USA-Pub | DEU-Tpt | USA-Obs |
| **3** | USA-Cst | DEU-Tpt | USA-Obs | USA-Cst | USA-Pub | USA-Fin | USA-Hth | USA-Pub | CHN-Met |
| **4** | USA-Hth | USA-Hth | USA-Cst | USA-Est | USA-Tpt | USA-Pub | USA-Est | CHN-Elc | USA-Pub |
| **5** | USA-Est | DEU-Cst | USA-Est | USA-Rtl | GBR-Hth | USA-Hth | CHN-Elc | USA-Hth | USA-Est |
| **6** | USA-Rtl | RUS-Hth | USA-Hth | CHN-Cst | ESP-Cst | CHN-Cst | USA-Rtl | CHN-Cst | CHN-Agr |
| **7** | USA-Fod | DEU-Fod | JPN-Htl | JPN-Cst | DEU-Hth | JPN-Cst | USA-Cst | CHN-Met | CHN-Fod |
| **8** | JPN-Pub | GBR-Cst | JPN-Met | USA-Fin | GBR-Cst | USA-Ocm | USA-Fin | USA-Tpt | USA-Fin |
| **9** | USA-Tpt | USA-Cst | USA-Met | USA-Tpt | USA-Cst | CHN-Met | CHN-Fod | ESP-Cst | CHN-Min |
| **10** | JPN-Est | FRA-Tpt | JPN-Obs | USA-Fod | USA-Obs | USA-Met | CHN-Mch | AUS-Cst | USA-Cok |
| **11** | JPN-Hth | USA-Fod | DEU-Cst | USA-Htl | FRA-Tpt | JPN-Obs | JPN-Cst | ITA-Hth | CHN-Elc |
| **12** | USA-Fin | GBR-Hth | JPN-Pub | USA-Ocm | TUR-Tex | JPN-Htl | USA-Fod | DEU-Hth | USA-Hth |
| **13** | USA-Htl | USA-Obs | JPN-Hth | JPN-Pub | USA-Est | CHN-Agr | USA-Htl | USA-Obs | CHN-Omn |
| **14** | JPN-Fod | JPN-Cst | JPN-Ocm | USA-Obs | AUS-Cst | USA-Cst | CHN-Tpt | RUS-Hth | CHN-Cok |
| **15** | JPN-Rtl | DEU-Mch | JPN-Fod | JPN-Est | USA-Fod | JPN-Pub | JPN-Pub | CHN-Tpt | CHN-Mch |
| **16** | DEU-Cst | ESP-Cst | JPN-Fin | JPN-Hth | ITA-Hth | USA-Agr | USA-Tpt | DEU-Mch | USA-Cst |
| **17** | JPN-Elc | JPN-Tpt | USA-Agr | USA-Whl | DEU-Cst | USA-Tpt | USA-Ocm | FRA-Cst | CHN-Chm |
| **18** | JPN-Whl | DEU-Met | USA-Fod | JPN-Tpt | DEU-Fod | JPN-Met | CHN-Pub | CHN-Tex | JPN-Obs |
| **19** | JPN-Tpt | USA-Elc | JPN-Whl | CHN-Elc | DEU-Mch | JPN-Hth | USA-Obs | GBR-Cst | USA-Ocm |
| **20** | JPN-Mch | USA-Est | USA-Pup | DEU-Tpt | CHN-Elc | CHN-Elc | USA-Whl | DEU-Met | JPN-Cst |

Table 4.7: **Top 20 Industries Identified by the Three Methods for Selected Years.** The first method is the final-demand-weighted backward linkage measure (see appendix A), **w**. The second is the PageRank centrality, $PR$. The third is the community coreness measure $\lvert dQ \rvert$.

### 4.5.2 Global Network Properties of the WION

Because the WION is directed, we can calculate the assortativity coefficient in three ways, namely, in-degree assortativity, out-degree assortativity, and total-degree assortativity. As shown in Figure 4.35, they all behave similarly over time.

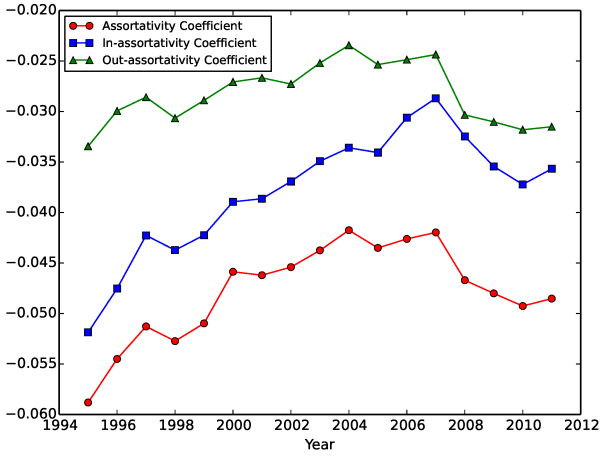
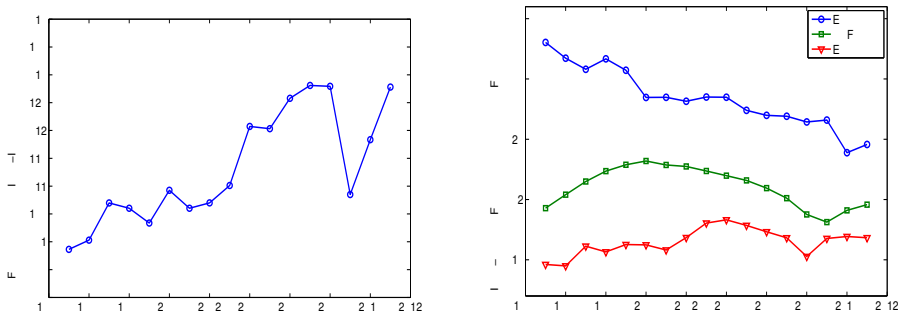



Figure 4.35: From top to bottom, we show the over time out-degree assortativity, in-degree assortativity, and total-degree assortativity of the WION, respectively.

First, they have all been negative throughout the whole period. Since assortativity measures the tendencies of nodes to connect with other nodes that have similar (or dissimilar) degrees as themselves, a negative coefficient means that dissimilar nodes are (slightly) more likely to be connected. One possible explanation of the negativity is that high-degree industries such as construction often take inputs (or supply outputs) from (or to) low-degree industries such as transport services. Moreover, the spatial constraints (each node has only few neighboring nodes in the same country) introduce degree-

degree anticorrelations (disassortativity) since high degree sectors are in different countries and the probability to connect decays with distance (Emmerich et al. (2014)). Second, all the coefficients show an increasing trend before 2007 and a significant decline after 2007. The behavior of the assortativity measures seems to be correlated with the trend of the foreign share in the inter-industrial transactions over time (Figure 4.36). That is, we can calculate a globalization indicator as the percentage of inputs from foreign origins (or equivalently, the percentage of outputs to foreign destinations) of the transactions matrix $\mathbf{Z}$ of the 40 WIOD economies. Same as observed in assortativity, the foreign share of $\mathbf{Z}$ had a steady growth (from 9.9% in 1995 to 12.8% in 2007) before 2007 and a sharp decrease after 2007. [2]

[2]While the most severely depressed domestic edges during 2008-2009 in terms of the magnitude of the reduced flows are mostly within USA, the top 3 most impacted foreign edges are all from the mining industry to the coke and fuel industry and are from Canada to USA, from Netherlands to Belgium, and from Mexico to USA, respectively.



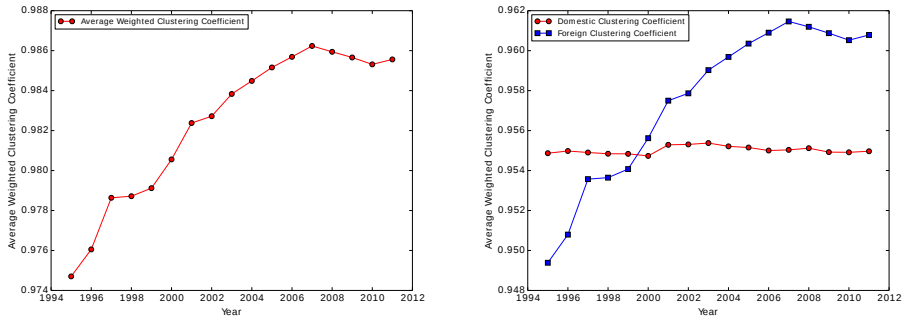

(a) All 40 economies (b) Selected regions

Figure 4.36: Panel (a) shows the foreign share of the transactions matrix **Z** over time. We calculate the percentage of inputs from foreign origins (or equivalently, the percentage of outputs to foreign destinations) of the transactions matrix **Z** of the 40 WIOD economies. It can be viewed as a globalization indicator becasue it measures how much inter-industrial transactions are made through international trade. Panel (b) considers the intra-region foreign share out of the total foreign share for some regions classified in Table 3.1 in chapter 3. For the three regions, Euro Zone relies on the intra-region foreign trade the most and East Asia the least. Moreover, while the intra-region share in the other two regions fluctuates over time, it almost always declines in Euro Zone. Finally, all the three regions became less dependent on the intra-region foreign trade before the 2008 crisis. After the crisis, East Asia increased the intra-region foreign trade immediately, which is followed by NAFTA, and then by Euro Zone.

The increase in the foreign share implies more interactions across economies and hence tends to make the WION less dissortative. The opposite happens when the foreign share goes down as a result of the global financial crisis. Third, we notice that the in-degree assortativity tends to be lower than the out-degree assortativity, but there is a tendency to close the gap between the two measures. We interpret this evidence as a clear signal of the globalization of production chains, that is to say, both global buying and selling hubs have

now a higher chance to be connected across borders.

The hump-shaped behavior is also observed in the clustering coefficient. Figure 4.37 shows that the average weighted clustering coefficient of the WION has been steadily increasing but was followed by a decline since 2007. Again, a possible explanation is that the booming economy before 2007 introduced more interactions between industries, hence higher clustering coefficient, and the financial crisis after 2007 stifled the excess relationships.




(a) Clustering coefficient (b) Domestic and foreign

Figure 4.37: Panel (a) shows the average weighted clustering coefficient of the WION over time. Panel (b) further decomposes the clustering coefficient into domestic clustering coefficient and foreign clustering coefficient. Clearly the behavior in Panel (a) is more explained by the foreign clustering coefficient.

We can also examine the global network properties of the WION by plotting its degree and strength distributions. As shown in Figure 4.38, unlike other network systems such as the internet, where the degree distributions follow the power law, the WION is characterized by the highly left-skewed degree distributions. Most nodes enjoy high-degree connections in the WION because the industries are highly aggregated. That is, it is hard to find two

completely disconnected industries given the high level of aggregation. Furthermore, the WION is almost complete, i.e., every node is connected with almost every node, if represented by unweighted edges.[3]

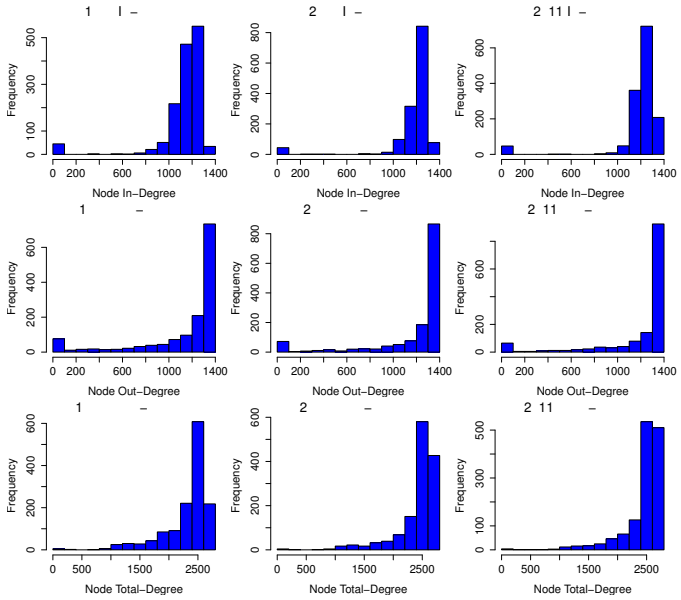



Figure 4.38: Figure shows the histogram of in-degree, out-degree, and total-degree distributions for selected years. For the selected years 1995, 2003, and 2011, the first row has the in-degree distributions while the second row and the third row have the out-degree and total-degree distributions respectively. The WION is characterized by the highly left-skewed degree distributions. Most nodes enjoy high-degree connections in the WION due to the aggregated industry classification.

We can also take into account the edge weights and examine the strength distributions of the WION. Figure 4.39 shows the in-strength, out-strength, and total-strength distributions for the years 1995, 2003, and 2011. Like the

[3]The same feature is also found in the input-output networks at the national level (McNerney et al. (2013)). Using a single-year (2006) data of the WIOD, Carvalho (2013) also reports the heavy-tailed but non-power-law degree distributions.

previous studies at the national level (McNerney et al. (2013)), the strength distributions can be well approximated by the log-normal distributions. As reasoned by Acemoglu et al. (2012), this asymmetric and heavy-tailed distribution of strength in the WION may serve as the origin of economic fluctuations.

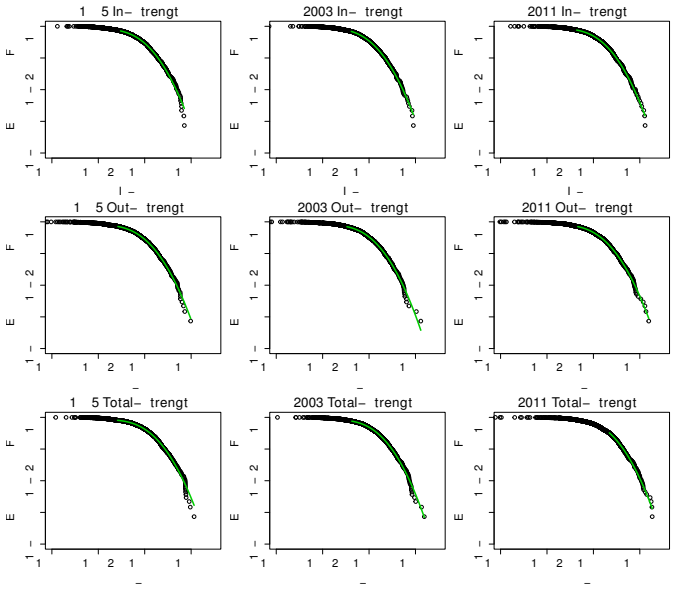



Figure 4.39: Figure shows the empirical counter-cumulative distribution functions of in-strength, out-strength, and total-strength for selected years. For the selected years 1995, 2003, and 2011, the first row has the in-strength distributions while the second row and the third row have the out-strength and total-strength distributions respectively. The observed data are in black circles while the green curve is the fitted log-normal distribution.

### 4.5.3 The Community Detection in the WION

Figures 4.40, 4.41, and 4.42 report the community detection results, obtained with Newman-Girvan modularity optimization, for the selected years 1995, 2003, and 2011, respec- tively. The 40 countries in the WIOD are arranged

by rows while the 35 industries are arranged by columns. Different colors indicate different communities detected. There are two interesting findings in our results. First, most communities were based on a single economy, i.e., the same color often goes through a single row. This echoes one of the features of the WION mentioned previously, i.e., most of the inter-industrial activities are still restricted in the country border. Second, for all the three years selected, we always color the community involving Germany in red and put it on the top. As a result, our algorithm captures a growing Germany-centered [4] input-output community.

Since the WIOD monetary goods flows are based on undeflated current prices, one possible reason for the emergence of the German community is that the community members may have experienced significantly more inflation and/or exchange rate volatility than other regions in the world. Referring to the World Bank inflation data and the exchange rate data used in the WIOD, we show that this is hardly the case.

[4] It is centered on Germany because the community core detection results below show that the cores of this red community are all within Germany.

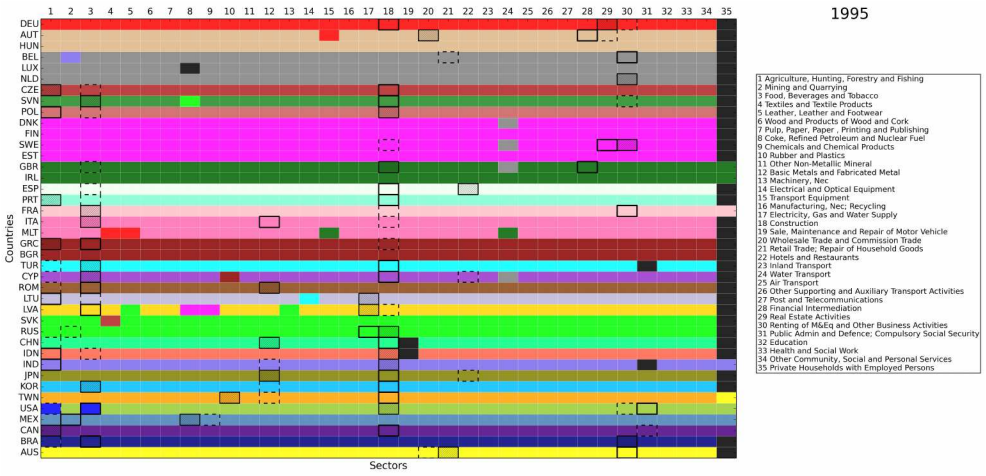



Figure 4.40: Figure shows community detection and community core detection results in 1995. The economies are arranged by rows and the industries are arranged by columns. Each color represents a community detected, except that the black color indicates the isolated nodes with only self-loop. Within each community, the top 3 core economy-industry pairs are identified. The first place is with thick and solid border. The second place is with thick and dashed border. The third place is with border and texture.

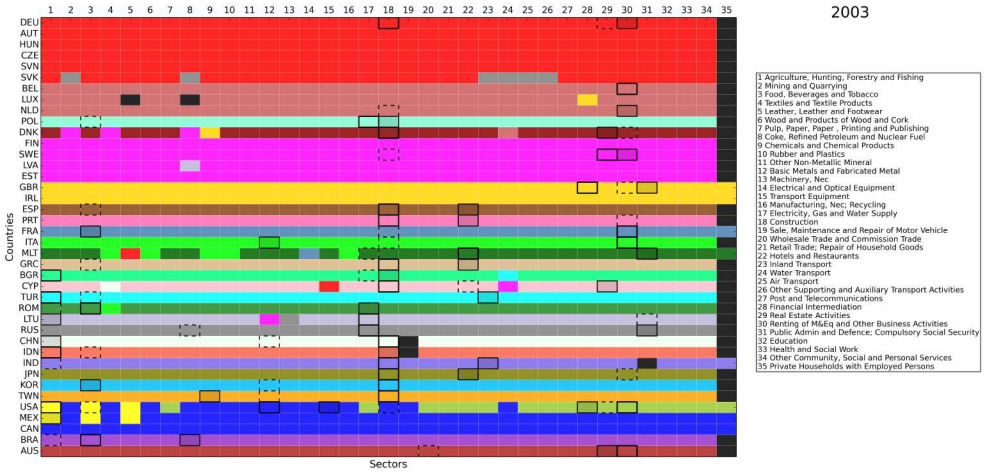



Figure 4.41: Figure shows community detection and community core detection results in 2003.

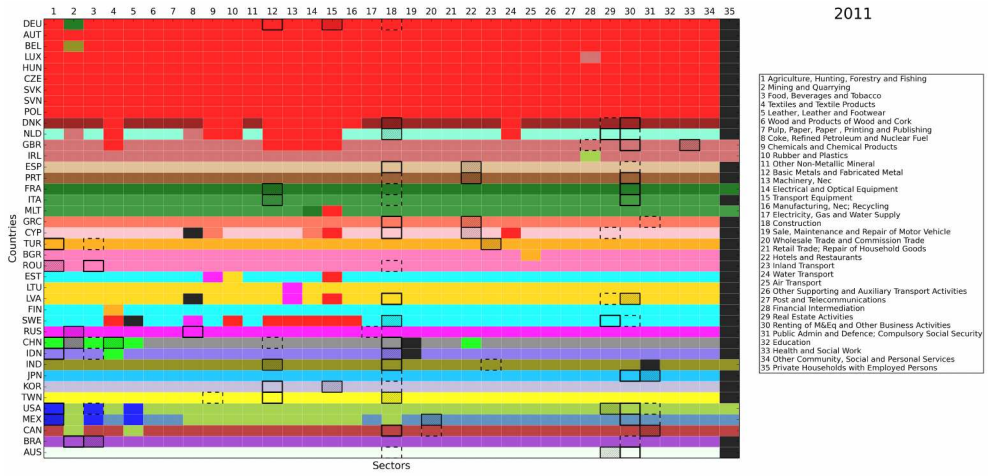



Figure 4.42: Figure shows community detection and community core detection results in 2011.

Since most of the 40 economies in the WIOD are in Europe, we cannot rule out the possibility that similar regional input-output communities are emerging in other continents. Indeed, we find also an integrated NAFTA community in North America. However, since many Asian economies are not included in the WIOD, we cannot argue if a similar trend is ongoing in the Far East.

Within each community, we also carry out the community core detection (De Leo et al. (2013)). In Figures 4.40-4.42, we identify the top 3 core economy-industry pairs for each community. The first place is with thick and solid border. The second place is with thick and dashed border. The third place is with border and texture. In general, the cores are mostly concentrated in the industries of agriculture (1), mining (2), food (3), metals (12), construction (18), and financial, business, and public services (28-31). Over time, while the services industries (28-31) have become the cores in more

and more developed economies, the primary industries (1-3) have become less central in the developed economies and have only remained as the cores in a few emerging economies, which is consistent with the Kuznets facts (Kuznets (1957),Kuznets (1973)). Furthermore, for the growing community centered on Germany, the cores are always identified in Germany (that is why we simply call it the German community) for the three selected years. It is also worth noting that, the German industry of transport equipment (15) is identified as a core in 2011 and the car industry is the most integrated in the German community, which spans over 17 economies.

### 4.5.4 The Network-Based Methods of Identifying the Key Industries

Since on a global scale the traditional assumption of stable input-output technical coefficients is violated due to the dynamics of international trade, the traditional final-demand-weighted backward linkage measure alone is insufficient to evaluate the importance of any given industry on the global economy. However, the networks approach provides us a holistic view of the global production system and we can compute various centrality measures to compare the nodes in the network. Here we focus on two network-based methods of identifying the key industries in the WION, PageRank centrality and community coreness measure. We choose PageRank over other centrality measures such as closeness and betweenness because the former systematically measures the influence of a given node and has been widely used in the previous literature to identify the key nodes (Acemoglu et al. (2012),Carvalho (2013)).

**Pagerank Centrality** Given a network, it is a problem of capital importance to bring order to its structure by ranking nodes according to their relevance. Among the many proposed, a successful and widely used centrality measure is PageRank ((Page et al., 1999), see chapter section 1.1.4).

Since the WION is weighted, we use here a weighted version of PageRank, which is computed iteratively as follows:

1. At t = 0, an initial probability distribution is assumed, usually $PR(i;0) = \frac{1}{N}$ where $N$ is the total number of nodes;

2. At each time step, the PageRank of node i is computed as:

$$PR(i,t+1) = \frac{1-d}{N} + d \sum_{j \in M(i)} \frac{PR(j,t)w_{ij}}{S(j)} \tag{4.11}$$

   where $M(i)$ are the in-neighbors of $i$, $w_{ij}$ is the weight of the link between the nodes $i$ and $j$, $S$ is the sum of the weights of the outgoing edges from $j$, and the damping factor $d$ is set to its default value, 0.85.

In Table 4.7, the second column of each year is produced by the PageRank centrality, which is denoted by PR. Unlike the final-demand-weighted backward linkage measure, where only 4 economies are among the top 20, the PageRank centrality recognizes 10 economies in the top 20 list for the three selected years.

**Community coreness measure** The other network-based method of identifying the key industries is the community coreness measure.

In the WION, once we have the $|dQ|$ for each industry, we can consider the one with the biggest $|dQ|$ the most important. We can also normalize

the $|dQ|$ to identify the most important nodes within each community. The results are shown in Figures 4.40, 4.41, and 4.42, where the first place in each community is with thick and solid border, the second place is with thick dashed border, and the third place is with both border and texture. In Table 4.7, the third column of each year is produced by the community coreness measure, which is denoted again by $|dQ|$. Interestingly, like the final-demand-weighted backward linkage measure, the community coreness measure also only includes China, Germany, Japan, and USA in the top 20 list for the selected years.

Now we have totally three methods to identify the key industries in the WION, the traditional final-demand- weighted backward linkage measure, the PageRank centrality measure, and the community coreness measure. They have different results from each other. For instance, the industry of transport equipment in Germany is captured by the PageRank but not by the other two while the industry of other business activities in USA is more important by $|dQ|$ than by the other two (see Table 4.7). Table 4.8 reports the correlation coefficient matrix among the three methods for the selected years 1995, 2003, and 2011. We find that all the three methods are positively correlated, while $\mathbf{w}$ and $|dQ|$ are correlated even more. Therefore, the network-based $|dQ|$ and especially PR can be used to complement, if not to substitute, $\mathbf{w}$ to identify the key industries in the WION.

| | **1995** | | | | **2003** | | | | **2011** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **w** | $PR$ | $\|dQ\|$ | | **w** | $PR$ | $\|dQ\|$ | | **w** | $PR$ | $\|dQ\|$ |
| **w** | 1 | 0.664224 | 0.819625 | **w** | 1 | 0.688819 | 0.724121 | **w** | 1 | 0.64281 | 0.754442 |
| PR | 0.664224 | 1 | 0.650459 | PR | 0.688819 | 1 | 0.596233 | PR | 0.64281 | 1 | 0.592057 |
| $\|dQ\|$ | 0.819625 | 0.650459 | 1 | $\|dQ\|$ | 0.724121 | 0.596233 | 1 | $\|dQ\|$ | 0.754442 | 0.592057 | 1 |

Table 4.8: Correlation coefficient matrix among the three key-industry-identification methods for selected years. The first method is the final-demand-weighted backward linkage measure, **w**. The second is the PageRank centrality, PR. The third is the community coreness measure $|dQ|$.

# Chapter 5

# Discussion

The aim of this work has been to use community detection methodology and develop methods and algorithm to be applied to the study of the global market and its functioning, in order to understand the origin of economic turmoils and critical events. In the first two works reviewed in this thesis (Cerina et al. (2012), De Leo et al. (2013) - see chapter 4 for further reference) we developed the theoretical tools that have been then applied to real world data.

Since in many cases space affects link formation by mixing in with some other attribute or masking it at all, there is the need to separate those different contributions. In section 4.1 we propose a simple model which allowed us to test community detection on spatial networks. Our model generates simple graphs that mix both geographical properties and attributes. In literature many other spatial network models have been introduced for which nodes are connected each other through a certain spatial rule. Examples range from the growth of street networks to the evolution of the territorial infrastructural

networks (see (Barthelemy, 2011) for an extensive list of this kind of models). Moreover a whole class of models that study node properties and their aggregation has recently been introduced and one of the most important of them is the stochastic block model in which a combination of various kind of node attributes are present. The novelty of our approach is to study at the same time these various aspects (geography and attributes), and, up to our knowledge, our model is the first one that considers simultaneously the two factors, space and attributes, in the context of community detection (Cerina et al. (2012)).

In particular, we explicitly show that the existence of correlations between attributes and space drastically affects the result of community detection. The results presented in this study show that community detection in spatial networks should be taken with great care, and that including space in community detection methods could lead to results difficult to interpret. We show that for weak correlations, most community detection methods work, while for stronger correlation community detection methods which remove the spatial component of the network can lead to incorrect results. It is thus important to have some information on the correlations between space and attributes in order to assess the validity of the results of community detection methods. In practical applications however, these attributes-space correlations are generally not known and this calls for the need of new approaches, such as community detection methods including in some tunable form the existence of such correlations.

Still, the main problems of all algorithms for community detection is the fact that the community definition does not provide any information about

the importance of a node inside its own community. Nodes of a community do not have all the same importance for the community stability: the removal of a node in the "core" of a network affects the partition much more than the deletion of a node that stays on the edge of the community (i.e. a node connected in the same way with nodes internal and external to its community). For this purpose, in section 4.2 we developed a novel way for detecting cores inside communities by using the properties the of modularity function (De Leo et al. (2013)).

Our application to transportation networks has been a kind of territorial benchmark for this novel approach, but the proposed method for detecting cores in communities through the optimization of the modularity function is very simple and quite general but indeed very powerful and has a variety of potential applications to other networked systems.

We applied community detection and the new "core detection" technique to three different cases with very interesting results:

**Network communities within and across borders** In section 4.3 we adopt a complex network approach to the study of the international realtionships and use core detection to define an outreach index able to measure the international openness of countries and their "internationalization" (Cerina et al. (2012)).

As already mentioned in chapter 4, the role of distance in spatially embedded complex networks has been recently investigated. Empirically speaking, it has been found that connectivity tends to decay with distance according to a power- law relationship (Onnela et al. (2011)) . This is in line with pre-

vious results in the economic literature, where an inverse power relationship has been repeatedly observed in gravity-like models of international trade, human migra- tion and foreign investment (Disdier and Head (2008)).

Despite this growing body of evidence regarding complex networks in space, still little is known about the role of national borders in the formation of cross-national networks. In this paper we aim at understanding more about this role by analyzing the structure of network communities within and across borders. We show that, while the connectivity of US scientific communities decays as a power of distance, European scientific communities tend to be confined within national borders. We introduce a new measure for the outreach of network communities across borders and confirm our results via simulations. All in all, our findings reveal that Europe is still a collection of national systems of innovation and the European Research Area is still far from becoming reality (Chessa et al. (2013)) . Our methodological approach can be used to keep track of the progress toward the integration of the European Research Area. More in general, the outreach index we discuss in this paper is worth using to detect the impact of borders on the formation and dynamic evolution of spatially embedded networks.

**The rise of China in the International Trade Network: A Community Core Detection Approach** By viewing the international trade system as an interdependent complex network and China's opening to world trade as a natural experiment, in section 4.4 we use community detection and community core detection techniques to examine both the global dynamics, i.e., communities disappear or reemerge, and the regional dynamics,

i.e., community core changes between community members, in the ITN over the period from 1995 to 2011. We find that the Asia-Oceania community has displayed rich dynamics both internally and externally. That is, the Asia-Oceania community was present during 1995-2001 and was led by Japan, and then it disappeared and was integrated with the America community during 2002-2004, and finally it reemerged during 2005-2011 and was led by China (Zhu et al. (2014)).

With a model of a dynamic-edge-weight mechanism for both inter- and intra-communities, we are able to explain the dynamics observed in the Asia-Oceania community. In a network with a fixed number of nodes and a preset initial community structure, each period a node will be selected and by chance it may increase its edge weight with an inter-community node (if the edge already exists; otherwise a new edge will be established). It will then increase its edge weight with an intra-community neighbor. Those neighbors with more inter-community strength will be preferred. Our simulation results show that the global dynamics, i.e., communities disappear or reemerge can be generated by this model setting.

In light of this simple mechanism, the interpretation of the dynamics in the Asia-Oceania community can be that, the community collapsed after China entered the WTO and built strong trade relationships with other communities, especially with the external cores, i.e., the United States and Germany, and China became regionally attractive due to the preference of external strength and restored the Asia-Oceania community as the community leader. We find some supporting evidence in the trade data. In particular, the behavior of the ratio of the inter-community trade to the intra-community

trade between the Asia-Oceania community and the America community coincides with the disappearance and reemergence of the Asia-Oceania community. Within the community, China surpassed Japan after 2003 in terms of both inter- and intra-community trade. In our simulation, the external strength can only be increased by chance. In reality, however, it can be achieved by a series of strategic moves in trade policy. This is evidenced by the surging number of RTAs that China formed since 2003. Moreover, the intra-community preference of the nodes with more inter-community strength can be understood as the incentive to have better market access through the regional big player in the global production chains.

**World Input-Output Network** In section 4.5 we investigate a MRIO system characterized by the recently available WIOD database. By viewing the world input-output system as an interdependent network where the nodes are the individual industries in different economies and the edges are the monetary goods ows between industries, we study the network properties of the so-called world input-output network (WION) and document its evolution over time. We are able to quantify not only some global network properties such as assortativity, clustering coefficient, and degree and strength distributions, but also its subgraph structure and dynamics by using community detection techniques. Over time, we trace the effects of globalization and the 2008-2009 financial crisis. We notice that national economies are increasingly interconnected in global production chains. Moreover, we detect the emergence of regional input-output community. In particular we see the formation of a large European community led by Germany. Finally, because

on a global scale the traditional assumption of stable input-output technical coefficients is violated due to the dynamics of international trade, we also use the network-based PageRank centrality and community coreness measure to identify the key industries in the WION and the results are different from the one obtained by the traditional final-demand-weighted backward linkage measure (Cerina et al. (2014)).

As mentioned elsewhere, due to the limited coverage of the WIOD, we cannot argue if the input-output integration is also observed in other continents. Therefore, in our future work, we will utilize a database having a wider coverage. Moreover, since each of the three methods of identifying the key industries captures a different aspect of the importance of any given industry, future work is also needed to compare the methods so as to identify the systematically important industries for the global economy with attention to the territorial aspects.

# Appendix A

# The Leontief-Inverse-Based Method of Identifying the Key Industries

The intuition behind the Leontief inverse is that an increase in the final demand of an industry's output will induce not only more production from the industry itself but also more from other related industries because more inputs are required. Therefore, the Leontief inverse takes into account both the direct and indirect effects of a demand increase. For instance, $\mathbf{L}_{ij}$ measures the total output produced in Industry $i$ given a one-unit increase in Industry $j$'s final demand [1]. As a result, $\mathbf{i}'\mathbf{L}$ sums up each column of $\mathbf{L}$ and each sum

[1]The Leontief inverse is demand-driven, i.e., a repercussion effect triggered by an increase in final demand. Another strand of the input-output economics literature is based on the supply-driven model, where a repercussion effect is triggered by an increase in value added (primary inputs) Ghosh (1958), Miller and Blair (2009).

measures the total output of all the industries given a one-unit increase in the corresponding industry's final demand. The vector $\mathbf{i}'\mathbf{L}$ is called the backward linkage measure [2] and can be used to rank the industries and identify the key ones in the economy Yotopoulos and Nugent (1973). However, as pointed out by Laumas (1976), the key assumption embedded in the backward linkage measure is that every industry is assigned with the same weight (or unweighted), which is far from the reality. The problem with the unweighted backward linkage measure can be demonstrated by using the hypothetical data from Table 4.6. The calculated $\mathbf{i}'\mathbf{L}$ is $\begin{bmatrix} 2.0688 & 1.8377 & 1.2223 & 1.1854 \end{bmatrix}$, which considers the industries in Economy 1 more important than the ones in Economy 2, despite the fact that Economy 2 is a lot larger than Economy 1 in terms of total outputs.

The industries of the 40 economies covered in the WIOD are very heterogeneous in terms of both total outputs and technical structure, which certainly makes the unweighted backward linkage measure invalid. In order to identify the key industries in the WIOD, we hence follow Laumas (1976) and use the final-demand-weighted backward linkage measure, which is denoted by $\mathbf{w}$ and is defined here as the Hadamard (element-wise) product of the vector of the unweighted backward linkage measure and the vector of the percentage shares of the total final demand across industries, i.e.,

$$\mathbf{w} = \mathbf{i}'\mathbf{L} \circ \frac{\mathbf{f}'}{\mathbf{i}'\mathbf{f}} \tag{A.1}$$

where $\circ$ is the element-wise multiplication operator.

[2] It is backward because the linkage is identified by tracing back to the upstream industries.

Table A.1: **Top 20 Industries Identified by the Three Methods for Selected Years.** The first method is the final-demand-weighted backward linkage measure, **w**. The second is the PageRank centrality, $PR$. The third is the community coreness measure $|\mathrm{d}Q|$.

| | **1995** | | | **2003** | | | **2011** | | |
|---|---|---|---|---|---|---|---|---|---|
| **Rank** | **w** | $PR$ | $\lvert\mathrm{d}Q\rvert$ | **w** | $PR$ | $\lvert\mathrm{d}Q\rvert$ | **w** | $PR$ | $\lvert\mathrm{d}Q\rvert$ |
| **1** | USA-Pub | USA-Pub | USA-Pub | USA-Pub | USA-Hth | USA-Obs | CHN-Cst | GBR-Hth | CHN-Cst |
| **2** | JPN-Cst | USA-Tpt | JPN-Cst | USA-Hth | DEU-Tpt | USA-Est | USA-Pub | DEU-Tpt | USA-Obs |
| **3** | USA-Cst | DEU-Tpt | USA-Obs | USA-Cst | USA-Pub | USA-Fin | USA-Hth | USA-Pub | CHN-Met |
| **4** | USA-Hth | USA-Hth | USA-Cst | USA-Est | USA-Tpt | USA-Pub | USA-Est | CHN-Elc | USA-Pub |
| **5** | USA-Est | DEU-Cst | USA-Est | USA-Rtl | GBR-Hth | USA-Hth | CHN-Elc | USA-Hth | USA-Est |
| **6** | USA-Rtl | RUS-Hth | USA-Hth | CHN-Cst | ESP-Cst | CHN-Cst | USA-Rtl | CHN-Cst | CHN-Agr |
| **7** | USA-Fod | DEU-Fod | JPN-Htl | JPN-Cst | DEU-Hth | JPN-Cst | USA-Cst | CHN-Met | CHN-Fod |
| **8** | JPN-Pub | GBR-Cst | JPN-Met | USA-Fin | GBR-Cst | USA-Ocm | USA-Fin | USA-Tpt | USA-Fin |
| **9** | USA-Tpt | USA-Cst | USA-Met | USA-Tpt | USA-Cst | CHN-Met | CHN-Fod | ESP-Cst | CHN-Min |
| **10** | JPN-Est | FRA-Tpt | JPN-Obs | USA-Fod | USA-Obs | USA-Met | CHN-Mch | AUS-Cst | USA-Cok |
| **11** | JPN-Hth | USA-Fod | DEU-Cst | USA-Htl | FRA-Tpt | JPN-Obs | JPN-Cst | ITA-Hth | CHN-Elc |
| **12** | USA-Fin | GBR-Hth | JPN-Pub | USA-Ocm | TUR-Tex | JPN-Htl | USA-Fod | DEU-Hth | USA-Hth |
| **13** | USA-Htl | USA-Obs | JPN-Hth | JPN-Pub | USA-Est | CHN-Agr | USA-Htl | USA-Obs | CHN-Omn |
| **14** | JPN-Fod | JPN-Cst | JPN-Ocm | USA-Obs | AUS-Cst | USA-Cst | CHN-Tpt | RUS-Hth | CHN-Cok |
| **15** | JPN-Rtl | DEU-Mch | JPN-Fod | JPN-Est | USA-Fod | JPN-Pub | JPN-Pub | CHN-Tpt | CHN-Mch |
| **16** | DEU-Cst | ESP-Cst | JPN-Fin | JPN-Hth | ITA-Hth | USA-Agr | USA-Tpt | DEU-Mch | USA-Cst |
| **17** | JPN-Elc | JPN-Tpt | USA-Agr | USA-Whl | DEU-Cst | USA-Tpt | USA-Ocm | FRA-Cst | CHN-Chm |
| **18** | JPN-Whl | DEU-Met | USA-Fod | JPN-Tpt | DEU-Fod | JPN-Met | CHN-Pub | CHN-Tex | JPN-Obs |
| **19** | JPN-Tpt | USA-Elc | JPN-Whl | CHN-Elc | DEU-Mch | JPN-Hth | USA-Obs | GBR-Cst | USA-Ocm |
| **20** | JPN-Mch | USA-Est | USA-Pup | DEU-Tpt | CHN-Elc | CHN-Elc | USA-Whl | DEU-Met | JPN-Cst |

Tables A.2 and A.3 provide an alternative way of viewing the key industries and economies over time. In particular, Table A.2 lists the most important economies by industry while Table A.3 lists the most important industries by economy.

Table A.2: **The Most Important Economies by Industry Over Time: Using the Final-Demand-Weighted Backward Linkage Measure.**

| Industry/Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Agr** | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN |
| **Min** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Fod** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | CHN |
| **Tex** | USA | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN |
| **Lth** | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN |
| **Wod** | JPN | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | CHN | CHN | CHN | CHN |
| **Pup** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Cok** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Chm** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Rub** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | CHN | CHN | CHN | CHN | CHN |
| **Omn** | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | USA | USA | DEU | CHN | CHN | CHN | CHN |
| **Met** | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | CHN | CHN | CHN |
| **Mch** | JPN | JPN | USA | USA | USA | USA | USA | USA | USA | USA | CHN | CHN | CHN | CHN | CHN | CHN | CHN |
| **Elc** | JPN | USA | USA | USA | USA | USA | USA | USA | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN | CHN |
| **Tpt** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | CHN | CHN | CHN |
| **Mnf** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Ele** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Cst** | JPN | JPN | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | CHN | CHN | CHN | CHN | CHN |
| **Sal** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Whl** | JPN | JPN | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Rtl** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Htl** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Ldt** | JPN | JPN | JPN | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | IND |
| **Wtt** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Ait** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Otr** | DEU | DEU | DEU | DEU | DEU | USA | USA | ITA | ITA | ITA | ITA | ITA | ITA | ITA | ITA | CHN | CHN |
| **Pst** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Fin** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Est** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Obs** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Pub** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Edu** | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | JPN | USA | CHN | CHN | CHN | CHN | CHN |
| **Hth** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Ocm** | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA | USA |
| **Pvt** | ITA | ITA | ITA | USA | USA | USA | USA | USA | ITA | ITA | ITA | ITA | ITA | ITA | ITA | ITA | ITA |

Table A.3: **The Most Important Industries by Economy Over Time: Using the Final-Demand-Weighted Backward Linkage Measure.**

| Economy/Year | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AUS** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **AUT** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **BEL** | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **BGR** | Fod | Fod | Fod | Fod | Agr | Agr | Fod | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **BRA** | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub |
| **CAN** | Pub | Pub | Cst | Pub | Tpt | Tpt | Pub | Pub | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **CHN** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **CYP** | Htl | Cst | Htl | Htl | Htl | Htl | Htl | Pub | Pub | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **CZE** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Tpt | Tpt |
| **DEU** | Cst | Cst | Cst | Cst | Cst | Cst | Tpt | Tpt | Tpt | Tpt | Tpt | Tpt | Tpt | Tpt | Tpt | Tpt | Tpt |
| **DNK** | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Hth | Hth | Hth | Cst | Cst | Cst | Hth | Hth | Hth | Hth |
| **ESP** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **EST** | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **FIN** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **FRA** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **GBR** | Pub | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Hth | Hth | Hth | Hth | Hth | Hth | Hth |
| **GRC** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Pub | Pub | Pub |
| **HUN** | Fod | Fod | Fod | Fod | Elc | Elc | Elc | Elc | Elc | Elc | Elc | Elc | Elc | Elc | Elc | Elc | Elc |
| **IDN** | Cst | Cst | Cst | Cst | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **IND** | Agr | Agr | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **IRL** | Fod | Fod | Fod | Elc | Elc | Elc | Elc | Elc | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Chm | Chm |
| **ITA** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **JPN** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **KOR** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **LTU** | Fod | Fod | Pub | Cst | Cst | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Fod | Fod | Fod |
| **LUX** | Cst | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin | Fin |
| **LVA** | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **MEX** | Fod | Fod | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **MLT** | Elc | Elc | Elc | Elc | Elc | Elc | Elc | Cst | Cst | Elc | Cst | Ocm | Ocm | Ocm | Ocm | Ocm | Ocm |
| **NLD** | Fod | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Pub | Pub | Pub | Pub | Cst | Cst | Pub | Pub | Pub |
| **POL** | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Fod | Fod | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Cst |
| **PRT** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **ROM** | Fod | Fod | Fod | Fod | Fod | Fod | Fod | Fod | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **RUS** | Fod | Cst | Cst | Cst | Fod | Fod | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **SVK** | Cst | Cst | Cst | Cst | Cst | Cst | Tpt | Tpt | Tpt | Tpt | Cst | Tpt | Tpt | Tpt | Cst | Cst | Cst |
| **SVN** | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **SWE** | Est | Est | Est | Est | Est | Hth | Hth | Hth | Hth | Hth | Hth | Hth | Hth | Hth | Hth | Hth | Hth |
| **TUR** | Cst | Cst | Cst | Cst | Fod | Fod | Fod | Tex | Tex | Tex | Fod | Fod | Tex | Fod | Fod | Fod | Fod |
| **TWN** | Cst | Cst | Pub | Pub | Elc | Elc | Elc | Elc | Elc | Cst | Cst | Cst | Cst | Cst | Cst | Cst | Cst |
| **USA** | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub | Pub |

# Bibliography

Traffic detector handbook. Fhwa operations material. URL http://www.fhwa.dot.gov/.

Censimento generale della popolazione e delle abitazioni - matrice origine destinazione degli spostamenti pendolari della sardegna. Technical report, Italian National Institute of Statistics (ISTAT), 1991. URL http://www.istat.it/it/archivio/3758.

The travel forecasting model set for the atlanta region - 2008 documentation. Technical report, Atlanta Regional Commission, 2008.

Daron Acemoglu, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The Network Origins of Aggregate Fluctuations. *Econometrica*, 80 (5):1977–2016, 09 2012. URL http://ideas.repec.org/a/ecm/emetrp/v80y2012i5p1977-2016.html.

Yong-Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multi-scale complexity in networks. *Nature*, 466:761+, October 2010. URL http://arxiv.org/abs/0903.3178.

R Albert and A.-L. Barabási. Statistical mechanics of com-

plex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002. ISSN 0034-6861. doi: 10.1103/RevModPhys.74.47. URL `http://www.springer.com/physics/theoretical,+mathematical+&+computational+physics/book/978-3-540-40372-2http://link.aps.org/doi/10.1103/RevModPhys.74.47`.

Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.

Martin Andersson and Urban Gråsjö. Spatial dependence and the representation of space in empirical models. *The annals of regional science*, 43(1): 159–180, 2009.

Iván Arribas, Francisco Pérez, and Emili Tortosa-Ausina. Measuring globalization of international trade: theory and evidence. *World Development*, 37(1):127–145, 2009.

S. Arunachalam and M.J. Doss. Mapping international collaboration in science in asia through coauthorship analysis. *Current Science*, 79(5):621, 2000.

Prema-chandra Athukorala. Product fragmentation and trade patterns in east asia*. *Asian Economic Papers*, 4(3):1–27, 2005.

Prema-chandra Athukorala and Nobuaki Yamashita. Production fragmentation and trade integration: East asia in a global context. *The North American Journal of Economics and Finance*, 17(3):233–256, 2006.

Richard Baldwin and Daria Taglioni. Gravity for dummies and dummies

for gravity equations. Technical report, National Bureau of Economic Research, 2006.

Richard E Baldwin. *The Spoke Trap: hub and spoke bilateralism in East Asia*. Oxford University Press, 2008.

A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

A. Barrat, M. Barthèlemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proc Natl Acad Sci USA*, 101 (11):3747–3752, 2004.

Marc Barthelemy. Spatial Networks. *Physics Reports*, 499:1–101, 2011. URL `http://arxiv.org/abs/1010.0302`.

M. Batty. *Urban Modelling: Algorithms Calibrations, Predictions*. Cambridge Earth Science Series. Cambridge University Press, 1976. ISBN 9780521208116. URL `http://books.google.it/books?id=-uRRAQAAIAAJ`.

Y. Berezin, A. Gozolchiani, O. Guez, and S. Havlin. Stability of climate networks with time. *Scientific Reports*, 2, 2012.

Jeffrey H Bergstrand. The gravity equation in international trade: some microeconomic foundations and empirical evidence. *The review of economics and statistics*, pages 474–481, 1985.

Florian Blöchl, Fabian J Theis, Fernando Vega-Redondo, and Eric O'N

Fisher. Vertex centralities in input-output networks reveal the structure of modern economies. *Physical Review E*, 83(4):046127, 2011.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008+, July 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/P10008. URL `http://dx.doi.org/10.1088/1742-5468/2008/10/P10008`.

D. Brockmann and D. Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342:1337–1342, 2013.

Guido Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology (Oxford Finance)*. Oxford University Press, USA, June 2007. ISBN 0199211515. URL `http://www.worldcat.org/isbn/0199211515`.

Guido Caldarelli, Alessandro Chessa, Fabio Pammolli, Andrea Gabrielli, and Michelangelo Puliga. Reconstructing a credit network. *Nature Physics*, 9: 125–126, 2013. ISSN 1745-2473. doi: doi:10.1038/nphys2575. URL `http://www.nature.com/nphys/journal/v9/n3/full/nphys2580.html`.

R. J. G. B. Campello. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, 28(7):833–841, May 2007. doi: 10.1016/j.patrec.2006.11.010. URL `http://dx.doi.org/10.1016/j.patrec.2006.11.010`.

Céline Carrere. Revisiting the effects of regional trade agreements on trade flows with proper specification of the gravity model. *European Economic Review*, 50(2):223–247, 2006.

Vasco M Carvalho. A survey paper on recent developments of input-output analysis. Technical report, Complexity Research Initiative for Systemic InstabilitieS, 2013. URL http://www.crisis-economics.eu/publication/deliverable-d3-1-a-survey-paper-on-recent-developments-of-input-output-analysis

Federica Cerina, Zhen Zhu, Alessandro Chessa, and Massimo Riccaboni. World input-output network.

Federica Cerina, Vincenzo De Leo, Marc Barthelemy, and Alessandro Chessa. Spatial correlations in attribute communities. *PLoS ONE*, 7(5):e37507, 05 2012. doi: 10.1371/journal.pone.0037507. URL http://dx.doi.org/10.1371%2Fjournal.pone.0037507.

Federica Cerina, Alessandro Chessa, Fabio Pammolli, and Massimo Riccaboni. Network communities within and across borders. *Sci. Rep.*, 4, 2014. doi: 10.1038/srep04546. URL http://dx.doi.org/10.1038/srep04546.

David Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, USA, 1 edition, September 1987. ISBN 0195042778.

A Chessa, A Morescalchi, F Pammolli, O Penner, AM Petersen, and M Riccaboni. Is europe evolving toward an integrated research area? *Science*, 339(6120):650–651, 2013.

Martha G. Alatriste Contreras and Giorgio Fagiolo. Propagation of Economic Shocks in Input-Output Networks: A Cross-Country Analysis. Papers 1401.4704, arXiv.org, January 2014. URL http://ideas.repec.org/p/arx/papers/1401.4704.html.

Riccardo Crescenzi, Andrés Rodríguez-Pose, and Michael Storper. The territorial dynamics of innovation: a europe–united states comparative analysis. *Journal of Economic Geography*, 7(6):673–709, 2007.

Leon Danon, Jordi Duch, Alex Arenas, and Albert Díaz-Guilera. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 9008:09008, 2005.

L. Daqing, K. Kosmidis, A. Bunde, and S. Havlin. Dimension of spatially embedded networks. *Nature Physics*, 7(6):481–484, 2011.

Luca De Benedictis and Lucia Tajoli. The world trade network. *The World Economy*, 34(8):1417–1454, 2011.

V. De Leo, G. Santoboni, F. Cerina, M. Mureddu, L. Secchi, and A. Chessa. Community core detection in transportation networks. *Physical Review E, 87 (2). ISSN 1539-3755; eprint arXiv:1304.0141*, March 2013.

Andrea De Montis, Simone Caschili, and Alessandro Chessa. Commuter networks and community detection: a method for planning sub regional areas. *arXiv preprint arXiv:1103.2467*, 2011.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Phase transition in the detection of modules in sparse networks. *Physical Review Letters*, 107(6):4, 2011. URL `http://arxiv.org/abs/1102.1182`.

L Denoeud, H Garreta, and A Gu. Comparison of distance indices between partitions. *Proceedings of IFCS2006 Data Science and Classification V Batagelj et al Eds Springer*, pages 21–28, 2006. URL `http://dx.doi.org/10.1007/3-540-34416-0_3`.

Anne-Célia Disdier and Keith Head. The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, 90(1): 37–48, 2008.

Georg Duernecker, Moritz Meyer, and Fernando Vega-Redondo. Being close to grow faster: A network-based empirical analysis of economic globalization. In *EUI Working Papers, ECO 2012/05*. Department of Economics, European University Insitute, 2012.

J. Eaton and S. Kortum. Technology, geography, and trade. *Econometrica*, 70(5):1741–1779, 2004.

Thorsten Emmerich, Armin Bunde, and Shlomo Havlin. Structural and functional properties of spatially embedded scale-free networks. *Phys. Rev. E*, 89:062806, Jun 2014. doi: 10.1103/PhysRevE.89.062806. URL `http://link.aps.org/doi/10.1103/PhysRevE.89.062806`.

Sven Erlander and Neil F. Stewart. *The gravity model in transportation analysis : theory and extensions / Sven Erlander and Neil F. Stewart.* VSP, Utrecht, 1990. ISBN 9067640891.

T. Evans. Dynamical Processes on Complex Networks, by A. Barrat, M. Barthelemy and A. Vespignani. *Contemporary Physics*, 51:187–188, March 2010. doi: 10.1080/00107510903084036.

Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences of the United States of America*,

108(19):7663–7668, 2011. URL http://www.ncbi.nlm.nih.gov/pubmed/21518910.

Giorgio Fagiolo, Javier Reyes, and Stefano Schiavo. World-trade web: Topological properties, dynamics, and evolution. *Physical Review E*, 79(3): 036115, 2009.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486 (3-5):75 – 174, 2010. ISSN 0370-1573. doi: DOI:10.1016/j.physrep.2009.11.002. URL http://www.sciencedirect.com/science/article/B6TVP-4XPYXF1-1/2/99061fac6435db4343b2374d26e64ac1.

Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *physics0607100*, 104(1):8, 2007. URL http://arxiv.org/abs/physics/0607100.

Diego Garlaschelli and Maria I Loffredo. Structure and evolution of the world trade network. *Physica A: Statistical Mechanics and its Applications*, 355 (1):138–144, 2005.

Guillaume Gaulier and Soledad Zignago. Baci: International trade database at the product-level the 1994-2007 version. 2010.

Alak Ghosh. Input-output approach in an allocation system. *Economica*, pages 58–64, 1958.

Sumantra Ghoshal and Christopher A Bartlett. The multinational corporation as an interorganizational network. *Academy of management review*, 15(4):603–626, 1990.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–7826, 2002. doi: 10.1073/pnas.122653799. URL `http://www.pnas.org/content/99/12/7821.abstract`.

Jacob Goldenberg and Moshe Levy. Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*, 2009.

Steve Gregory. Ordered community structure in networks. *Physica A*, 391 (8), 2011. URL `http://arxiv.org/abs/1104.0923`.

Gene M Grossman and Elhanan Helpman. Trade wars and trade talks. In *Trade and Tax Policy, Inflation and Exchange Rates*, pages 171–214. Springer Berlin Heidelberg, 1997.

R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102(22):7794–7799, May 2005. ISSN 0027-8424. doi: 10.1073/pnas.0407994102. URL `http://dx.doi.org/10.1073/pnas.0407994102`.

Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70, 2004.

Alexander Halavais. National borders on the world wide web. *New Media & Society*, 2(1):7–28, 2000.

Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

Jarno Hoekman, Thomas Scherngell, Koen Frenken, and Robert Tijssen. Acquisition of european research funds and its effect on international scientific collaboration. *Journal of Economic Geography*, 13(1):23–52, 2013. doi: 10.1093/jeg/lbs011.

Dandan Hu, Peter Ronhovde, and Zohar Nussinov. Phase transitions in random potts systems and the community detection problem: spin-glass type and dynamic perspectives. *Philosophical Magazine*, 0(0):1–40, 2011a.

Dandan Hu, Peter Ronhovde, and Zohar Nussinov. A replica inference approach to unsupervised and multi-scale image segmentation. *eprint arXiv:1106.5793*, 2011b. URL `http://arxiv.org/abs/1106.5793`.

Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. ISBN 0-13-022278-X. URL `http://portal.acm.org/citation.cfm?id=42779`.

Raja Kali and Javier Reyes. The architecture of globalization: a network approach to international economic integration. *Journal of International Business Studies*, 38(4):595–620, 2007.

Pablo Kaluza, Andrea Kolzsch, Michael T Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of the Royal Society Interface the Royal Society*, 7(48):1093–1103, 2010. URL `http://arxiv.org/abs/1001.2172`.

Brian Karrer, Elizaveta Levina, and M E J Newman. Robustness of community structure in networks. *Physical Review E*, 77(4):10, 2007. URL `http://arxiv.org/abs/0709.2108`.

Sangmoon Kim and Eui-Hang Shin. A longitudinal analysis of globalization and regionalization in international trade: A social network approach. *Social Forces*, 81(2):445–468, 2002.

Maksim Kitsak, Massimo Riccaboni, Shlomo Havlin, Fabio Pammolli, and H Eugene Stanley. Scale-free models for the structure of business firm networks. *Physical Review E*, 81(3):036117, 2010.

Simon Kuznets. Quantitative aspects of the economic growth of nations: Ii. industrial distribution of national product and labor force. *Economic Development and Cultural Change*, 5(4):pp. 1–111, 1957. ISSN 00130079. URL `http://www.jstor.org/stable/1151943`.

Simon Kuznets. Modern Economic Growth: Findings and Reflections. *American Economic Review*, 63(3):247–58, June 1973. URL `http://ideas.repec.org/a/aea/aecrev/v63y1973i3p247-58.html`.

R. Lambiotte, V.D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical Review E - Statistical, Non-*

*linear and Soft Matter Physics*, 80(5 Pt 2):056117, 2009. URL `http://arxiv.org/abs/0908.1062`.

Prem S Laumas. The weighting problem in testing the linkage hypothesis. *The Quarterly Journal of Economics*, pages 308–312, 1976.

Anna Lewis, Nick Jones, Mason Porter, and Charlotte Deane. The function of communities in protein interaction networks at multiple scales. *BMC Systems Biology*, 4(1):100, 2010. ISSN 1752-0509. doi: 10.1186/1752-0509-4-100. URL `http://www.biomedcentral.com/1752-0509/4/100`.

Chunguang Li and Philip K Maini. An evolving network model with community structure. *Journal of Physics A: Mathematical and General*, 38(45): 9741, 2005.

Ettore Majorana and R.N. Mantegna. The value of statistical laws in physics and social sciences. In GiuseppeFranco Bassani, editor, *Ettore Majorana Scientific Papers*, pages 237–260. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-48091-4. doi: 10.1007/978-3-540-48095-2_11. URL `http://dx.doi.org/10.1007/978-3-540-48095-2_11`.

Rosario Nunzio Mantegna, Harry Eugene Stanley, et al. *An introduction to econophysics: correlations and complexity in finance*, volume 9. Cambridge university press Cambridge, 2000.

S. Maraut, H. Dernis, C. Webb, V. Spiezia, and D. Guellec. The oecd regpat database: a presentation. Technical report, OECD Publishing, 2008.

James McNerney, Brian D Fath, and Gerald Silverberg. Network structure of inter-industry flows. *Physica A: Statistical Mechanics and its Applications*, 392(24):6427–6441, 2013.

Marina Meila. Comparing clusterings, 2002.

Ronald E Miller and Peter D Blair. *Input-output analysis: foundations and extensions*. Cambridge University Press, 2009.

E. J. Mishan. *Elements of cost-benefit analysis, by E. J. Mishan*. Allen and Unwin London, 1972. ISBN 0043000355 0043000363 0043000665.

Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995. ISSN 1098-2418. doi: 10.1002/rsa.3240060204. URL `http://dx.doi.org/10.1002/rsa.3240060204`.

M E J Newman and M Girvan. Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, 69(2 Pt 2):16, 2004a. URL `http://arxiv.org/abs/cond-mat/0308217`.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, Feb 2004b. doi: 10.1103/PhysRevE.69.026113.

J.P. Onnela, S. Arbesman, M.C. González, A.L. Barabási, and N.A. Christakis. Geographic constraints on social network groups. *PLoS one*, 6(4): e16939, 2011.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web, 1999.

Fabio Pammolli and Massimo Riccaboni. Technological regimes and the growth of networks: An empirical analysis. *Small Business Economics*, 19 (3):205–215, 2002.

B.P. Pashigian. *Price Theory and Applications*. McGraw-Hill, New York, 1995.

Carlo Piccardi and Lucia Tajoli. Existence and significance of communities in the world trade web. *Physical Review E*, 85(6):066119, 2012.

M. A. Porter, J. P. Onnela, and P. J. Mucha. Communities in networks. *ArXiv*, 902(6), 2009.

W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

J.E. Rauch. Business and social networks in international trade. *Journal of Economic Literature*, pages 1177–1203, 2001.

Javier Reyes, Rossitza Wooster, and Stuart Shirrell. Regional trade agreements and the pattern of trade: A networks approach. *The World Economy*, 2014.

Massimo Riccaboni and Stefano Schiavo. Structure and growth of weighted networks. *New Journal of Physics*, 12(2):023003, 2010.

Massimo Riccaboni, Alessandro Rossi, and Stefano Schiavo. Global networks of trade and bits. *Journal of Economic Interaction and Coordination*, 8 (1):33–56, 2013.

Moliterni R Riccaboni M. Managing technological transitions through r&d alliances. *R&D Management*, 39:124–135, 2009. doi: 10.1111/j.1467-9310.2009.00545.x.

Schiavo S Riccaboni M. Stochastic trade networks. *Journal of Complex Networks*, 2014.

Thomas Scherngell and Rafael Lata. Towards an integrated european research area? findings from eigenvector spatially filtered spatial interaction models using european framework programme data*. *Papers in Regional Science*, 2012. ISSN 1435-5957. doi: 10.1111/j.1435-5957.2012.00419.x.

Christian M Schneider, André A Moreira, José S Andrade, Shlomo Havlin, and Hans J Herrmann. Mitigation of malicious attacks on networks. *Proceedings of the National Academy of Sciences*, 108(10):3838–3841, 2011.

Ma Angeles Serrano and Marián Boguná. Topology of the world trade web. *Physical Review E*, 68(1):015101, 2003.

Jungyul Sohn. Evaluating the significance of highway network links under the flood damage: An accessibility approach. *Transportation Research Part A: Policy and Practice*, 40(6):491–506, 2006. URL `http://EconPapers.repec.org/RePEc:eee:transa:v:40:y:2006:i:6:p:491-506`.

Ricard V Solé, Martí Rosas-Casals, Bernat Corominas-Murtra, and Sergi

Valverde. Robustness of the european power grids under intentional attack. *Physical Review E*, 77(2):026102, 2008.

Christian Thiemann, Fabian Theis, Daniel Grady, Rafael Brune, and Dirk Brockmann. The structure of borders in a small world. *PloS one*, 5(11): e15422, 2010.

Marcel Timmer, AA Erumban, J Francois, A Genty, R Gouma, B Los, F Neuwahl, O Pindyuk, J Poeschl, JM Rueda-Cantuche, et al. The world input-output database (wiod): Contents, sources and methods. *WIOD Background document available at www. wiod. org*, 2012.

AA Tsonis and PJ Roebber. The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications*, 333:497–504, 2004.

Arnold Tukker and Erik Dietzenbacher. Global multiregional input–output frameworks: an introduction and outlook. *Economic Systems Research*, 25 (1):1–19, 2013.

Irena Tzekina, Karan Danthi, and Daniel N Rockmore. Evolution of community structure in the world trade web. *The European Physical Journal B*, 63(4):541–545, 2008.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world' networks. *nature*, 393(6684):440–442, 1998.

C Webb, H Dernis, D Harhoff, and K Hoisl. Analysing european and international patent citations: A set of epo patent database building blocks, ocde science. *Technology and industry working paper*, 9, 2005.

Thomas Wiedmann, Harry C Wilting, Manfred Lenzen, Stephan Lutter, and Viveka Palm. Quo vadis mrio? methodological, data and institutional requirements for multi-region input–output analysis. *Ecological Economics*, 70(11):1937–1945, 2011.

Alan Geoffrey Wilson. Land-use/transport interaction models: Past and future. *Journal of Transport Economics and Policy*, pages 3–26, 1998.

Pan A Yotopoulos and Jeffrey B Nugent. A balanced-growth version of the linkage hypothesis: a test. *The Quarterly Journal of Economics*, 87(2): 157–171, 1973.

Zhen Zhu, Federica Cerina, Alessandro Chessa, Guido Caldarelli, and Massimo Riccaboni. The rise of china in the international trade network: A community core detection approach. *PLoS ONE*, 9(8):e105496, 08 2014. doi: 10.1371/journal.pone.0105496. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0105496`.

# Acknowledgements

Federica Cerina gratefully acknowledges Sardinia Regional Government for the financial support of her PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources, Objective l.3, Line of Activity l.3.1.).