Community Detection Algorithms and Clustering of Input-Output Linkages — Analysis of their interconnections*

Nadia Garbellini†

May 20, 2012

Abstract The problem of finding groups of industries particularly connected to each other and relatively disconnected from the rest of the inter-industry network is not new in input-output literature — starting from Leontief (1986[1963]), who translated it into the problem of block-partitioning the inter-industry transactions table in order to isolate blocks of non-zero elements.

Since an inter-industry transactions table can be seen as a weighted, directed graph representing a network, network theory can be very useful in providing efficient techniques for the identification of groups of industries — or, using network theory terminology, for the detection of communities. A wide range of community detection algorithms have been developed which allow to isolate clusters. However, typically different algorithms lead to different partitions.

The great majority of these algorithms have been developed with no reference to specific real networks. Moreover, different algorithms rely on the translation into mathematical terms of different operational definitions of what a cluster is. Some of these definitions might be appropriate for certain kinds of networks but not for others. The definition of a cluster might be very different in the context of a biological network than within a social network.

In order to define an appropriate method for partitioning I-O networks, contributions coming from traditional I-O and network theory need to be synthesised. The present paper aims first at giving a precise definition of

*Paper prepared for the 20th IIOA Conference, Bratislava, June 2012. Draft Version. Please do not cite without Author's permission.

†Università degli Studi di Pavia, nadia.garbellini@unipv.it.

a cluster of industries. In the second place, the possible economic interpretation of the community detection algorithms most commonly used in I-O applications will be scrutinised, in order to assess their appropriateness for I-O networks. Finally, they will be related to the ideas at the basis of earlier attempts coming from traditional I-O clustering literature. All these issues will be illustrated with an empirical application.

1 Introduction

The problem of identifying groups of industries conforming strongly connected economic 'sub-systems' is at least as old as Leontief's (1986[1963]) pioneering contribution. Many attempts have been made at answering this question, though without settling a common ground as a point of departure. In fact, no shared definition of what these 'sub-systems' actually are has been given; as a consequence, no unitary, and sometimes not even comparable, answer has been provided.

Analytically, the problem emerged as a way of making the intermediate transaction matrix block diagonal, or triangular, in order to make its inversion easier. As a by-product, a relation between the form took by such modified matrix and the structure of the economic system was singled out:

> It was the labor of computation that prompted the first systematic studies of the structural characteristics of an economy as they are displayed in an input-output table. During the late 1940s [...] the U.S. Air Force undertook to rearrange the rows and columns in a table of the U.S. economy in such a way as to minimize the computation required to yield numerical solutions. Such rearrangement brought into sharper relief the interindustry and intersectoral transactions that tie industries and sectors together in the *subunits of the total structure of the ecomomy.*
>
> (Leontief 1986[1963], p. 166), our italics

It is quite clear that Leontief seeks *subunits of the total structure of the ecomomy* as emerging from their absolute direct intermediate transactions; a problem distinct from that of finding subsystems in the sense of Sraffa (1960), as it was sometimes stated. In fact, some approaches define the kind of linkages for individuating industry clusters as direct and indirect ones, as emerging from the power series of the input coeficient matrix, from industries closely connected to final demand backwards to industries basically playing

the role of intermediate inputs providers.[1]

An important difference with respect to Leontief's analysis is the level of disaggregation of the tables he used to consider. Tables with that same disaggregation are nowadays provided by a very small number of countries.[2] The immediate consequence is that it is quite unlikely to have zero entries; to use Leontief's words, such economies will result completely interdependent, and no triangularisation or block-diagonalisation of the intermediate deliveries table can e performed. In such a context, it is necessary to find a different procedure for the identification of industry clusters, or subunits of the complete economic system.

A common solution in the literature has been that of defining some criterion on the basis of which to identify *significant* relations, and then of transforming the corresponding table into a binary, or boolean, matrix where ones indicate the existence of a significant flow, while zeros its absence. In this way, the transaction table is forced to display zero entries, and some rearrangement of rows and columns can be performed in order to isolate groups. Of course, such a procedure suffers from arbitrariness as regards the chosen criterion.

Another possible solution is that of performing a hierarchical clustering of the activities, consisting in building a tree whose 'roots' are the system as a whole and whose tips are the individual industries. Branches consecutively connect the most closely connected tips down to the roots. This procedure, though considering the magnitude of all intermediate flows, and not the mere existence of significant ones, still include some degree of arbitrariness; in fact, clusters can be individuated starting from sections of the tree that can be cut at different heights. Different choices of course induce different partitions of the IO network. Moreover, there are many ways of performing hierarchical clustering, the difference basically consisting in the way in which connections between agglomerated activities are updated. The choice of the method to follow is of course not trivial and generally leads to different results.

In what follows, different approaches to significant linkages or industry clusters detection will be considered, both from a theoretical and from an

[1] Though being distinct, the two problems are closely related, and interesting to be analysed in their connection. They can be referred to as the singling out of the horizontal and vertical structure of an economic system, respectively. Both perspectives, and their interactions, are fundamental for the purpose of the proper definition of industrial, labour and fiscal policies.

[2] For example: UK (123 activitiews), India (), Japan (190 activities) and US ().

empirical point of view, by applying each of them to the case of Italy for the year 2008. In particular, section 2 introduces Slater's (1977) hierarchical clustering methodology. Section 3 deals with an approach which had, and still have, great fortune: Qualitative IO Analysis (QIOA); in particular, we will focus on Aroche-Reyes Important Coefficients (ICs) analysis (section 3.1), and Schnabl's Elasticity Coefficient (EC, section 3.2) and Minimal Flow Analysis (MFA, section 3.3). Section 4 analyses the approach proposed by Oosterhaven, Eding & Stelder's (2001) for the detection of regional clusters in the Netherlands. Finally, section 5 describes Hoen's (2002) attempt at block-diagonalising a boolean matrix built based upon a set of restriction on the intermediate direct and indirect flows. This last method will also be applied starting from binary adjacency matrix as obtained with all the other procedures.

Sections 6 and 7 introduce concepts coming from recent developments of network theory, analysing their connection with traditional IO approaches and singling out their economic interpretations, along with their appropriateness or not for the study of IO networks. Section 8 provides an alternative interpretation of the second method and a reformulation of the first one including the developed hints.

In the last section, some very brief concluding remarks are followed by an essential sketch of possible further lines of research. compares the outcomes of different approaches and singles out their advantages and drawbacks. On the basis of this, conclusions are drawn on the methodology which is considered more appropriate and dealing to the most satisfactorily results. As a way of conclusion, some remarks on possible further lines of research taking advantage of network theory will be outlined.

2 Hierarchical clustering based on Functionally Integrated Industries

Slater's (1977) paper proposes an agglomerative hierarchical clustering criterion based on what he calls *Functionally Integrated Industries* (FII). Translated into graph-theoretical terms, Slater's (1977) procedure consists in obtaining a symmetric flows matrix $\mathbf{F} = [f_{ij}]$ s.t.:

$$f_{ij} = \max\{x_{ij}, x_{ji}\}, \quad i, j = 1, 2, \ldots, n$$

and then computing, on the basis of it, a distance matrix to be used to

get a spanning tree where hierarchical clustering is performed on the basis of the single linkage, or minimum distance, criterion. In a few words, this criterion consists in picking the two nodes i and j corresponding to $\max\{\mathbf{F}\}$ and merging them; then, the second greatest element of the flow matrix is considered, and the two corresponding nodes merged. If one of the nodes already belongs to a merge, the whole group is merged. The procedure continues up to the point where all the original nodes are connected. The resuting spanning tree, or dendrogram, can in pronciple be cut at different heights, each corresponging to a successive merge — alternatively, the cut can correspond to a pre-defined number of clusters; the sections obtained in this way from the roots of the tree up to the tips would identify an increasing number of clusters: from one big cluster including all nodes in corresponence of the root, to n clusters, one for each distinct node, in correspondence of the tips.

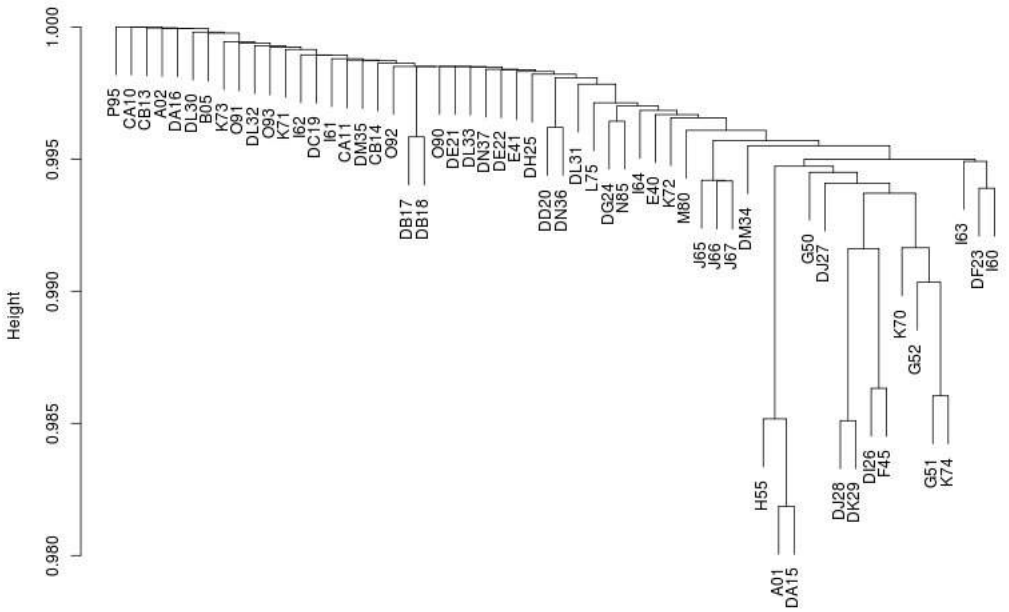
The dendrogram obtained by applying this procedure to the Italian case in year 2008 is shown in Figure 1a.

As stated in the Introduction, choosing a different criterion for hierarchical clustering would lead to very different results. As a way of example, consider Figure 1b, which displays the results of clustering based on the complete, rather than single, linkages method.

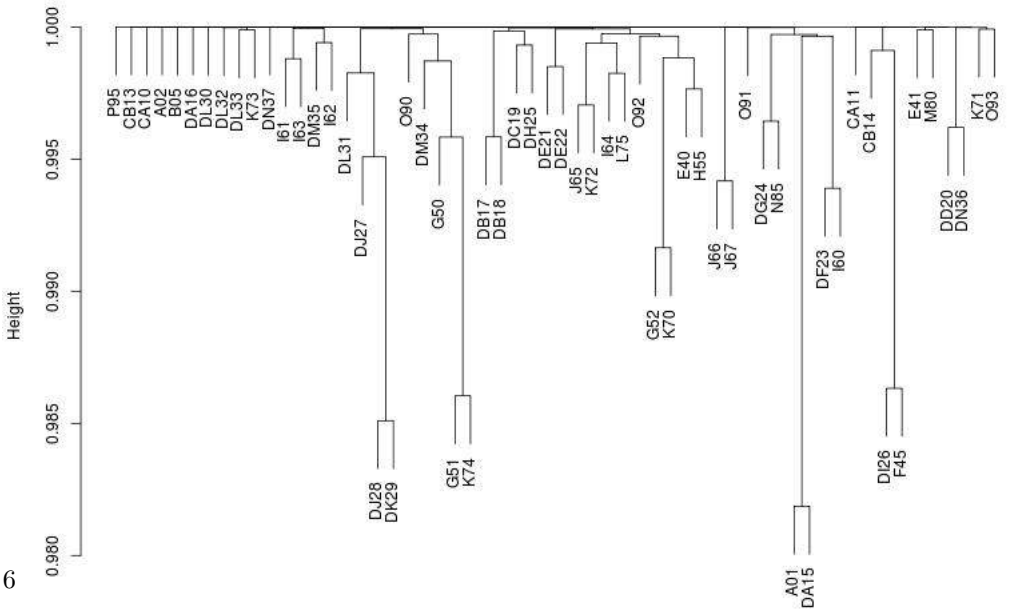
The difference between two methods consists of the way in which the flow matrix is updated after each successive merge. The first two industries to be merged in both cases are i=A01 and j=DA15. With the single linkages method, the flows, to be considered for further merges, between this newly formed group and any other industry $k \neq i, j$ is given by $\max\{f_{ik}, f_{jk}\}$; on the contrary, with the complete linkages method it is given by $\min\{f_{ik}, f_{jk}\}$. The two procedures are summarised in Table 1.

While the first procedure, adopted by Slater (1977), does not lead, in the Italian (2008) case, to the identification of well-defined industry clusters — neither at any height nor correspondingly to any pre-defined number of clusters — the complete linkages procedure does, at least to some extent. Tables 2 and 3 show the results of cutting the corresponding trees in order to isolate 15 industry clusters.

The single linkages method leads to the identification of one big cluster, including 44 industries; the remaining ones stand isolated, each conforming a separated cluster. Increasing/decreasing the pre-defined number of cluster to be identified would lead to a decrease/increase of the number of industries entering the 'mega cluster', with the others still being separated.



(a) Single linkages method



(b) Complete linkages method

Single linkages method	Complete linkages method
(1) Pick $f_{ij} = \max\{\mathbf{F}\}$	(1) Pick $f_{ij} = \max\{\mathbf{F}\}$
(2) Group 1: $G_1 = \{i, j\}$	(2) Group 1: $G_1 = \{i, j\}$
(3) Merge rows/columns (i, j) of matrix $\mathbf{F}$: $f_{G_1 k} = \max\{f_{ik}, f_{jk}\}$	(3) Merge rows/columns (i, j) of matrix $\mathbf{F}$: $f_{G_1 k} = \min\{f_{ik}, f_{jk}\}$
(4) Go through steps (1)-(3) until the last group found includes all nodes in the network	(4) Go through steps (1)-(3) until the last group found includes all nodes in the network

Table 1: Single and complete linkages methos for hierarchical clustering

On the contrary, the complete linkages method leads to identifying some meaningful clusters. Cluster 7 in Table 3 includes Stone-sand-clay-minerals, Glass-clay-cement-ceramic, and Construction; Cluster 10 Wood, Furniture-Sports-Toys, Renting-equipment and Personal-services; Cluster 11 Office-machinery-computer, ICT-equipment, Medical-precision-equip. and R-D; finally, Cluster 12 includes Ships-railway-aircrafts, Transport-water, Transport-air and Storage-travel-agencies.

3 Qualitative Input-Output Analysis (QIOA)

QIOA is based on obtaining, by adopting some — arbitrary — definition of 'significant' edges, a boolean adjacency matrix from either $\mathbf{A}$ or $\mathbf{X}$, and then applying a multi-layer approach looking for the shortest-path 'significant' connections between industries.

More specifically, after defining significant edges and thus some criterion for their identification — usually fixing the minimum value of a significant edge and denoting it by F — a boolean adjacency matrix $W^0 = [w^0_{ij}]$ is built s.t. $w^0_{ij} = 1$ iff $a_{ij} > F$ and $w^0_{ij} = 0$ otherwise. Subsequent layers are then obtained as:

$$\mathbf{W}^k = \mathbf{W} \dot{\times} \mathbf{W}^{k-1}, \quad k = 1, 2, \ldots, n-2$$

where $\dot{\times}$ denotes boolean multiplication and $n-2$ is the last layer — since $n-1$ is the maximum possible length of a shortest-path between two industries.

The $\mathbf{W}^k$s are then condensed, through boolean summation, in a single matrix $\mathbf{W} = [w_{ij}]$, s.t. $w_{ij} = 1$ if at least a path of length $\lambda \in [1, n-1]$ exists between industries i and j. As a last step, the so-called *connectivity matrix*

Single linkages method		
cluster 1	Construction	Coal Mining
Agriculture	Sale-repair-vehicles	**cluster 5**
Petroleum-gas-extraction	Wholesale-trade	Metal-mining
Stone-sand-clay-minerals	Retail-trade	**cluster 6**
Food-beverages	Hotel-restaurant	Tobacco
Textiles	Transport-land	**cluster 7**
Clothing	Transport-water	Leather
Wood	Storage-travel-agencies	**cluster 8**
Paper	Post-telecomm.	Office-machinery-computer
Publishing-printing	Finance	**cluster 9**
Petroleum-refinery	Insurance	ICT-equipment
Chemicals-pharma	Brokerage-credit-cards	**cluster 10**
Rubber-plastics	Real-estate	Transport-air
Glass-clay-cement-ceramic	Computer-services	**cluster 11**
Iron-steel-aluminium-tubes	Business-services	Renting-equipment
Structural-metal-products	Public-admin.	**cluster 12**
Mechanical-machinery	Education	R-D
Electrical-machinery	Health	**cluster 13**
Medical-precision-equip.	Refuse-disposal	Membership-organisations
Motor-vehicles	Arts-entertainment	**cluster 14**
Ships-railway-aircrafts	**cluster 2**	Personal-services
Furniture-Sports-Toys	Forestry	**cluster 15**
Recycling	**cluster 3**	Household-services
Electricity-gas	Fishing	
Water	**cluster 4**	

Table 2: Single linkages method; 15 clusters level

Complete linkages method		
cluster 1	Tobacco	Arts-entertainment
Agriculture	**cluster 9**	**cluster 10**
Food-beverages	Textiles	Wood
Petroleum-refinery	Clothing	Furniture-Sports-Toys
Chemicals-pharma	Leather	Renting-equipment
Transport-land	Paper	Personal-services
Insurance	Publishing-printing	**cluster 11**
Brokerage-credit-cards	Rubber-plastics	Office-machinery-computer
Health	Iron-steel-aluminium-tubes	ICT-equipment
Membership-organisations	Structural-metal-products	Medical-precision-equip.
cluster 2	Mechanical-machinery	R-D
Forestry	Electrical-machinery	**cluster 12**
cluster 3	Motor-vehicles	Ships-railway-aircrafts
Fishing	Electricity-gas	Transport-water
cluster 4	Sale-repair-vehicles	Transport-air
Coal Mining	Wholesale-trade	Storage-travel-agencies
cluster 5	Retail-trade	**cluster 13**
Petroleum-gas-extraction	Hotel-restaurant	Recycling
cluster 6	Post-telecomm.	**cluster 14**
Metal-mining	Finance	Water
cluster 7	Real-estate	Education
Stone-sand-clay-minerals	Computer-services	**cluster 15**
Glass-clay-cement-ceramic	Business-services	Household-services
Construction	Public-admin.	
cluster 8	Refuse-disposal	

Table 3: Complete linkages method; 15 clusters level

H is computed as:

$$\mathbf{H} = \mathbf{W} + \tilde{\mathbf{W}} = [h_{ij}]$$

s.t. $h_{ij} = 2$ if there is at least one direct or indirect path both between i and j and between j and i; $h_{ij} = 1$ if at least one path exists in only one direction; and $h_{ij} = 0$ otherwise.

Different approaches to QIOA basically consist of different definitions of significant edges, and thus of different ways of computing the filter value F — but also, as we will see in a moment, of different ways of consolidating successive layers.

The main drawback of QIOA is the fact of disregarding the magnitude of the flows once the minimum threshold for edges to be deemed as significant is defined and the boolean adjacency matrix is consistently obtained. However, as Leontief himself noticed,

> [t]he triangulation of a real input-output table — that is, the discovery of its peculiar structural properties — is a challenging task. It is complicated by the fact that one must take into account not only the distinction between zero and nonzero entries *but also the often more important difference between their actual numerical magnitudes.*
>
> (Leontief 1986[1963], p. 169, our italics)

A similar argument has been put forwrd by Mesnard:

> Boolean methods lead to lost of information and to a larger volume of computation than do quantitative methods. With regard to layers-based methods, [...] I have demonstrated that one layer provides the same information as the next, proving that such an approach is not informationally discerning. The above developments prove that if the aim of structural analysis is to detect the paths of influence, quantitative analysis is preferable to qualitative analysis.
>
> (Mesnard 2001, p. 278)

In the remainder of the section, we will consider three different approaches to QIOA: Important Coefficients (ICs) analysis, Elasticity Coefficients (ECs) analysis, and Minimal Flow Analysis (MFA).

3.1 Important Coefficients (ICs) Analysis

Aroche-Reyes (1996) adopts the ICs approach to the definition of significant edges. Such an approach originates from a seminal work by Jilek (1971),

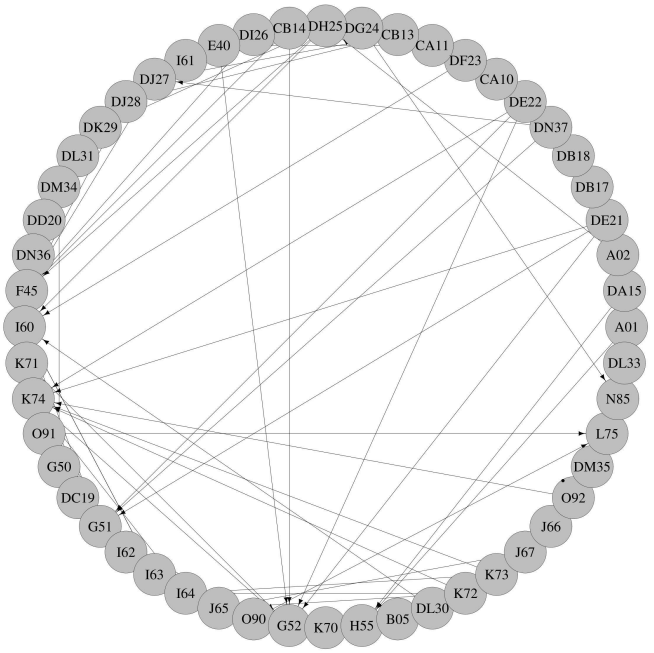
extending a contribution by Sherman & Morrison (1950) — dealing with the issue of how a change Δm_{ij} in one cell of a matrix $\mathbf{M}$ induces changes in all elements of the inverse $\mathbf{M}^{-1}$ — by including the concept of *tolerable limits*. In a few words, a *coeteris paribus* change in a technical coefficient a_{ij} modifies, via the direct and indirect linkages of both industries i ad j (i.e. via the induced modifications in the Leontief inverse), the gross output vector $\mathbf{g}$. The greater the 'importance' of the paths connecting industries i and j, the smaller the variation r_{ij}, or 'tolerable limit', able to induce a percentage change p change in g_i and/or g_j. After defining a filter value F for the r_{ij}s, the a_{ij}s associated to a $r_{ij} \leq F$ are deemed as ICs. The r_{ij} are computed according to the formula:

$$r_{ij} = a_{ij} \left(l_{ji} p + l_{jj} \frac{g_j}{g_i} \frac{1}{p} \right)$$

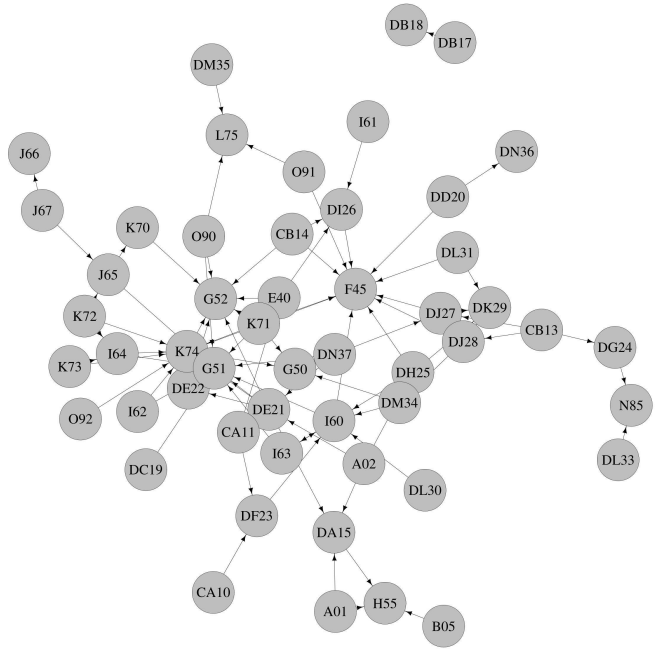
Aroche-Reyes choose the tolerable limit as 0.2. Accordingly, a boolean adjacency matrix $\mathbf{W} = [w_{ij}]$ is computed by setting $w_{ij} = 1$ if $r_{ij} \leq 0.2$, $w_{ij} = 0$ otherwise. Subsequent layers are obtained as detailed above, i.e. by computing boolean powers of matrix and $\mathbf{W}$ and then through boolean summation of such powers.

The same procedure has been here applied for Italy (2008); results are shown in Figures 2 and 3, displaying the first layer and all the cumulated layers, respectively. Figures 2a and 3a graphically organise ICs in the same way as Aroche-Reyes (1996) did, while Figures 2b and 3b organise them as structured graphs, in order to single out particularly connected groups of industries.

As it was straightorward to expect, significant connections in the first layer only are much less numerous than in the consolidated figure. However, the general picture does not change dramatically. Construction (F45), Wholesale-trade (G51), Retail-trade (G52) an Business-services (K74) are in both cases collecting many inflows, showing their 'centrality' in the inter-industry network, while Sale-repair-vehicles (G50) appears as much more central in the multi-layer picture than in the direct-layer one; Textiles (DB17) and Clothing (DB18) stand isolated as closely connected to each other and to no other industry in both cases. Agriculture (A01), Fishing (B05), Food-beverages (DA15) and Hotel-restaurant (H55) display significant connections in both figures, and the same holds for Finance (J65), Insurance (J66), Brokerage-credit-cards (J67), Real-estate (K70) and for Chemical-pharma (DG24), Medical-precision-equip. (DL33), Health (N85).

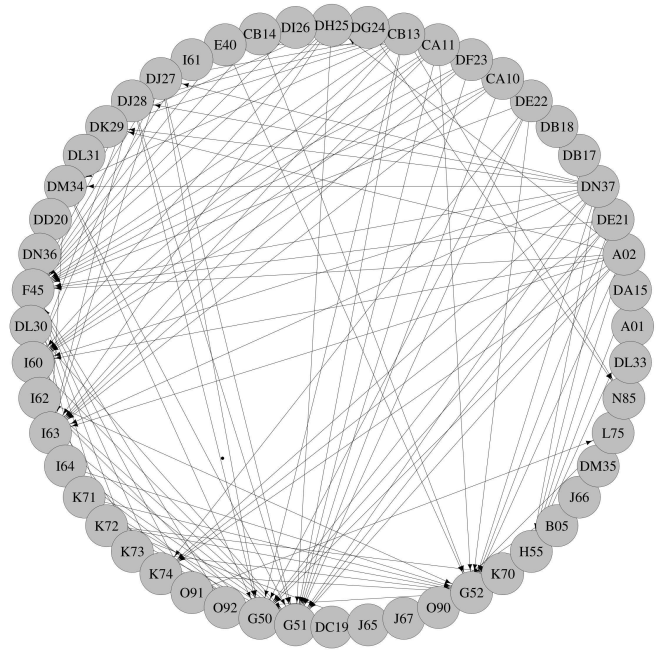


(a)

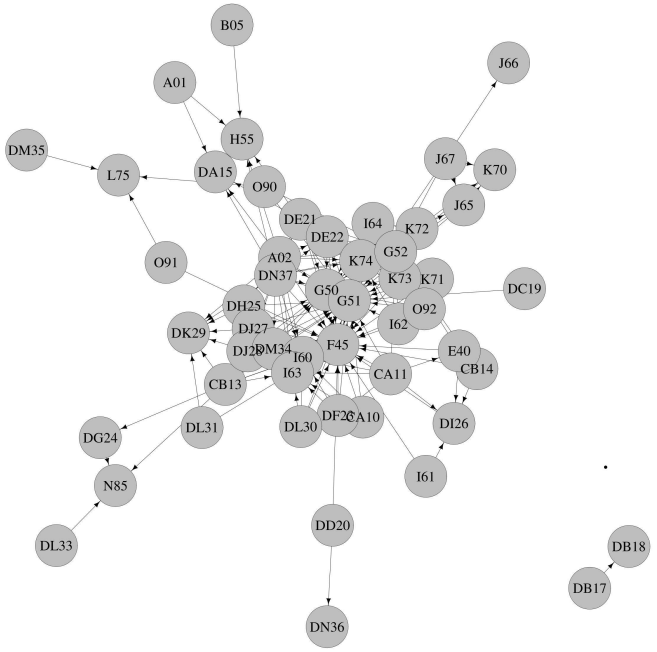


(b)

Figure 2: ICs for Italy (2008). First layer only



(a)



(b)

Figure 3: ICs for Italy (2008). Cumulated layers

The ICs method is based upon a definition of significant edges explicitly relying on direct input coefficients as appearing in matrix **A**. In other words, the very definition is based on the relative importance of each individual element of matrix **A** in determining the magnitude of the elements of the Leontief inverse. Given such definition, extending it to accordingly identify singificant ICs in the indirect layers of the (boolean) power series does not seem very consistent; the analysis of the direct layer only seems much more appropriate from this point of view. Of course, there still is the issue of arbitrariness in the choice of the threshold for the tolerable limit; the choice is influenced by the need of identifying a number of ICs big enough to single out a structure, or a complex path, linking the different nodes of the network, but at the same time small enough to keep the picture readable. The identification of industry clusters is then somewhat arbitrary too, since there is not a criterion for partitioning the network into communities. In order to find a possible way to overcome such difficulty, after introducing, in Section 5, Hoen's (2002) block-diagonalisation method, we will apply it to Aroche-Reyes boolean matrix to see whether it leads to identifying consistent industry clusters.

3.2 Elasticity Coefficient (ECs) Analysis

The ECs method was originally developed by Maaß (1980) and then applied by Schnabl (1995a) in comparison to the ICs method.

The procedure consists in computing the elasticity of each single element l_{ij} of the Leontief inverse with respect to each single element a_{kl} of input-coefficient matrix **A**, given a change $\mathrm{d}a_{kl} = a_{kl}(1+p)$:

$$\varepsilon_{l_{ij}a_{kl}} = \frac{\mathrm{d}l_{ij}}{\mathrm{d}a_{kl}}\frac{a_{kl}}{l_{ij}} = \frac{\mathrm{d}l_{ij}}{a_{kl}(1+p)}\frac{a_{kl}}{l_{ij}}, \quad i,j,k,l = 1,2,\ldots,n$$

and then taking, as an index of the *significance* of the connection between industries k and l, the maximum of all such elasticities:

$$Ec(p)_{kl} = \max_{ij}\{\varepsilon_{l_{ij}a_{kl}}\}$$

As Schnabl (1995a) pointed out, Maaß (1980) showed that the maximum is attained in correspondence of l_{kl}, and thus the correspondent value of EC_{kl} is given by:

$$Ec(p)_{kl} = \frac{l_{kk}a_{kl}l_{ll}}{l_{kl}(1 - pa_{kl}l_{lk}}$$

(see Schnabl 1995a, p. 497, equation (8))

As a second step, to introduce a measure more easily comparable to ICs, the the effect on gross output of a change $\mathrm{d}a_{kl} = a_{kl}(1+p)$, $k, l = 1, 2, \ldots, n$ is computed as:

$$Ec(p)^{\star}_{kl} = \frac{l_{kk} a_{kl} \sum_{j=1}^{n} d_j}{(1 - p a_{kl} l_{lk}) g_k}$$

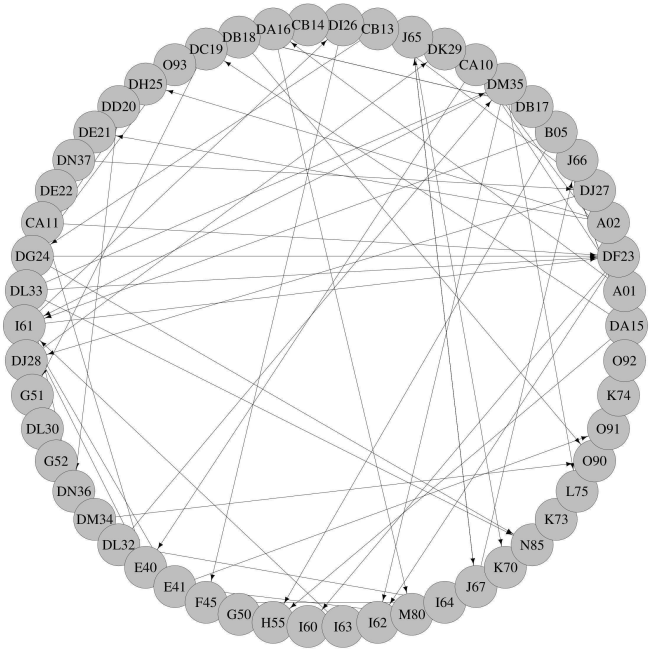
(see Schnabl 1995a, p. 498, equation (10))

The problem remains of choosing a filter level for the identification of singificant ECs; Schnabl (1995a) chose $F^{\star} = 0.05$ — as they pointed out (p. 501), value practically identical to the one normally adopted by ICs analysis. No reference is made to the value they chose for F;[3] the choice we made here is that of chosing a threshold leading to the identification of a number of ECs as close as possible to that of ICs. Since in the previous exercise we identificated, in the first layer alone, 87 such ICs, corresponding to 97.4% of total edges, we fixed F at the level of the corresponding quantile, i.e. $F \cong 0.89$.

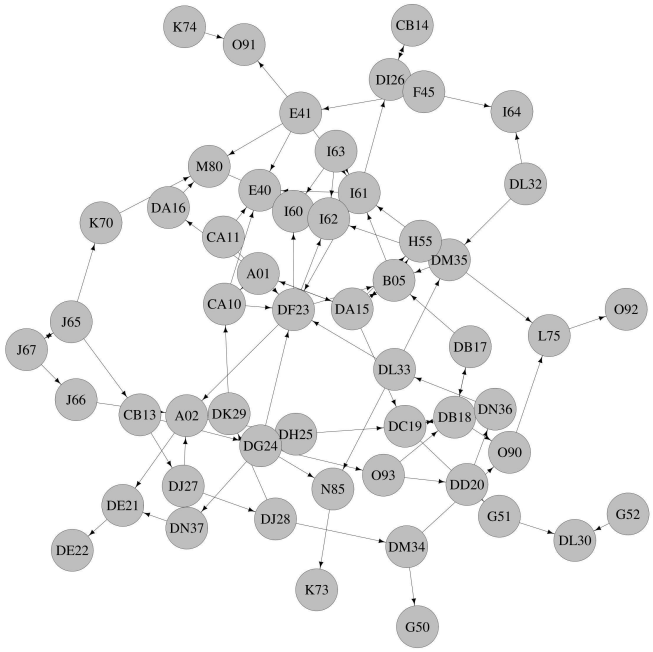
The results obtained for the case of Italy (2008) for ECs and EC*s are shown in Figures 4 and 5, respectively.

Let us look first at EC* in order to compare them to ICs. Construction (F45) and Retail-trade (G52) does not display particularly numerous significant ECs nor EC*s anymore. On the contrary, both Wholesale-trade (G51) and Business-services (K74) are still the center of many important EC*s (but not ECs), but interestingly such edges are now principally outgoing rather than in-going, as it was for ICs. Textiles (DB17) and Clothing (DB18) are still closely connected to each other, but they are not isolated anymore: they are now connected to the rest of the network via their connection to Electricity-gas (E40) when we consider EC*s; they appear as being much more connected to the whole inter-industry network when ECs are considered. Agriculture (A01), Food-beverages (DA15) and Hotel-restaurant (H55) also continue being significantly connected, but their EC*s with Fishing (B05) are now unsignificant (while ECs are significant), being the latter now strongly connected to Petroleum-refinery (DF23) and Electricity-gas (E40). Significant EC*s survive between Finance (J65), Insurance (J66) and Brokerage-credit-cards (J67) — Real-estate (K70) being also included in the

[3] "As in ICA, this is in the end a deliberate choice that can be made by taking some average that provides a reasonable pattern, somewhat similar to the endogenized structure of the MFA with respect to the number of depicted sectors" (Schnabl 1995a, p. 499).

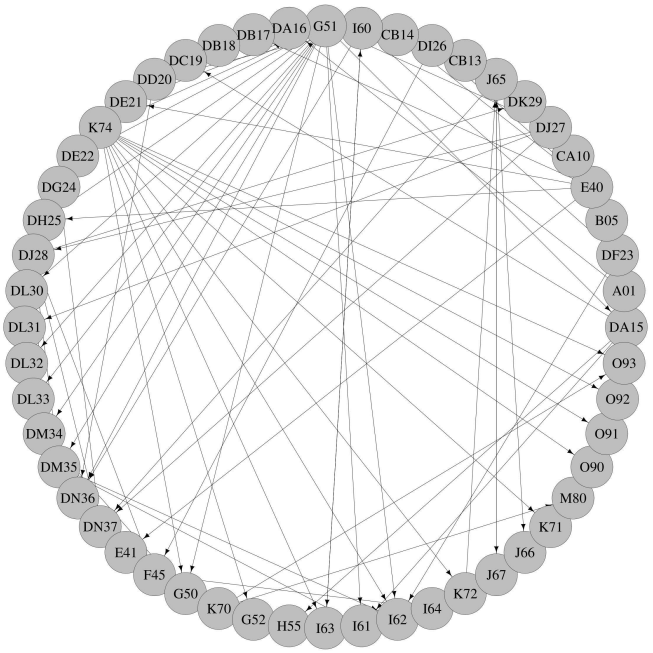


(a)

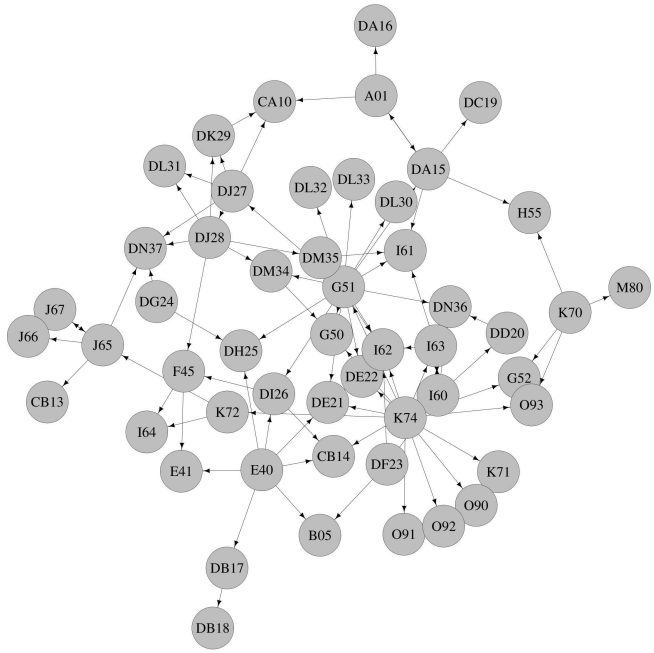


(b)

Figure 4: ECs for Italy (2008)



(a)



(b)

Figure 5: EC⋆s for Italy (2008)

case of ECs while being excluded from the group, and significantly connected to Retail-trade (G52), Hotel-restaurant (H55), Education (M80) and Personal-Services (O93), if we consider EC*s. In its turn, Chemical-pharma (DG24) is here significantly connected to Recycling (DN37) and Rubber-plastics (DH25) in the case of EC*s; to Personal-services (O93), Rubber-plastics (DH25), etroleum-refinery (DF23), Forestry (A02) and Recycling (DN37) in the case of ECs.

The ECs approach shares with the ICs one the basic idea of finding sificant flows between industry according to their relative influence on the magnitude of the elements of the Leontief inverse. In the introduction to Schnabl (1995a), it is stated that

> Aroche-Reyes (1996) introduced a new type of formal cutting rule in order to differentiate important/unimportant links, respectively sectors, and thus to determine intertemporal structural changes [...]. According to his arguments he looked for a simpler solution than MFA or QIOA [...] that had to solve the problem of deliberateness in choosing an adequate filter threshold.
>
> (Schnabl 1995a, p. 495)

The objection was however risen that

> Aroche-Reyes with his escape into IC-Analysis (ICA) and threshold-fixing did not really solve the problem of *deliberateness*, but only *shifted* it to the question, *why* 5.0 ($\sim r_{ij} = 0.2$) should be the appropriate filter value, besides the feature of being 'conventional', which explains nothing. The convention may well have practical use but lacks any theoretical reason so far.
>
> (Schnabl 1995a, p. 496)

Though being a consistent objection, the alternative procedure based on ECs does not solve the problem neither. A specular choice is necessary, and Schnabl's (1995a) solution was that of choosing a threshold equivalent to that chosen by Aroche-Reyes himself. Moreover, the rational at the basis of choosing ECs or EC*s in order to single out significant connetions is not clearly stated. This is not at all a trivial issue, since the two procedures have different theoretical, as well as practical, implications. In particular, it seems necessary to clearly state whether or not the connections of the inter-industry network to the boundaries, i.e. to final demand, needs to be considered when attempting at singling out the basic structure of intermediate flows.

3.3 Minimal Flow Analysis (MFA)

As mentioned above, QIOA was harshly criticised by Mesnard (Mesnard 1995, Mesnard 2001) on the basis of the argument that applying the multi-layer approach by using as the terms of the power series the powers of the boolean adjacency matrix induce to single out significant edges that turn out not to be so. Imagine to have two significant direct flows $a_{ik} > F$ and $a_{kj} > F$. Then the standard procedure would conclude that a significant indirect path of length two does exist between sectors i and j, even though the corresponding element of matrix $\mathbf{A}^2$ might in fact be below the filter level.

Schanbl's MFA (Schnabl 1994, Schnabl 2001) overcome this limitation by using, as the terms of the power series, a matrix $\mathbf{W}_k = [w_{ij}]$ for each layer k built in such a way as to make $w_{k_{ij}} = 1$ iff $t^k_{ij} \geq F$[4] and $w_{k_{ij}} = 0$ otherwise $(i, j = 1, 2, \ldots, n)$. The relevant flow matrix for each layer, $\mathbf{T}^k$, is built s.t.:

$$\mathbf{T}^0 = \mathbf{T}; \qquad \mathbf{T}^k = \mathbf{A}\mathbf{T}^{k-1}, \quad k = 1, 2, \ldots, n-2$$

and the corresponding boolean power series is computed in the *dependency matrix*:

$$\mathbf{D} = [d_{ij}] = \sum_{k=0}^{\bullet n-2} \mathbf{W}^k$$

where

$$\mathbf{W}^0 = \mathbf{W}_0, \quad \mathbf{W}^1 = \mathbf{W}_1\mathbf{W}^0 = \mathbf{W}_1\mathbf{W}_0, \ldots, \quad \mathbf{W}^K = \mathbf{W}_K\mathbf{W}^{K-1} = \prod_{k=0}^{\bullet K} \mathbf{W}_k$$

$d_{ij} = 1$ if at least one (directed) path of length $\lambda \leq n-2$ exists from industry i to j; $h_{ij} = 0$ otherwise.

Finally, the *connettivity matrix* is obtained as:

$$\mathbf{H} = [h_{ij}] = \mathbf{D} + \tilde{\mathbf{D}}$$

where $h_{ij} = 2$ if the connection between i and j is bidirectional; $h_{ij} = 1$ if it is unidirectional; $h_{ij} = 0$ if no connection exists.

[4] $T = [t_{ij}] = \{\mathbf{A}, \widehat{\mathbf{Ad}}\}$ according to the initial choice made on which IO flows to use for the analysis.

As to the choice of the flows to consider, or equivalently of the structure to uncover, Schnabl (2001) talks about *actual structure* when the influence of final demand is considered, and therefore $\mathbf{T} = \mathbf{A}\widehat{\mathbf{d}}$; of *technological structure* if only input coefficients are considered and therefore $\mathbf{T} = \mathbf{A}$.

In both cases, the boolean matrices $\mathbf{W}_k$ are obtained with reference to the chosen filter value F. In order to reduce the arbitrariness associated to the choice of the threshold, Schnabl (2001) endogenises it in the following way. First, the whole procedure is carried out with increasing filter values; in this way, the number of bilateral significant connections progressively decreases. F^{max} is found when no bilateral significant connections exist anymore.

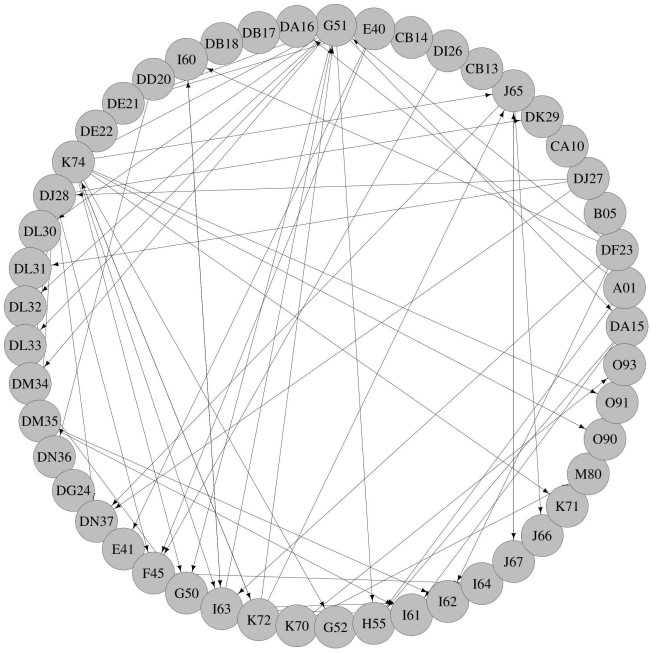
Secondly, the exercise is performed again for a sequence of 50 equidistant filters values $\in [0, F^{max}]$. For each of them, the frequency of 2, 1 and 0 entries in the connectivity matrix is computed. The optimal filter, $F^{\star}$, is that element of the sequence in correspondence of which the three frequencies are as similar as possible.

By applying the procedure above to the case of Italy (2008) we get $F^{\star} \cong 0.0087$ for the technological structure, and $F^{\star} \cong 273.99$ for the technological stucture. However, adopting such filter values would by definition lead to identify a very high number of significant connections.[5] This makes it hard to represent the associated graph as we did above for ICs and ECs. In order to be able to do so, we updated the procedure choosing the optimal filter value as the one leading to a number of non-zero entries in the dependency matrix as close as possible to 80.

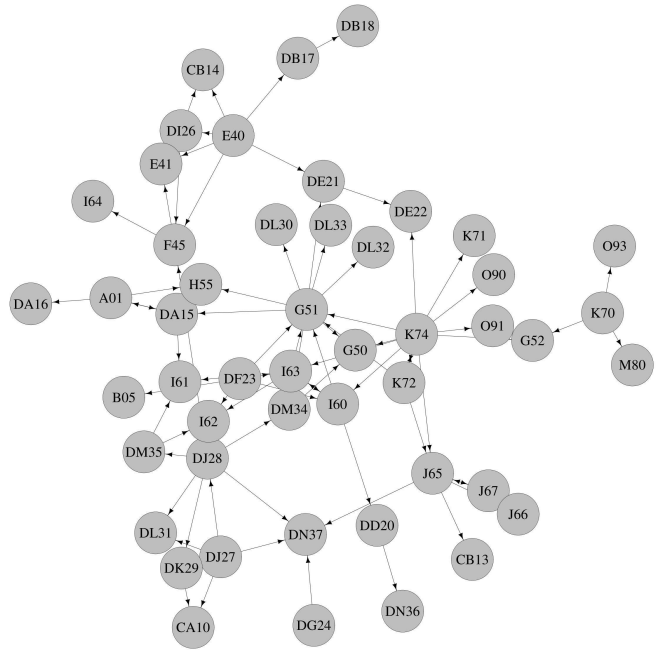
Results of the updated MFA analysis of Italian 2008 IO flows are shown in Figure 6 for the technological structure and 7 for the actual structure.

Let us have a look at Figures 6 and 7. By looking at the technological structure, the 'centrality' of Wholesale-trade (G51) and Business-services (K74) and, to a smaller extent, Sale-repair-vehicles (G50), Structural-metal-products (DJ28) and Construction (F45). Textiles (DB17) and Clothing (DB18) are also in this case closely connected to each other and to the rest of the network via the edge between the former and Electricity-gas (E40). The latter is also significantly connected to Glass-clay-cement-ceramic (DI26), Stone-sand-clay-minerals (CB14) and Construction (F45). Agriculture (A01), Food-beverages (DA15), Tobacco (DA16) and Hotel-restaurant

[5] More precisely, we found 798 unidirectional and 984 bidirectional edges in the case of actual structure; 800 unidirectional and 1020 bidirectional edges in the case of technological structure.

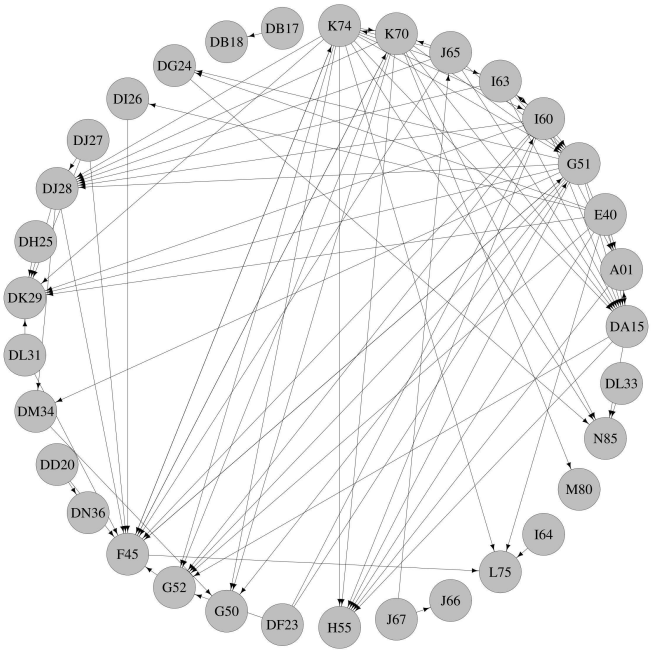


(a)

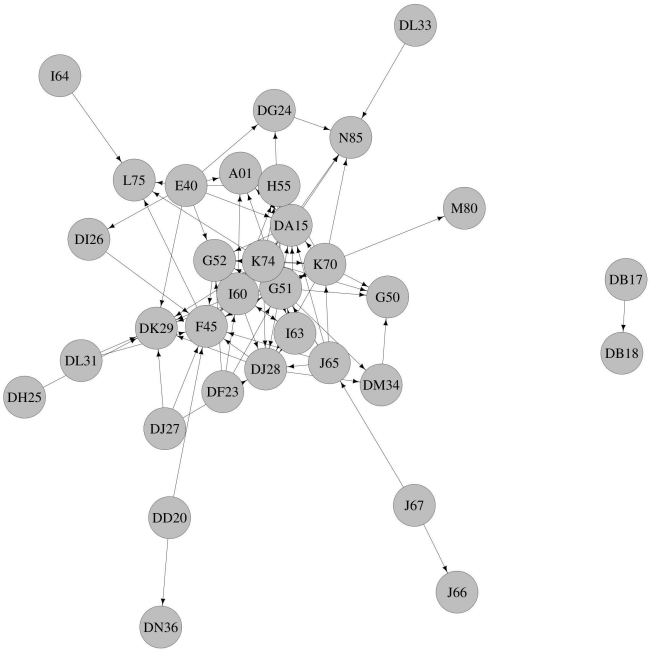


(b)

Figure 6: MFA for Italy (2008). Technological structure. Filter: $F^\star \cong 0.0539$



(a)



(b)

Figure 7: MFA for Italy (2008). Actual structure. Filter: $F^{\star} = 1795.635$

(H55) also conform a quite interconnected group of industries. The same holds for Finance (J65), Insurance (J66), Brokerage-credit-cards (J67) and for Real-estate (K70), Education (M80), Personal-services (O93) and Retail-trade (G52).

The actual structure displays a somewhat more intricated structure, with many industries in the centre of a quite intricated web[6] and some others more marginally connected to the rest of the network: Post-telecomm. (I64) via the flow to Public-admin. (L75); Medical-precision-equip. (DL33) via its deliveries to Health (N85); Rubber-plastics (DH25) delivering to Mechanical-machinery (DK29); Furniture-Sports-Toys (DN36) through it purchase to Wood (DD20), which in its turn sells intermediate inputs to Construction (F45); Education (M80) through its purchases from Real-estate (K70). Finally, we again observe Textiles (DB17) and Clothing (DB18) standing isolated from the rest of the network as an independent group.

The same criticisms raised at the end of the previous section to the ICs and ECs approach hold in the case of MFA; the same attempt at exploiting Hoen's (2002) block-diagonalisation will be performed in Section 5.

4 Clusters, Linkages and Interregional Spillovers

Oosterhaven et al.'s (2001) paper advances a method for singling out 'which direct linkages are important enough to be considered as potentially cluster-building' (Oosterhaven et al. 2001, p. 813); in particular, clusters are defined as

> industries [...] most closely tied together, in the sense that changes in any industry in a certain cluster are most likely to be passed to other industries in the same cluster instead of being passed on to industries in other clusters.
>
> (Oosterhaven et al. 2001, p. 812)

In order to identify such particularly strong direct linkages, three *quantitative* criteria are adopted, taking into account the magnitude of both absolute and relative intermediate flows, with the explicit aim of capturing

[6] Agriculture (A01), Food-beverages (DA15), Petroleum-refinery (DF23), Structural-metal-products (DJ28), Mechanical-machinery (DK29), Construction (F45), Sale-repair-vehicles (G50), Wholesale-trade (G51), Retail-trade (G52), Hotel-restaurant (H55), Transport-land (I60), Storage-travel-agencies (I63), Finance (J65), Real-estate (K70) and Business-services (K74).

differences between dependency and interdependency, which means that the *direction* of such flows is taken into account too.

More specifically, the three criteria concern the absolute size of intermediate flows, which must be a factor α larger than the average intermediate transaction between any two ibdustries; the relative size of intermediate sales, which must be a factor β_r larger than the average; finally, the intermediate sales, which must be a factor β_c larger than average. Formally, industries i and j are candidate to enter the same cluster when:

$$\begin{aligned} x_{ij} &> \alpha \mathbf{e}^T \mathbf{X} \mathbf{e} \\ a_{ij} &> \beta_r \mathbf{e}^T \mathbf{A} \mathbf{e} \\ b_{ij} &> \beta_c \mathbf{e}^T \mathbf{B} \mathbf{e} \end{aligned}$$

Oosterhaven et al. (2001) also state the secondary importance of the last two criteria, being them imposed upon relative rather than absolute intermediate flows; moreover, the threshold α is 'set as low as possible, subject to the requirement that the information may still be summarized and plotted visually' (Oosterhaven et al. 2001, p. 813). In order to do so, a natural break in the rank-size of the entries of matrix $\mathbf{X}$ around 40-60 linkages is looked for; α is then set equal to the ratio between such break and the average of matrix $\mathbf{X}$ entries. Linkages also satisfying the additional requirement of being at least a factor $\beta_r = 10$ or $\beta_c = 10$ larger than average in relative terms are deemed as specially significant and thus distinguished from those satisfying the main requirement only.

The exercise has been performed for the Italian case (2008).[7] Results are plotted in Figure 8.

Before looking at the results, it is worth stressing first that the natural break we looked for was around 60-80 linkages rather than 40-60, since the disaggregation adopted includes 10 industries more than those considered by Oosterhaven et al. (2001); this lead to setting, $\alpha = 9.04$, while in the case of Netherlands it has been set equal to 20 — which probably means that absolute intermediate flows are smoother in Italy than in the Neterlands. 74 linkages has been thus singled out as being α times larger than average; 27 linkages are also above average in relative terms.

[7] The procedure followed by Oosterhaven et al. (2001) is slightly different due to the fact that they use regional data and therefore set up a multiregional I-O matrix whose blocks correspond to the 13 Dutch regions. On the contrary, here we are solely using national data.

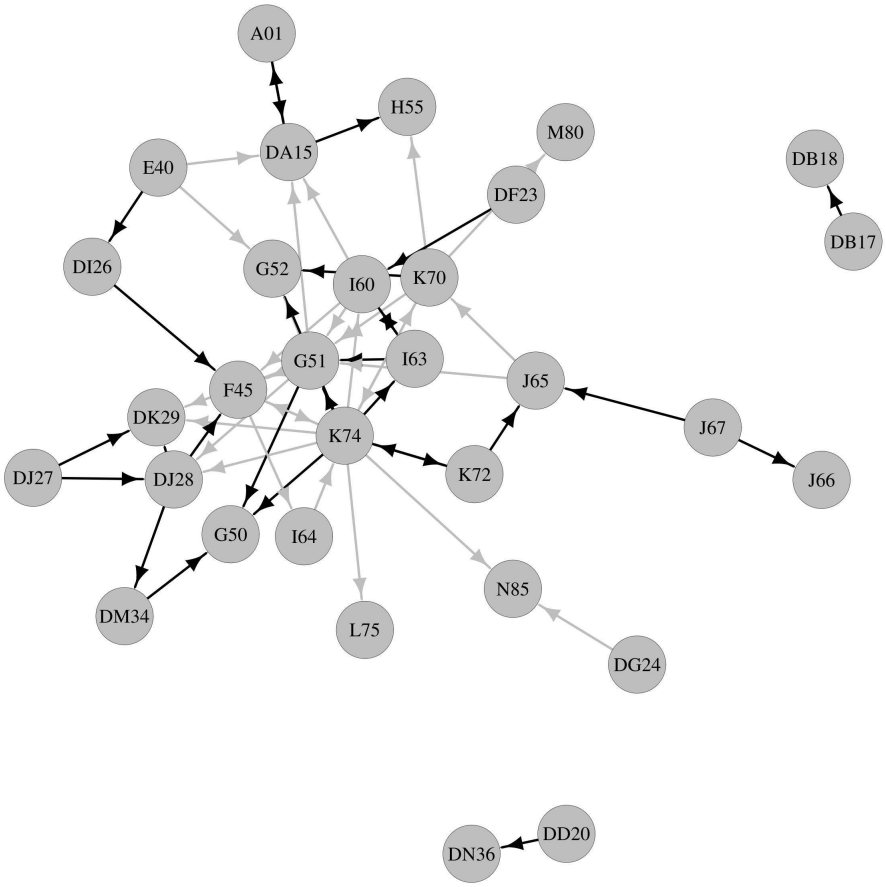


Figure 8: Grey arrows display linkages above average in absolute terms only; black ones linkages above average in relative terms too.

We can now inspect Figure 8. The feature worth beig stressed is that only 31 out of 58 industries display above average deliveries/purchases. Textiles (DB17) and Clothing (DB18) on the one side and Wood (DD20) and Furniture-Sports-Toys (DN36) on the other stand isolated from the rest of the network, their interrelations being above the average both in absolute and in relative terms. Insurance (J66) and Brokerage-credit-cards (J67) are connected to the rest of the network via the deliveries from the latter to Finance (J65); in the same way, Health (N85) and Chemicals-pharma (DG24) are connected to the rest of the network through the purchases of the former from Business-services (K74); Agriculture (A01) via its bilateral intermediate flows from and to Food-beverages (DA15). Iron-steel-aluminium-tubes (DJ27) sells intermediate inputs to Mechanical-machinery (DK29) and Structural-metal-products (DJ28); (DM34) buys from the latter and sells to Sale-repair-vehicles (G50). (I64) buys from (F45) and sells to (K74). Public-admin. (L75) is connected to the rest of the network through it purchases from Business-services (K74), while Education (M80) via its purchases from Real-estate (K70). Finally, Computer-services (K72) shows bilateral exchanges with Business-services (K74) and then sells its output to Finance (J65).

5 Block-diagonalisation (B-D) of adjacency matrix

This procedure has been introduced by Hoen (2002) as a way of identifying industry clusters starting from a boolean adjancency matrix **W** obtained by imposing certain restrictions on the connections in order to deem them as significant.

More specifically, $w_{ij} = 1$ iff $P(a_{kl} \leq a_{ij}) \leq 0.95 \cup P(x_{kl} \leq x_{ij}) \leq 0.95 \cup P(l_{kl} \leq l_{ij}) \leq 0.95$; $w_{ij} = 0$ otherwise. Rows and columns of the adjcency matrix are then permuted in order to block-diagonalise it; blocks correspond to industry clusters.

This methodology is explicitly inroduced by Hoen (2002) overcome the typical drawbacks of other methodologies, which he describes in the first part of the article. In particular, he deals with maximisation procedures which basically consist in hierarchical clustering successively selecting the maximum off-diagonal element, along the same lines as Slater (1977). By using Hoen's (2002) words,

> [t]he method continues in this way until an exogenously specified num-

> ber of clusters has been found, after which it terminates. Obviously, this general method has two important drawbacks: the method uses only one data source (i.e. the matrix of intermediate deliveries, the input matrix, the output matrix, or the Leontief inverse) and the number of clusters has to be specified in advance.
>
> (Hoen 2002, pp. 134-5)

According to Hoen (2002), some authors[8] implemented a 'restricted maximisation procedure' in order to overcome such drawbacks, consisting in imposing a set of restrictions on a set of IO matrices, and more specifically:

> (1) The intermediate delivery itself has to be larger than a constant α multiplied by the average of all intermediate deliveries.
> (2) The input coefficient has to be larger than a constant β multiplied by the average of all input coefficients.
> (3) The output coefficient has to be larger than a constant β multiplied by the average of all output coefficients.
>
> (Hoen 2002, p. 135)

However, this last method can still e criticised for the need of specifying in advance the desired number of clusters, and for the excessive dependence of the choice of the threshold values. Moreover, all the above mentioned methods lead to the identification of either mega clusters or mini clusters (groups of two industries only), which is not a desirable outcome.

Figure 9 shows the graph corresponding to the significant edges as emerging from Hoen's (2002) restriction matrix, as emerging from the Italian case for year 2008. 78 elements of the boolean restriction matrix are non-zero. Paper (DE21) and Publishing-printing (DE22) on the one side, and Finance (J65), Insurance (J66) and Brokerage-credit-cards (J67) on the other stand isolated in a separated group — and in fact Hoen's (2002) procedure identifies them as two (mini) clusters. Textiles (DB17) and Clothing (DB18) are still significantly connected to each other, but they are also connected to the rest of the network via the purchases of the latter from Electricity-gas (E40), which in its turn sells intermediate inputs to Rubber-plastics (DH25) and Glass-clay-cement-ceramic (DI26) as well. Wood (DD20) and Furniture-Sports-Toys (DN36) are still stronlgy connected to each other, the latter purchasing inputs from Transport-land (I60) and Wholesale-trade (G51).

[8] Hoen (2002) cites Eding, Oosterhaven & Stelder (1999). In fact, the three restrictions are equivalent to those chosen by Oosterhaven et al. (2001) in selecting the potential clusters-building intermediate flows.

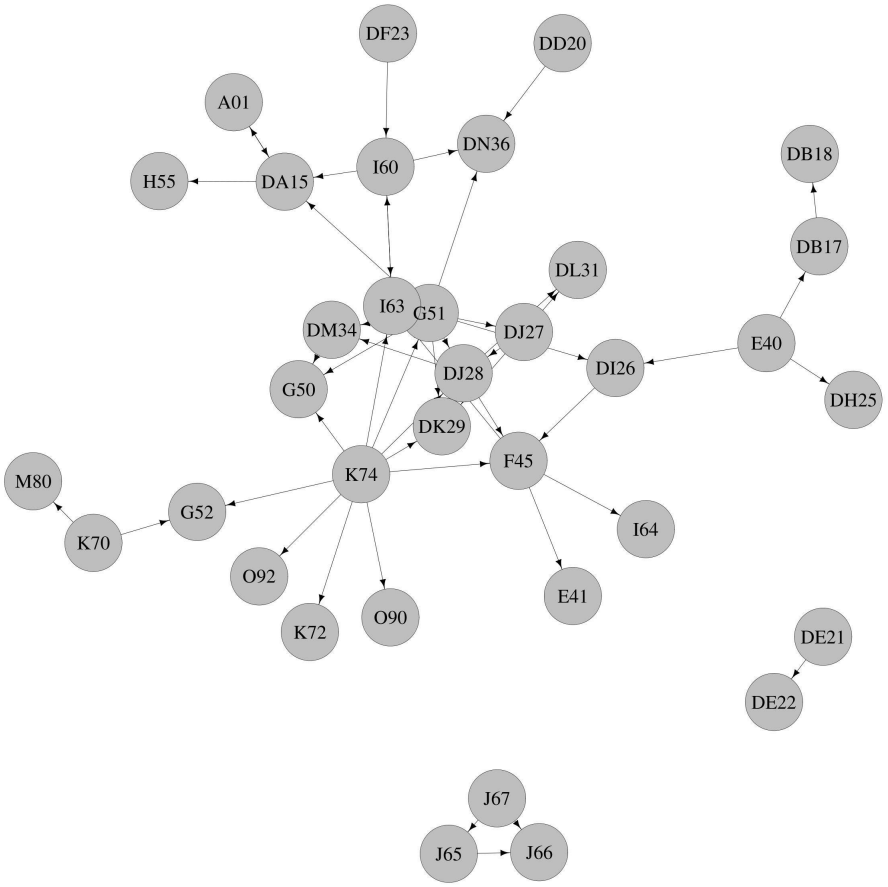


Figure 9: Hoen block diagonalisation method, significant linkages

Business-services (K74) and Construction (F45) are also in this case central activities in the network. The latter provides inputs to Post-telecomm. (I64) and Water (E41), which are 'terminal nodes' in this connection. Real-estate (K70) sells to Education (M80) and Retail-trade (G52), whose purchases from Business-services (K74) represent the connection of this group of industries to the rest of the network. It is still possible to identify as a group of closely connected activities Agriculture (A01), buying from and selling to Food-beverages (DA15), the latter then providing inputs to Hotel-restaurant (H55).

By looking at the corresponding column of table 4, and as anticipated above, 30 out of 58 activities are grouped in a 'megacluster'. Then we have two small clusters including Paper (DE21) - Publishing-printing (DE22) and Finance (J65) - Insurance (J66) - Brokerage-credit-cards (J67), respectively. The remaining 23 industries stand isolated.

Hoen (2002)		
cluster 1	Textiles	Real-estate
Agriculture	Clothing	Computer-services
Food-beverages	Rubber-plastics	Education
Wholesale-trade	Construction	Refuse-disposal
Glass-clay-cement-ceramic	Water	Arts-entertainment
Iron-steel-aluminium-tubes	Sale-repair-vehicles	**cluster 2**
Structural-metal-products	Hotel-restaurant	Paper
Mechanical-machinery	Transport-land	Publishing-printing
Electrical-machinery	Petroleum-refinery	**cluster 3**
Motor-vehicles	Storage-travel-agencies	Finance
Furniture-Sports-Toys	Post-telecomm.	Insurance
Wood	Business-services	Brokerage-credit-cards
Electricity-gas	Retail-trade	

Table 4: Block-diagonalisation, Hoen's procedure

We can now apply the same methodology to the boolean adjacency matrices as obtained following the approaches described above.

Table 5 shows the results of block-diagonalising Aroche-Reyes's adjacency matrix (direct layer only). 50 activities are grouped in one single megacluster; Textiles (DB17) and Clothing (DB18) conform an independent (mini) cluster, and the remaining six industries stand isolated.

Table 6 and 7 show the results for ECs and EC*s approaches, respectively. In the first case, all industries are grouped into one mega cluster, with the exception of Electrical-machinery (DL31), Renting-equipment (K71),

Aroche-Reyes (1996)		
cluster 1	Structural-metal-products	Storage-travel-agencies
Agriculture	Mechanical-machinery	Post-telecomm.
Forestry	Office-machinery-computer	Finance
Fishing	Electrical-machinery	Insurance
Coal Mining	Medical-precision-equip.	Brokerage-credit-cards
Petroleum-gas-extraction	Motor-vehicles	Real-estate
Metal-mining	Ships-railway-aircrafts	Renting-equipment
Stone-sand-clay-minerals	Furniture-Sports-Toys	Computer-services
Food-beverages	Recycling	R-D
Leather	Electricity-gas	Business-services
Wood	Construction	Public-admin.
Paper	Sale-repair-vehicles	Health
Publishing-printing	Wholesale-trade	Refuse-disposal
Petroleum-refinery	Retail-trade	Membership-organisations
Chemicals-pharma	Hotel-restaurant	Arts-entertainment
Rubber-plastics	Transport-land	**cluster 2**
Glass-clay-cement-ceramic	Transport-water	Textiles
Iron-steel-aluminium-tubes	Transport-air	Clothing

Table 5: Block-diagonalisation, Aroche-Reyes's procedure

Schnabl's ECs		
cluster 1	Mechanical-machinery	Transport-water
Agriculture	Medical-precision-equip.	Transport-air
Coal Mining	Motor-vehicles	Storage-travel-agencies
Food-beverages	Ships-railway-aircrafts	Post-telecomm.
Fishing	ICT-equipment	Finance
Tobacco	Furniture-Sports-Toys	Insurance
Textiles	Wood	Brokerage-credit-cards
Clothing	Recycling	Real-estate
Leather	Electricity-gas	Public-admin.
Petroleum-refinery	Water	Education
Forestry	Construction	Health
Petroleum-gas-extraction	Glass-clay-cement-ceramic	R-D
Paper	Stone-sand-clay-minerals	Refuse-disposal
Publishing-printing	Sale-repair-vehicles	Membership-organisations
Chemicals-pharma	Wholesale-trade	Business-services
Metal-mining	Office-machinery-computer	Arts-entertainment
Rubber-plastics	Retail-trade	Personal-services
Iron-steel-aluminium-tubes	Hotel-restaurant	
Structural-metal-products	Transport-land	

Table 6: Block-diagonalisation, Schnabl's ECs

Schnabl's EC⋆s		
cluster 1	Textiles	Transport-air
Agriculture	Transport-water	Storage-travel-agencies
Coal Mining	Clothing	Post-telecomm.
Food-beverages	Paper	Finance
Tobacco	Publishing-printing	Metal-mining
Leather	Petroleum-refinery	Insurance
Iron-steel-aluminium-tubes	Glass-clay-cement-ceramic	Brokerage-credit-cards
Structural-metal-products	Water	Real-estate
Mechanical-machinery	Construction	Retail-trade
Electrical-machinery	Sale-repair-vehicles	Computer-services
Motor-vehicles	Wholesale-trade	Business-services
Ships-railway-aircrafts	Office-machinery-computer	Renting-equipment
Recycling	ICT-equipment	Education
Chemicals-pharma	Medical-precision-equip.	Refuse-disposal
Rubber-plastics	Furniture-Sports-Toys	Membership-organisations
Electricity-gas	Wood	Arts-entertainment
Fishing	Hotel-restaurant	Personal-services
Stone-sand-clay-minerals	Transport-land	

Table 7: Block-diagonalisation, Schnabl's EC⋆s

Computer-services (K72) and Household-services (P95) standing isolated. In the second case, the megacluster involves 52 activities; those standing isolated are Forestry (A02), Petroleum-gas-extraction (CA11), R-D (K73), Public-admin. (L75), Health (N85) and Household-services (P95).

The results are not different when looking at the clusterisation obtained through MFA (Schnabl 1994, Schnabl 2001), presented in Tables 8 and 9. In the case of the technological structure, we find a megacluster including all activities, with the exception of Household-services (P95) — which of course does not have any linkage with any other activities, since it uses as an input direct labour only. Similarly, the actual structure shows Fishing (B05) , Coal Mining (CA10), Metal-mining (CB13), Tobacco (DA16), Office-machinery-computer (DL30) and Household-services (P95) as isolated industries, all the others being grouped together.

Finally, Table 10 shows the results of applying Hoen's (2002) procedure to Oosterhaven et al.'s (2001) approach. As already appearing from Figure 8, the 27 industries constituting the core of the network are grouped together in one mega cluster. Outside this group, we find Wood (DD20) and Furniture-Sports-Toys (DN36) on the one side and Textiles (DB17) and Clothing (DB18) constituting two separate (mini) clusters.

Schnabl's MFA (Technological structure)		
cluster 1	Structural-metal-products	Storage-travel-agencies
Agriculture	Mechanical-machinery	Post-telecomm.
Coal Mining	Electrical-machinery	Finance
Petroleum-gas-extraction	ICT-equipment	Metal-mining
Food-beverages	Medical-precision-equip.	Insurance
Fishing	Motor-vehicles	Brokerage-credit-cards
Tobacco	Ships-railway-aircrafts	Real-estate
Textiles	Furniture-Sports-Toys	Renting-equipment
Clothing	Recycling	Computer-services
Leather	Electricity-gas	R-D
Wood	Office-machinery-computer	Business-services
Paper	Water	Public-admin.
Publishing-printing	Construction	Education
Petroleum-refinery	Sale-repair-vehicles	Health
Forestry	Wholesale-trade	Refuse-disposal
Stone-sand-clay-minerals	Retail-trade	Membership-organisations
Chemicals-pharma	Hotel-restaurant	Arts-entertainment
Rubber-plastics	Transport-land	Personal-services
Glass-clay-cement-ceramic	Transport-water	
Iron-steel-aluminium-tubes	Transport-air	

Table 8: Block-diagonalisation, Schnabl's MFA, Technological Structure

Schnabl's MFA (Actual structure)		
cluster 1 Electrical-machinery	Post-telecomm.	
Agriculture	ICT-equipment	Finance
Petroleum-gas-extraction	Medical-precision-equip.	Insurance
Stone-sand-clay-minerals	Motor-vehicles	Brokerage-credit-cards
Food-beverages	Ships-railway-aircrafts	Real-estate
Textiles	Furniture-Sports-Toys	Renting-equipment
Clothing	Recycling	Computer-services
Leather	Electricity-gas	R-D
Wood	Water	Business-services
Paper	Construction	Public-admin.
Publishing-printing	Sale-repair-vehicles	Education
Petroleum-refinery	Wholesale-trade	Health
Chemicals-pharma	Retail-trade	Refuse-disposal
Rubber-plastics	Hotel-restaurant	Membership-organisations
Glass-clay-cement-ceramic	Transport-land	Arts-entertainment
Iron-steel-aluminium-tubes	Transport-water	Personal-services
Structural-metal-products	Transport-air	
Mechanical-machinery	Storage-travel-agencies	

Table 9: Block-diagonalisation, Schnabl's MFA, Actual Structure

Oosterhaven et al. (2001)		
cluster 1	Retail-trade	Public-admin.
Agriculture	Hotel-restaurant	Education
Food-beverages	Transport-land	Health
Electricity-gas	Petroleum-refinery	Chemicals-pharma
Glass-clay-cement-ceramic	Storage-travel-agencies	**cluster 2**
Construction	Post-telecomm.	Textiles
Structural-metal-products	Finance	Clothing
Iron-steel-aluminium-tubes	Brokerage-credit-cards	**cluster 3**
Mechanical-machinery	Insurance	Wood
Motor-vehicles	Real-estate	Furniture-Sports-Toys
Sale-repair-vehicles	Computer-services	
Wholesale-trade	Business-services	

Table 10: Block-diagonalisation, Oosterhaven et al.'s (2001) procedure

The main drawback of Hoen's (2002) procedure is that it seeks for groups of industries having significant connections among themselves only, while being totally disconnected from the rest of the network; this is the indirectly provided definition of an industry cluster. However, it is much more reasonable to expect the presence of relevant flows between different clusters too; or at least, a procedure for community detection should not rule out such a possibility. In other words, we are going to provide a definition of industry clusters different than Hoen's (2002), in order to allow for interdependence between industries belonging to different clusters.

6 Spectral Bisection (SB)

The SB algorithm for unweighted, directed graphs was presented by Leicht & Newman (2008) as a generalisation of Newman's (2006) algorithm for unweighted, undirected graphs, and can be straightforwardly generalised to take weighted flows into account. In all three cases, the logic at the basis of the algorithm is the same, and runs as follows.[9]

The starting point is the optimal partition of a network, which is defined as a division into indivisible subgraphs. In other words, the 'true' partition of a network into communities is found when all distinct communities have been detected, and thus none of them can be further divided into

[9]For analytical details on the most simple case and on the extension to the most general case, see Appendix A.1.

sub-communities. We are therefore indirectly provided with a definition of *community* as an *indivisible subgraph*. A quantitative measure of how good a division of a network into communities is is given by the associated modularity — a measure of 'statistical surprise' — its maximum value being attained in corresponding of the optimal partition.

Take a network with n nodes (industries) and k communities (industry clusters) and represent it with a transaction (square, industry by industry I-O) matrix $\mathbf{F}$; each cell f_{ij} represents the flows going from node i to node j (commodities evaluated at current prices sold by industry i to industry j) in the actual network. Now, imagine we have some limited information describing such network: the totals by row $\mathbf{s}_{out} = \mathbf{Fe}$ — i.e. total intermediate deliveries by industry of origin, or industries' *out-strength*; the totals by column $\mathbf{s}_{in}^T = \mathbf{e}^T\mathbf{F}$ — i.e. total intermediate purchases by industry of destination, or industries' *in-strength*; total inter-industry flows $m = \mathbf{e}^T\mathbf{Fe} = \sum_i s_{i,out} = \sum_i s_{i,in}$.

On the basis of such information, we can compute the *expected value* f_{ij}^e of each inter-industry flow f_{ij}:

$$f_{ij}^e = mP(i,j) = mP(i,\cdot)P(\cdot,j) = m\frac{s_{i,out}}{m}\frac{s_{j,in}}{m} = \frac{s_{j,out}s_{j,in}}{m}$$

f_{ij}^e is given by the monetary value of total deliveries times the probability $P(i,j)$ that one of such monetary units flows from industry i to j.[10]

The idea at the basis of modularity maximisation is that purchases and/or deliveries will be in general[11] greater than average within industries in the same cluster, and below average otherwise. The *modularity matrix* $\mathbf{B} = [b_{ij}]$ is given by the difference between actual and expected flows:

$$\mathbf{B} = \mathbf{F} - \frac{\mathbf{s}_{out}\mathbf{s}_{in}^T}{m}$$

[10] $P(i,\cdot)$ and is the probability that one unit of total intermediate production is delivered by industry i, and $P(\cdot,j)$ that one unit of total intermediate production is purchased by industry j. The resulting matrix of probabilities is not a transition matrix, whose rows sum to 1; here it is the sum of all the elements of the matrix that is equal to 1. In fact, the element t_{ij} of a transition matrix $\mathbf{T}$ is the probability of going to node j given the fact that we are starting from node i.

[11] There might be exceptions, especially for industries with exceptionally low or exceptionally high purchases/deliveries. The industries providing business services, for example, might deliver higher-then-average flows of inputs even to industries not belonging to their communities. On the contrary, industries with very few inter-industry connections can display lower-than-average exchanges even with industries in their same cluster.

Each element b_{ij}s will be positive if flows between industry i and j are above average, and negative otherwise. In the case of directed graphs, $\mathbf{B}$ is not symmetric and each element b_{ij} only takes into account deliveries from industry i to j ad not viceversa. To overcome this limitation, we compute the *generalised modularity matrix*

$$\tilde{\mathbf{B}} = \mathbf{B} + \mathbf{B}^T$$

whose elements $\tilde{b}_{ij} = \tilde{b}_{ji} = b_{ij} + b_{ji}$ correctly take into account flows going in both directions.

Take now an initial (tentative) subdivision of the network in two communities α and β, and the *membership vector* $\mathbf{m} = [m_i]$, with $m_i = 1$ for $i \in \alpha$ and $m_i = -1$ for $i \in \beta$. *Modularity* can then be computed as a weighted sum of the $\tilde{b}_{ij}$'s

$$Q = \mathbf{m}^T \tilde{\mathbf{B}} \mathbf{m}$$

the weights being $m_i m_j$, where $m_i m_j = +1$ if industries i and j are assigned to the same community, $m_i m_j = -1$ otherwise. Therefore, correctly assigning i) to the same group two industries which actually belong to the same cluster and ii) to different groups industries actually belonging to different clusters improve modularity; on the contrary, incorrectly i) separating industries belonging to the same cluster and ii) grouping industries belonging to different clusters reduces modularity.[12]

It can be shown (see Appendix A.1) that each industry can be assigned to community α or β according to the sign of the corresponding element of the leading eigenvector of matrix $\tilde{\mathbf{B}}$. After the first subdivision, the algorithm proceeds by further bisecting each resulting community, as long as such bisections lead to positive contributions to modularity.

After each consecutive bisection, a fine tuning of the result is performed; it consists in moving each node, one at a time and only once, to the other group; then, within the set of intermediate states occupied by network, the one associated to the maximum value of modularity is chosen. The procedure of repeatedly moving nodes from one group to the other is then repeated until no increase in modularity can be reached.

[12] Consider the case of three industries i, j and k connected by strong linkages between i and j and between i and k, while those between j and k are weak. Then, $\tilde{b}_{jk} < 0$. However, if $\tilde{b}_{ij} > |\tilde{b}_{jk}|$ separating j and k would increase modularity, but at the cost of a decrease greater than such increase due to separating i and j. In such a case, SB would group the three industries together, due to both their direct and indirect linkages.

McNerney (2009) stresses the fact that the SB algorithm gets a *partition*, in the mathematical sense of the word, of the network: overlapping communities are not allowed. It is our contention that this is not a drawback, at least for the case of I-O networks. There might be industries closely connected to all the others, which would fall in more than one community if overlaps were allowed. However, they are particularly connected to none of them. SB would identify them as isolated nodes. This does not necessarily mean that they have weak connections with all the others, but also that they are strongly connected with industries belonging to different communities. Moreover, a partition of the network is useful in order to make it possible to take advantage of the applications of linear operators — e.g. to compute the subsystems corresponding to different clusters.

SB		
cls 1, *Agri-Food*	**cls 4, *Transport Serv.***	Glass-clay-cement-ceramic
Agriculture	Petroleum-gas-extraction	Construction
Fishing	Petroleum-refinery	**cls 7, *Fashion-Arts***
Food-beverages	Office-machinery-computer	Textiles
Tobacco	Ships-railway-aircrafts	Clothing
Hotel-restaurant	Sale-repair-vehicles	Retail-trade
cls 2, *Printed Media*	Transport-land	Real-estate
Forestry	Transport-water	Business-services
Leather	Transport-air	Education
Paper	Storage-travel-agencies	Membership-organisations
Publishing-printing	Renting-equipment	Arts-entertainment
Wholesale-trade	**cls 5, *Financial Services***	**cls 8, *Wood Products***
cls 3, *Heavy Machinery*	Metal-mining	Wood
Coal Mining	ICT-equipment	Furniture-Sports-Toys
Rubber-plastics	Post-telecomm.	**cls 9, *Bio-Tech***
Iron-steel-aluminium-tubes	Finance	Chemicals-pharma
Structural-metal-products	Insurance	Medical-precision-equip.
Mechanical-machinery	Brokerage-credit-cards	Electricity-gas
Electrical-machinery	Computer-services	Water
Motor-vehicles	R-D	Health
Recycling	Public-admin.	Personal-services
	cls 6, *Construction*	
	Stone-sand-clay-minerals	

Table 11: SB

Table 11 shows the results of applying SB to the Italian case for 2008. Two industrie, namely Refuse-disposal (O90) and Household-services (P95),

stand isolated; the others are grouped into 9 clusters.

The same results are presented graphically in Figures 10 — showing the whole web of above average connections between nodes, each cluster being assigned a different colour — and 11, showing the internal structure of each cluster found.

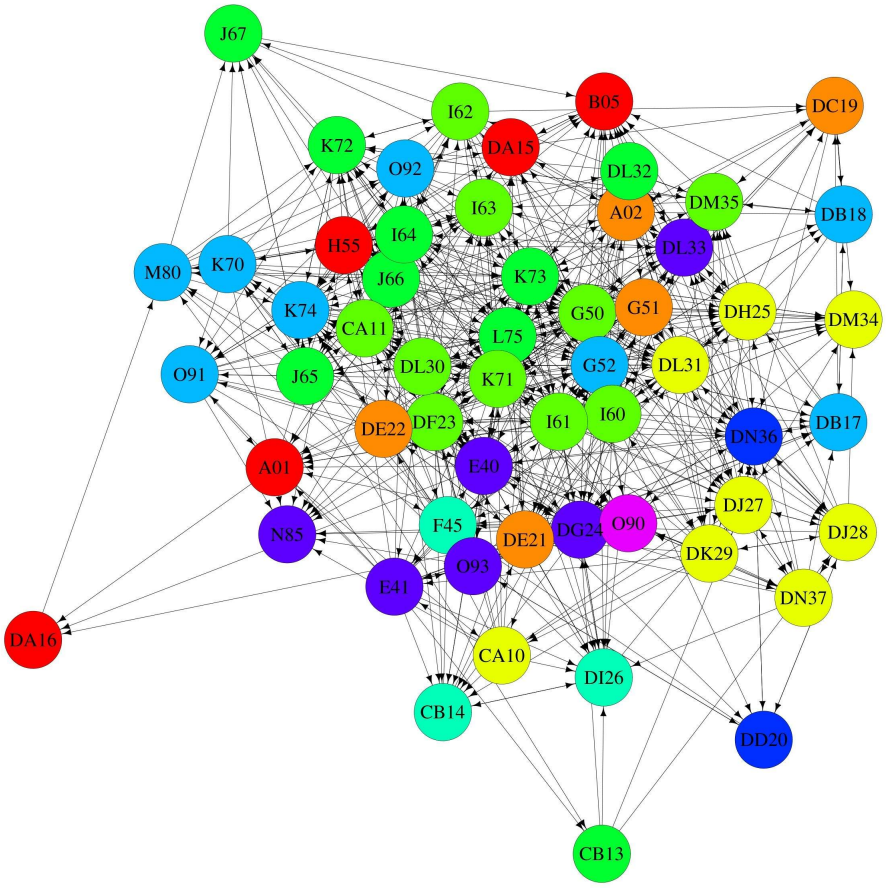
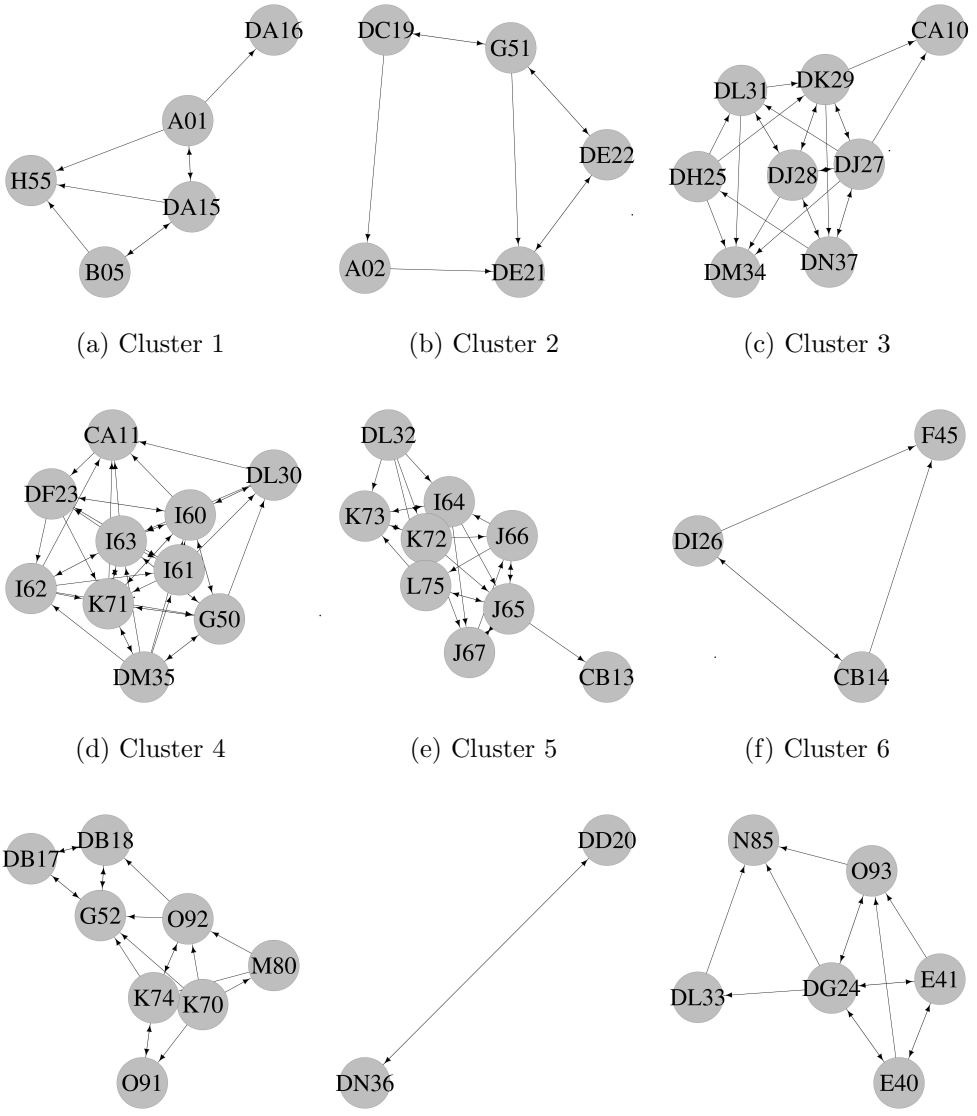


Figure 10: Above-average connections, spectral bisection



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3

(d) Cluster 4 (e) Cluster 5 (f) Cluster 6

(g) Cluster 7 (h) Cluster 8 (i) Cluster 9

Figure 11: Clusters' graphs, SB. Italy (2008)

7 Random walks and Markov chains

Piccardi's (2011) approach is based on a LMC model of a random walk. The definition of a community is based on the concept of *persistence probability*, i.e. the probability that a random walker currently in a node of a cluster remains in the cluster itself in the following step. Calling α such a probability, an α-community is a group of nodes whose persistence probability is not less than α. The partition of a network into α-communities is an α-partition. Given a certain partition of a network into communities and a level of α which is considered significant, computing the persistence probability of each community and comparing it to α is a way of testing the significance of each community. Based on this idea, Piccardi (2011) also proposes an algorithm for optimal partition of a network.

The basic idea is that a group of industries can be considered as a cluster when there is a high relatively probability that a flow of money, once it reaches the group, iteratively goes from one industry to the others in a relatively persistent loop. Notice that this definition implies a particular topology that a group of industries must show in order to be defined as a cluster; more specifically, all the industries must have a relatively symmetric role. A vertically integrated production chain might be seen as a scarcely significant community: industry i at the top of the chain, in fact, might display a close relation with those at the bottom as input providers, but no relations in the opposite direction, delivering all its output to other industries/clusters as inputs or to the final sector as final demand. In this case, a flow reaching industry i would immediately leave the community, thus lowering its persistence probability. On the contrary, it might be recognised as a community by other approaches, such as SB, which simply defines a cluster as a group of industries with above-than-average connections in either direction.

First, we have to compute the *transition matrix* $\mathbf{M}^T$. Define:

$$\mathbf{W} \equiv \begin{bmatrix} \mathbf{X} & \mathbf{d} \\ \mathbf{z}^T & 0 \end{bmatrix}$$

Then

$$\mathbf{M} = \mathbf{W}\widehat{\mathbf{W}}^{-1}$$

Formally, matrix $\mathbf{M}^T$ is a *Markov* — i.e. row-stochastic[13] — matrix: it is the transition matrix associated to an N-state Markov chain, and its LHS

[13]By construction, the sums of the rows are all equal to one, and thus the dominant eigenvalue is unitary.

leading eigenvector $\boldsymbol{\pi}^T$ is the stationary Markov chain state probability distribution.[14]

Based on this intermediate result, and given a certain candidate partition of the network, Piccardi (2011) then proposes a methodology for providing a quantitative measure of how significant each community of the partition — consistent with his own definition of significant community — and thus the partition itself, is. Such a methodology is then crucial for the community finding algorithm which he proposes in the second part of the paper.

First of all, define a *candidate* partition $\mathbb{P}_m$, partitioning the network in m communities, and the corresponding $(n \times m)$ *collecting matrix* $\mathbf{H} = [h_{i\gamma}]$, $(i = 1, \ldots, n, \gamma = 1, \ldots, m)$, where $h_{i\gamma} = 1$ if node i belongs to community γ, and $h_{i\gamma} = 0$ otherwise. It is therefore possible to define a *meta-network* composed by m *meta-nodes* given by the m communities; such a meta-network is characterised by a transition, or *Lumped Markov*, matrix, $\mathbf{U}$ defined as:

$$\mathbf{U} = (\widehat{\boldsymbol{\pi}^T\mathbf{H}})^{-1}\mathbf{H}^T\widehat{\boldsymbol{\pi}}\mathbf{M}^T\mathbf{H} \qquad \text{s.t.} \qquad \boldsymbol{\Pi}^T = \boldsymbol{\Pi}^T\mathbf{U}$$

where $\boldsymbol{\Pi}^T$ is the stationary LMC state probability distribution.

The elements of matrix $\mathbf{U} = [u_{\gamma\beta}]$ can be written as:

$$\mathbf{U} = [u_{\gamma\beta}] = \frac{\sum_{i\in\gamma}\pi_i \sum_{j\in\beta} m_{ji}}{\sum_{i\in\gamma}\pi_i}$$

i.e. the probability that a random walker currently in community γ finds itself in community β in the following step. In other words, and recalling the interpretation of the random walker and of probabilities π_i given above, $u_{\gamma\beta}$ is the fraction in *in a stationary state* — i.e. after a number of identical repetitions of the production process such that the weighted inter-industry transaction matrix finally converges — of the monetary value, at current prices, of cluster γ's total purchases which are paid as the counterpart of purchases from cluster β.

Morover, each element $u_{\gamma\gamma}$ $(\gamma = 1, \ldots, m)$ of the main diagonal of matrix $\mathbf{U}$ is the persistence probability associated to the corresponding community. By comparing it with the benchmark value attributed to α, it is possible to assess whether the communities are α-communities and thus whether the

[14] Since $\mathbf{M}^T$ is non-negative — and irreducible, by construction, being our network strongly connected — for Perron-Frobenius Theorems (PFTs) the dominant eigenvector is not repeated and strictly positive.

partition is an α-partition. Moreover, it is possible in this way to assess the significativity of the single communities rather than that of the whole partition, which means that it is possible to conclude that some communities are actually significant while some others are not.

Let us now see how this methodology can be applied in order to detect communities in a network. The basic idea is that of iteratively grouping the closest nodes of a network in order to get a tree — called *dendrogram* — which has the single nodes at one extreme and the whole network at the other. Each possible section of the dendrogram is a candidate partition of the network, and the above described methodology can be useful in choosing the most significant one.[15]

It is therefore necessary to define a measure of the distance between nodes consistent with the initial definition of a community. Piccardi (2011) provides a definition according to which, given an arbitrarily short RW of length T, the more often the random walker reaches i starting from j and j starting from i, the closer nodes i and j are. In other words, the greater the proportion of monetary flows going from industry i and industry j, the closer the two industries are. Formally, a measure of how close i an j are to each other is given by

$$\sigma_{ij} = \sigma_{ji} = \sum_{t=1}^{T} [P]^t(i,j) + [P]^t(j,i) \tag{7.1}$$

where $[P]^t(i,j) = (\mathbf{M}^T)^t_{ij} + \mathbf{M}^t_{ij}$ is the probability that a random walker starting from i is in j after t steps, and thus the sum in equation (7.1) is the expected number of total passages from i to j or from j to i during a RW of length T. Thus, the *symmetric* closeness matrix $\mathbf{\Sigma}$ is given by:

$$\mathbf{\Sigma} = \sum_{t=1}^{t} \left((\mathbf{M}^T)^t + \mathbf{M}^t \right)$$

Starting from matrix $\boldsymbol{\sigma}$ is then possible to draw the dendrogram and its optimal section can be picked up by evaluating the persistence probabilities of the communities in each corresponding partition.

[15] The usual choice is that of picking the partition with the greatest number of communities among those considered significant.

8 Alternative interpretation

The fact that IO networks are absorbing prevents us from following the procedure described by Piccardi (2011); using the 'complete' transition matrix $\mathbf{M}^T$ would not make sense, since the boundaries are not 'ordinary' nodes: they are the entrance (demand) or the exit from (value added) the network. We are only interested in the submatrix where the last row and column are eliminated:

$$\tilde{\mathbf{M}} = \mathbf{X}\left(\widehat{\mathbf{e}^T\mathbf{X} + \mathbf{z}^T}\right)^{-1}$$

Now the transition matrix is not stochastic anymore; the leading eigenvalue will be less than one, and the leading eigenvector cannot be given the same interpretation as Piccardi's (2011).[16]

However, it is still possible to give $\boldsymbol{\pi}$ an economic interpretation. Consider the power method for computing the leading eigenvector of a matrix: it consists of choosing an arbitrary vector $\boldsymbol{\pi}_0 \neq \mathbf{0}$ and then iteratively computing the product $\boldsymbol{\pi}_{t+1}^T = \tilde{\mathbf{M}}\boldsymbol{\pi}_t$ up to the point where $\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t = \boldsymbol{\pi}$. Whatever the initial choice, the resulting vector will always be the same (up to a scalar multiple): the leading eigenvector of matrix $\tilde{\mathbf{M}}$.

In economic terms, each element $\pi_{i,0}$ of the chosen initial vector can be interpreted as an additional flow of money entering industry i as a consequence of an increase in the demand faced by it. This additional flows will circulate throughout the network; at each successive passage, a part of it will be absorbed by the final sector in the form of value added, and thus that remaining in the inter-industry network will become smaller and smaller at each passage, finally converging to zero. Whatever the industry/industries facing the shock, and whatever its magnitude, after a certain number of stages the proportions of the resulting monetary flows going to the different industries converge; such proportions depend on the particular structure of the inter-industry relative flows: the more 'central' an industry is, i.e. the more it is connected to the others, the higher its absorption of indirect flows. This measure of how 'central' each industry is in the corresponding network is known in literature as *eigenvector centrality*. Table 12 summarises the

[16] Computing the transition matrix after eliminating the last row and column, thus still obtaining a stochastic matrix, would not be appropriate, because in that way we would be disregarding the fact that a part of any monetary flow associated to intermediate transactions immediately leaves the inter-industry network reaching the boundaries in the form of value added.

results for Italy in 2008.

Eigenvector Centrality; $\lambda_M = 0.44$					
Brokerage-credit-cards	4.12	Petroleum-refinery	2.18	Motor-vehicles	0.95
Computer-services	3.79	Transport-air	2.17	Leather	0.93
Renting-equipment	3.64	Arts-entertainment	2.04	Construction	0.93
Stone-sand-clay-minerals	3.6	Rubber-plastics	2.04	Textiles	0.85
Petroleum-gas-extraction	3.6	Wholesale-trade	1.95	Hotel-restaurant	0.8
Recycling	3.5	Office-machinery-computer	1.92	Furniture-Sports-Toys	0.79
Business-services	3.41	Metal-mining	1.89	Retail-trade	0.74
Paper	3.24	Structural-metal-products	1.72	Insurance	0.73
Storage-travel-agencies	2.97	Membership-organisations	1.68	Fishing	0.65
Electricity-gas	2.87	Iron-steel-aluminium-tubes	1.61	Medical-precision-equip.	0.56
Forestry	2.86	Ships-railway-aircrafts	1.51	Mechanical-machinery	0.55
Publishing-printing	2.84	ICT-equipment	1.39	Clothing	0.36
Post-telecomm.	2.64	Water	1.37	Education	0.25
Transport-land	2.57	Electrical-machinery	1.29	Tobacco	0.09
Finance	2.52	Chemicals-pharma	1.22	Personal-services	0.06
R-D	2.36	Food-beverages	1.11	Health	0.03
Glass-clay-cement-ceramic	2.26	Transport-water	1.1	Public-admin.	0.02
Wood	2.23	Sale-repair-vehicles	1.04	Household-services	0
Refuse-disposal	2.23	Coal Mining	1.02		
Agriculture	2.19	Real-estate	0.99		

Table 12: Eigenvector centrality

In this context, the dominant eigenvalue also has a specific interpretation: given the particular structure of the IO network, after a demand shock taking place in all industries in the same relative proportions as the elements of the dominant eigenvector, a proportion $1 - \lambda_M$ of the additional monetary flows would be immediately absorbed by the final sector as value added, while a proportion λ_M would continue to flow in the IO network as indirect flows. The smaller such eigenvalue with respect to 1, the more important the indirect flows (with respect to direct ones) in the specific IO network considered.

This interpretation of the dominant eigenvalue of matrix $\tilde{\mathbf{M}}$ suggests another measure of industries centrality, based on the variation of the value of λ_M associated to the removal of a row and the corresponding column. The higher the reduction of the dominant eigenvalue following such extraction, the greater the consequent change in the *structure* of inter-industry flows, and thus the greater the importance of the corresponding industry in determining it. Table 13 shows the measure for the Italian case in year 2008.

A comparison of these two Tables may be helpful in understanding some characteristics of the Italian inter-industry structure.

As stated above, eigenvector centrality provides a measure of the relative importance of the monetary flows channeled by inter-industry relations and reaching each industries; in other words, of the importance of each industry in (directly and indirectly) providing inputs to the others. On the contrary, a *change* in the dominant eigenvalue induced by the removal of each industry is a measure of the change in the whole *structure* of inter-industry relations induced by such removals.

The five most central industries in Table 12 are Brokerage-credit-cards (J67), Computer-services (K72), Renting-equipment (K71), Stone-sand-clay-minerals (CB14) and Petroleum-gas-extraction (CA11), attracting, respectively, the 4.12%, 3.79%, 3.64%, 3.6% and 3.6% of the inter-industry money flows. Among these, only Computer-services (K72) also induce a significant change in the dominant eigenvalue (2.73%), ranking eigth in the second Table. The removal of the others would induce very small changes in the whole inter-industry structure, λ_M changing less than 1%.

Let us now look at the first five position in the table concerning maximum eigenvalue reduction. Here, we find Wholesale-trade (G51), Business-services (K74), Transport-land (I60), Construction (F45) and Storage-travel-agencies (I63) — the change in λ_M being 8.56%, 7.92%, 5.85%, 4.65% and 4.39%, respectively. Business-services (K74), Storage-travel-agencies (I63) and Transport-land (I60) are also quite central — the corresponding entries in the first Table being 3.41%, 2.97% and 2.57%, respectively. Wholesale-trade (G51) and Construction (F45) are much less central (1.95% and 0.93%, respectively) though shaping in a strong way the structure of inter-industry relations.

Removing Coal Mining (CA10) or Metal-mining (CB13) would leave the dominant eigenvalue unchanged; however, they are not totally irrelevant as to their eigenvector centrality (1.02% and 1.89%, respectively).

Of course, the two measures have a quite distinct nature, since eigenvector centrality is an 'absolute' one, while the second is a 'relative' one, stressing changes rather thn absolute values. This observation suggests the possibility of getting a relative measure out of eigenvector centrality as well. More specifically, one could compute the percentage change in components of the dominant eigenvector resulting form the extraction of each industries. Such variations could be stored into a square matrix $\mathbf{\Pi} = [\pi_{ij}]$, where $\pi_{ij} < 0$ if removing industry i would reduce industry j's centrality; positive otherwise.

Maximum Eigenvalue Reduction (%)					V2
Wholesale-trade	8.56	Rubber-plastics	1.35	ICT-equipment	0.29
Business-services	7.92	Chemicals-pharma	1.32	Insurance	0.29
Transport-land	5.85	Motor-vehicles	1.17	Clothing	0.27
Construction	4.65	Electrical-machinery	1.1	R-D	0.24
Storage-travel-agencies	4.39	Brokerage-credit-cards	0.94	Medical-precision-equip.	0.18
Structural-metal-products	3.13	Refuse-disposal	0.91	Membership-organisations	0.13
Food-beverages	3	Ships-railway-aircrafts	0.89	Office-machinery-computer	0.07
Computer-services	2.73	Arts-entertainment	0.81	Petroleum-gas-extraction	0.07
Post-telecomm.	2.7	Furniture-Sports-Toys	0.78	Health	0.05
Hotel-restaurant	2.61	Real-estate	0.74	Education	0.04
Electricity-gas	2.55	Renting-equipment	0.74	Public-admin.	0.03
Retail-trade	2.34	Petroleum-refinery	0.73	Fishing	0.02
Sale-repair-vehicles	2.23	Transport-air	0.59	Personal-services	0.01
Glass-clay-cement-ceramic	2.1	Wood	0.55	Forestry	0.01
Finance	1.97	Leather	0.51	Metal-mining	0
Mechanical-machinery	1.79	Stone-sand-clay-minerals	0.41	Tobacco	0
Iron-steel-aluminium-tubes	1.75	Textiles	0.4	Coal Mining	0
Publishing-printing	1.59	Transport-water	0.37	Household-services	0
Agriculture	1.58	Water	0.34		
Paper	1.46	Recycling	0.33		

Table 13: Maximum eigenvalue reduction

The usefulness of such a matrix is straightforward: it could be used clusters identification, by using it for computing modularity. This means that the algorithm, instead of running starting from modularity matrix $\mathbf{B}$, would start from a 'modified' modularity matrix defined as:

$$\mathbf{B}^{\star} \equiv \mathbf{O} - \mathbf{\Pi}$$

SB based on eigenvector centralities reduction		
cls 1, *Agri-Food*	Iron-steel-aluminium-tubes	**cls 10, *Research-ICT***
Agriculture	Structural-metal-products	ICT-equipment
Fishing	Mechanical-machinery	Post-telecomm.
Food-beverages	Electrical-machinery	Computer-services
Wholesale-trade	Recycling	R-D
Hotel-restaurant	**cls 6, *Construction***	Business-services
Health	Stone-sand-clay-minerals	Education
cls 2, *Printed Media*	Glass-clay-cement-ceramic	**cls 11, *Vehicles***
Forestry	Construction	Motor-vehicles
Paper	Transport-water	Refuse-disposal
Publishing-printing	**cls 7, *Clothing***	**cls 12, *Air Transp.***
cls 3, *Energy*	Textiles	Ships-railway-aircrafts
Coal Mining	Clothing	Transport-air
Electricity-gas	Leather	**cls 13, *Equipm. Serv.***
cls 4 *Transport-HiTech*	Retail-trade	Sale-repair-vehicles
Petroleum-gas-extraction	Personal-services	Renting-equipment
Petroleum-refinery	**cls 8, *Wood Prods***	**cls 14, *Finance***
Office-machinery-computer	Wood	Finance
Medical-precision-equip.	Furniture-Sports-Toys	Insurance
Transport-land	**cls 9, *Chemicals***	Brokerage-credit-cards
Storage-travel-agencies	Chemicals-pharma	Real-estate
cls 5, *Heavy Mach.*	Rubber-plastics	**cls 15, *PA-Arts***
Metal-mining		Public-admin.
		Arts-entertainment

Table 14: Clusters based on maximum eigenvalue reduction

The results of applying such procedure to our data for Italy are displayed in Table 14 and graphically represented in Figures 13 and 12. We are now in aposition to compare them with those obtained by the standarb SB algorithm.

First of all, besides the four industries standing isolated — Household-services (P95), Tobacco (DA16), Water (E41) and Membership-organisations (O91), we can identify 15 clusters, while the standard procedure detected 9 communities.

The former *Agri-Food* cluster is now larger, also including Health (N85) and Wholesale-trade (G51). *Prited Media* becomes smaller, and do not include anymore Leather (DC19) — which goes into the new *Clothing* cluster — and Wholesale-trade (G51). The former *Heavy Machinery* also becomes smaller, including Iron-steel-aluminium-tubes (DJ27), Structural-metal-products (DJ28), Mechanical-machinery (DK29), Electrical-machinery (DL31), Recycling (DN37) with the addition of Metal-mining (CB13). Rubber-plastics (DH25) conforms the new *Chemicals* cluster together with Chemicals-pharma (DG24); Motor-vehicles (DM34) also conforms a new cluster, *Vehicles*, together with Refuse-disposal (O90). Cluster 4 *Transport Services* breaks too: a core, including Petroleum-gas-extraction (CA11), Petroleum-refinery (DF23), Office-machinery-computer (DL30), Transport-land (I60) and Storage-travel-agencies (I63) conform, together with Medical-precision-equip. (DL33), the new *Transport-HiTech* cluster; Transport-water (I61) goes into contruction. Ships-railway-aircrafts (DM35) and Transport-air (I62) on the one side, and Sale-repair-vehicles (G50) and Renting-equipment (K71) on the other side, separate from the core constituting two separate clusters: *Air Transport* and *Vehicles*, respectively.

Another formerly big cluster, namely *Financial Services*, also breaks. As mentioned above, Metal-mining (CB13) is part of the new *Heavy Machinery* cluster, while Public-admin. (L75) and Arts-entertainment (O92) form the new *PA-Arts* one. Finance (J65), Insurance (J66), and Brokerage-credit-cards (J67) join Real-estate (K70) in the new *Finance* cluster; finally, ICT-equipment (DL32), Post-telecomm. (I64), Computer-services (K72) and R-D (K73), together with Business-services (K74) and Education (M80), form the new *Research-ICT* cluster.

As to *Construction*, we already mentioned that the eigenvector-based algorithm finds a bigger cluster than the traditional one, also including Transport-water (I61). The *Fashion-Arts* cluster disappears, its components going into different communities. Textiles (DB17), Clothing (DB18) and Retail-trade (G52) join Leather (DC19) and Personal-services (O93) into the new *Clothing* cluster; as already mentioned, Real-estate (K70) enters the *Finance* one, while Business-services (K74) and Education (M80) the *Research-ICT* and Arts-entertainment (O92) the *PA-Arts* one. Membership-organisations (O91) stands isolated.

The *Bio-Tech* cluster disintegrates, with Chemicals-pharma (DG24) joining Rubber-plastics (DH25) into a specific *Chemicals* group; Medical-precision-equip. (DL33) entering the *Heavy Machinery* cluster and Electricity-gas (E40)

the *Energy* one. Health (N85) goes into the new *Agri-Food* cluster and Personal-services (O93) into *Clothing*, while Water (E41) stands isolated.

The only community that appears the same as a result of both procedures is *Wood Products*.

9 Conclusions

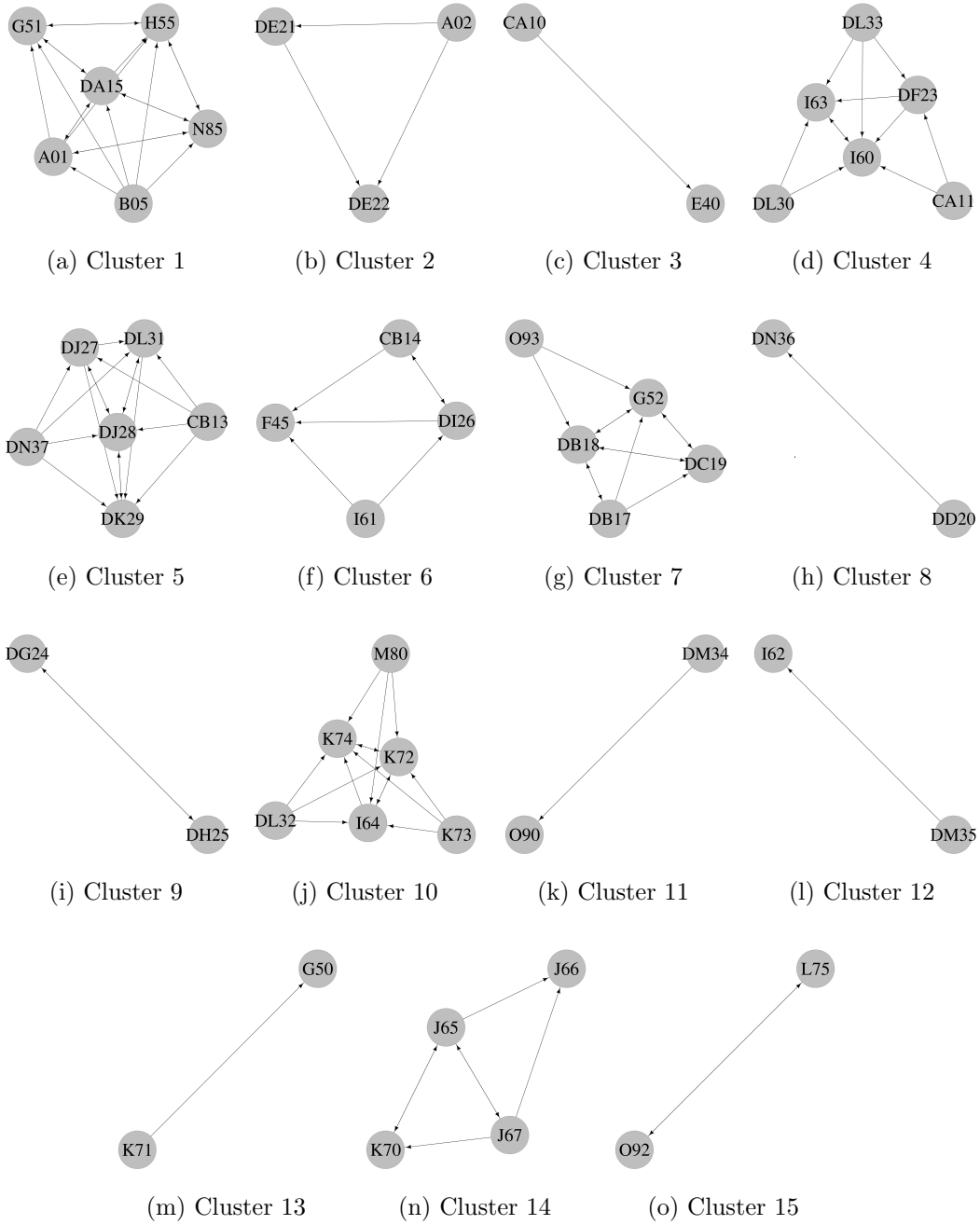
Our updated SB method leads to the identification of smaller, and often more meaningful, clusters than the traditional one. The modified modularity matrix on which it is based has a clearcut economic meaning, and it somehow includes many of the ideas at the basis of the methodologies reviewed in the present paper. It tries to combine approaches coming from traditional IO literature and graph-theoretical hints and procedures. It takes full advantage of the magnitude of the direct, absolute inter-industry transactions, but it also takes into account their circular character.

This last observation leads us back to the Introduction. The goodness of a method cannot but be evaluated in light of the definition of communities which one wants to search for. I share Oosterhaven et al.'s (2001) that an industry cluster should be identified on the basis of direct, rather than indirect flows, and that absolute transactions rather than input coefficients should be taken into account. Other procedures, such as Pasinetti's (1988) vertical hyper-integration, focus on the relation between inter-industry structure and the composition of final demand for consumption commodities.[17] Such analyses are of outmost importance for the understanding of the structure of an economic system, and for the implementation of income, labour and fiscal policies. However, it is my contention that they are *complementary* to community detection, not alternative.

In this direction, further lines of research could be oriented to uncovering the possible complementarities between these two levels of analysis. In particular, spectral decompositions are an essential tools for the analysis of subsystems, and could provide a bridge between the two approaches.

Another important bridge is the analysis of international trade, which can nowadays take advantage of an umprecedented data availability, and in which the interaction of 'horizontal' and vertical integration plays an essential role.

[17] Or final demand as traditionally intended, thinking of Pasinetti (1973).



(a) Cluster 1 (b) Cluster 2 (c) Cluster 3 (d) Cluster 4

(e) Cluster 5 (f) Cluster 6 (g) Cluster 7 (h) Cluster 8

(i) Cluster 9 (j) Cluster 10 (k) Cluster 11 (l) Cluster 12

(m) Cluster 13 (n) Cluster 14 (o) Cluster 15

Figure 12: Clusters' graphs, eigenvector-centrality SB. Italy (2008)

A Appendix

A.1 SB algorithm

Newman's (2006) original article dealt with unweighted, undirected graphs. In such case, a complete description of the flows of the network is provided by the so-called *adjacency matrix* $\mathbf{A} = [a_{ij}]$, where $a_{ij} = k$ is the number of connections between node i and j exists. Since we exclude, for simplicity, multiple connections, $a_{ij} = \{0, 1\}$. The sums by row (or, equivalently, the sums by column) of matrix $\mathbf{A}$ give the number of connections involving the corresponding node, or *nodes' degrees* $\mathbf{k} = \mathbf{Ae}$. The total number of connections is $m = (1/2)\sum_i k_i$ (since flows are symmetric, and self-loops are excluded.)

In this case, the expected number of connections between nodes i and j is given by twice the total number of connections times the probability of having a connection between i and j:

$$a^e_{ij} = 2mP(i,j) = 2mP(j,i) = 2mP(i,\cdot)P(j,\cdot) = 2m\frac{k_i}{2m}\frac{k_j}{2m} = \frac{k_i k_j}{2m}$$

and thus

$$\mathbf{B} = \mathbf{A} - \frac{\mathbf{k}\mathbf{k}^T}{2m}$$

When we consider weighted, though still undirected, flows — as Leicht & Newman (2008) did — we have to slightly change the above definitions. Instead of the adjacency matrix we have the (still symmetric) *weights matrix* $\mathbf{W} = [w_{ij}]$ giving the volume of the flows between nodes. The vector of nodes degrees is given by $\mathbf{k} = \mathbf{We} = (\mathbf{e^T W})^T$, m is defined as above and the expected value of w_{ij} is:

$$w^e_{ij} = 2mP(i,j) = 2mP(j,i) = 2mP(i,\cdot)P(j,\cdot) = 2m\frac{k_i}{2m}\frac{k_j}{2m} = \frac{k_i k_j}{2m}$$

and thus

$$\mathbf{B} = \mathbf{W} - \frac{\mathbf{k}\mathbf{k}^T}{2m}$$

Matrix $\mathbf{B}$ is a symmetric matrix where all the rows (and columns) sum to zero.[18] This means that it is singular and therefore that it has an eigenvalue

[18] In fact, $\mathbf{W}^e$ has been computed starting from $\mathbf{We}$ and $\mathbf{e}^T\mathbf{W}$, and therefore by construction $\mathbf{We} = \mathbf{W}^e\mathbf{e}$.

$\lambda_0 = 0$ with associated eigenvector $\mathbf{v}_0 = \mathbf{e}$. Moreover, being symmetric, its eigenvectors are orthogonal, thus matrix $\mathbf{V} = [\mathbf{v}_i]$ is orthogonal and it holds that:

$$\mathbf{V}^T = \mathbf{V}^{-1} \qquad \mathbf{V}^T\mathbf{B}\mathbf{V} = \mathbf{\Lambda}$$

where $\mathbf{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues λ_i of matrix $\mathbf{B}$.

Given the properties of the modularity matrix, Q can be written as:

$$Q = \mathbf{m}^T\mathbf{V}\mathbf{V}^{-1}\mathbf{B}\mathbf{V}\mathbf{V}^{-1}\mathbf{m} = \mathbf{m}^T\mathbf{V}\mathbf{V}^T\mathbf{B}\mathbf{V}\mathbf{V}^T\mathbf{m} = \mathbf{m}^T\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T\mathbf{m} = \sum_i (\mathbf{u}_i^T\mathbf{s})^2\lambda_i \tag{A.1}$$

According to equation (A.1) states, modularity can be written as a linear combination of the eigenvalues of matrix $\mathbf{B}$, the weights being $(\mathbf{u}_i^T\mathbf{s})^2$. Choosing $\mathbf{m}$ in order to maximise modularity thus means maximising the weight associated to the leading eigenvalue, λ_M, while minimising the others. The optimal solution would be that of choosing $\mathbf{m}$ proportional to the leading eigenvector, $\mathbf{v}_M$; in so doing, being the eigenvectors of $\mathbf{B}$ orthogonal, all the other weights would be zero. But since we are subject to the constraint $m_i = \pm 1$, the best we can do is choosing $\mathbf{m}$ so as to maximise $\mathbf{u}_i^T\mathbf{s}$.

If $\mathbf{B}$ has some positive eigenvalues, $\lambda_M \neq \lambda_0$ and thus $\mathbf{v}_M \neq \mathbf{v}_0$. Being the wo eigenvectors orthogonal, we are sure that $\mathbf{v}_M$ has both positive and negative components, and therefore maximising $\mathbf{u}_i^T\mathbf{s}$ means fixing $m_i = +1$ if $\mathbf{v}_{Mi} > 0$, and $m_i = -1$ if $\mathbf{v}_{Mi} < 0$.

If, on the contrary, all eigenvalues are non-positive, then λ_0 is the leading eigenvalue, $\mathbf{v}_0$ the leading eigenvector, and $\mathbf{m} = \mathbf{v}_0 = \mathbf{e}$: the network is an indivisible graph, and thus the only community is the whole network itself.

It was therefore proved that the network can be bisected by dividing the nodes into communities α and β according to te sign of the elements of the leading eigenvector of the modularity matrix.

The following steps consist in iteratively bisecting the existing groups, until no bisection can further increase modularity. However, after the first iteration the procedure needs to be slightly modified. If we treated the subgraphs in the same way as we did with the whole network, we would be disregarding inter-modular flows, and thus we would not be able to compute the modularity if the whole network, which depends on $\mathbf{B}$ and thus on the flows connecting *all* nodes. Rather, as stressed by Newman (2006), we compute the *additional contribution* to modularity ΔQ of each further bisection.

Let us take subgraph α as an example. In order to further bisect it into α_1 and α_2, define a new membership vector $\mathbf{m}_\alpha$ and a reduced modularity matrix $\mathbf{B}_\alpha$ extracting from $\mathbf{B}$ the rows and columns corresponding to nodes belonging to subgraph α. Each cell b_{ij} in α previously entered the sum in equation (A.1) with positive sign; with the new bisection, they will have to keep the positive sign if i and j both belong to α_1 or α_2; on the contrary, their sign has to turn negative if $i \in \alpha_1$ and $j \in \alpha_2$. Therefore, ΔQ only depends on the flows connecting nodes which were separated by the further bisection; such nodes have to be subtracted twice from Q: the first time to annihilate their previously positively accounted contribution; the second time to correctly consider the corresponding flow with negative sign:

$$\Delta Q = \mathbf{m}_\alpha^T \mathbf{B}_\alpha \mathbf{m}_\alpha - \mathbf{m}_\alpha^T \widehat{(\mathbf{B}_\alpha \mathbf{e})} \mathbf{m}_\alpha = \mathbf{m}_\alpha^T \left(\mathbf{B}_\alpha - \widehat{(\mathbf{B}_\alpha \mathbf{e})} \right) \mathbf{m}_\alpha = \mathbf{m}_\alpha^T \mathbf{B}^{(\alpha)} \mathbf{m}_\alpha$$

$\mathbf{B}^{(\alpha)}$ is symmetric, since the sums by row have been subtracted to the main diagonal. Therefore, the problem of finding the vector $\mathbf{m}_\alpha$ which maximises ΔQ — and thus the bisection of α wit maximum contribution to modularity — is formally the same as that of finding the vector $\mathbf{m}$ which maximises Q, and can be analogously solved by choosing the $m_{\alpha,ij} = \pm 1$ according to the sign of the elements of the leading eigenvector $\mathbf{v}_M^{(\alpha)}$ of matrix $\mathbf{B}^{(\alpha)}$.

References

Aroche-Reyes, F. (1996). Important Coefficients and Structural Change: A Multi-layer Approach. *Economic Systems Research*, 8(3):235–246.

——— (2001). The Question of Identifying Industrial Complexes Revisited: A Qualitative Perspective. In *Input-Output Analysis: Frontiers and Extensions*, pp. 280–296. Palgrave.

——— (2002). Structural Transformations and Important Coefficients in the North American Economies. *Economic Systems Research*, 14(3):257–273.

——— (2003). A Qualitative Input-Output Method to Find Basic Economic Structures. *Papers Reg. Sci.*, 82:581–590.

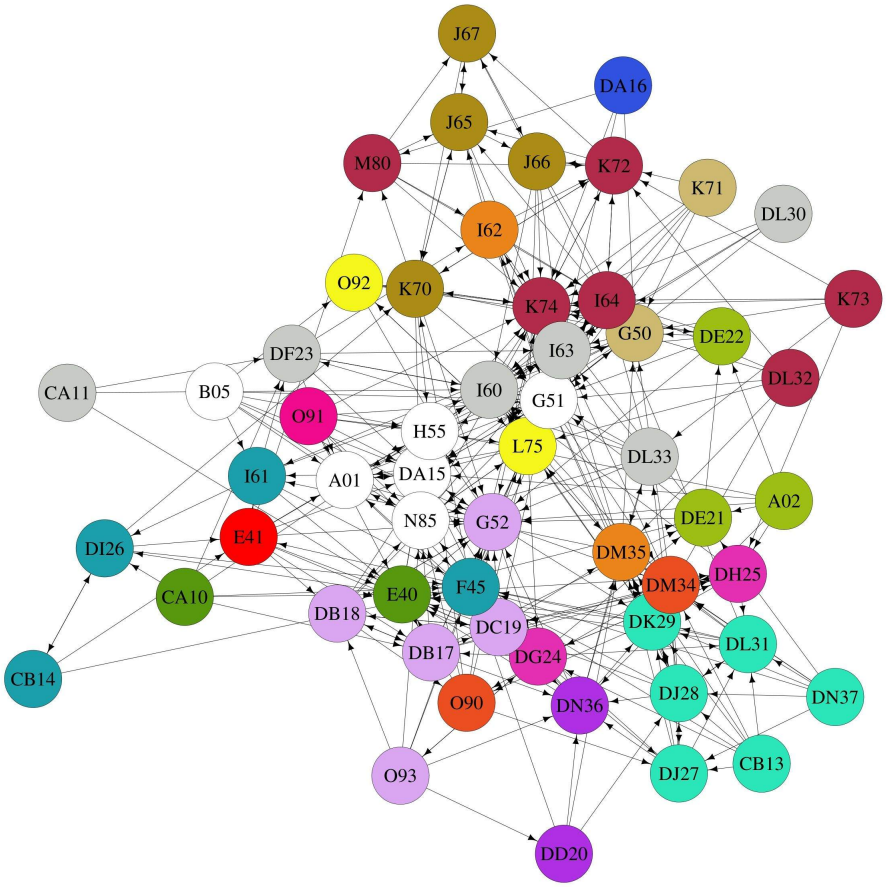


Figure 13: Important connections, based on variations in the maximum eigenvector, spectral bisection

——— (2006). Trees of the Essential Economic Structures: a Qualitative Input-Output Method. *Journal of Regional Science*, 46(2):333–353.

Bon, R. (1989). Qualitative Input-Output Analysis. In *Frontiers of Input-Output Analysis*, pp. 222–231. Oxford University Press.

Defourny, J. & Thorbecke, E. (1984). Structural Path Analysis and Multiplier Decomposition within a Social Accounting Matrix Framework. *The Economic Journal*, 94(373):111–136.

Drejer, I. (18-22 May 1998). Technological Interdependence in the Danish Economy — A Comparison of Methods for Identifying Knowledge. In *Twelfth International Conference on Input-Output Techniques*. New York.

Drejer, I., Kristensen, F.S., & Laursen, K. (10-11 October 1997). Studies of Clusters as a Basis for Industrial and Technology Policy in the Danish Economy. In *OECD Workshop on Cluster Analysis and Cluster Policies*. Amsterdam.

Eding, G.J., Oosterhaven, J., & Stelder, D. (1999). *Clusters en Linkages in Beeld, Een toepassing op de regio's Noord-Nederland, Groot-Amsterdam/NZKG en Groot-Rijnmond*. Stichting Ruimtelijke Economie Groningen, Groningen.

Ghosh, S. & Roy, J. (1998). Qualitative Input-Output Analysis of the Indian Economic Structure. *Economic Systems Research*, 10(3):263–274.

Gregori, T. & Schachter, G. (1999). Assessing Aggregate Structural Change. *Economic Systems Research*, 11(1):67–82.

Gurgul, H. & Majdosz, P. (2008). The Modified Diagonalization Method for Analysing Clusters within Economies. *Managing Global Transitions*, 6(1):53–73.

Hoen, A.R. (2002). Identifying Linkages with a Cluster-based Methodology. *Economic Systems Research*, 14(2).

Jilek, J. (1971). The Selection of the Most Important Coefficients. *Economic Bulletin for Europe*, 23:86–105.

Lahr, M.L. & Dietzenbacher, E. (Eds.) (2001). *Input-Output Analysis: Frontiers and Extensions*. Palgrave, New York.

Lamel, J., Richter, J., & Teufelsbauer, W. (1972). Patterns if Industrial Structure and Economic Development — triangularisation of Input-Output tables of ECE countries. *European Economic Review*, 3:47–63.

Lantner, R. (2001). Influence Graph Theory Applied to Structural Analysis. In *Input-Output Analysis: Frontiers and Extensions*, pp. 297–317. Palgrave.

Leicht, E. & Newman, M. (2008). Community Structure in Directed Networks. *Phys. Rev. Lett.*, 100.

Leontief, W.W. (1986[1963]). The Structure of Development. In *Input-Output Economics*, pp. 162–188. Oxford University Press. Second Edition.

Lequeux, F. (10-15 October, 2002). Thinking about Structural Decomposition in the Influence Graphs Theory. In *Proceedings of the 14th International Conference on Input-Output Techniques*. Montreal.

Maaß, M. (1980). *Die Reagibilität von Prognosen mittels Input-Output-Modellen auf Fehler im Datenmaterial*. Athenäum Verlag, Berlin.

McNerney, J. (2009). Network Properties of Economic Input-Output Networks. Technical report, International Institute for Applied Systems Analysis, Laxenburg, Austria. Interim Report.

Mesnard, L. (1995). A Note on Qualitative InputOutput Analysis. *Economic Systems Research*, 7(4):439–448.

——— (2001). On Boolean Topological Methods of Structural Analysis. In *Input-Output Analysis: Frontiers and Extensions*, pp. 268–279. Palgrave.

Morillas, A. & Díaz, B. (2008). Key Sectors, Industrial Clustering and Multivariate Outliers. *Economic Systems Research*, 20(1):57–73.

Newman, M. (2006). Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci. USA*, 103:8577–8582.

Oosterhaven, J., Eding, G.J., & Stelder, D. (2001). Clusters, Linkages and Interregional Spillovers: Methodology and Policy Implications for the Two Dutch Mainports and the Rural North. *Regional Studies*, 35(9):809–822.

Pasinetti, L.L. (1973). The Notion of Vertical Integration in Economic Analysis. *Metroeconomica*, 25:1–29.

——— (1988). Growing subsystems, vertically hyper-integrated sectors and the labour theory of value. *Cambridge Journal of Economics*, 12(1):125–34.

Piccardi, C. (2011). Finding Communities by Lumped Markov Chains. *PLoS ONE*, 6(11):1–13.

Rosenblatt, D. (1957). On Linear Models and the Graphs of Minkowski-Leontief Matrices. *Econometrica*, 25(2):325–338.

Schnabl, H. (1994). The Evolution of Production Structures, Analyzed by a Multi-layer Procedure. *Economic Systems Research*, 6(1):51–68.

——— (1995a). The ECA-method for Identifying Sensitive Reactions within an IO Context. *Economic Systems Research*, 15(4):495–504.

——— (1995b). The Subsystem-MFA: A Qualitative Method for Analyzing National Innovation Systems — The Case of Germany. *Economic Systems Research*, 7(4):383–396.

——— (2001). Structural Development of Germany, Japan and the USA, 1980-90: A Qualitative Analysis Using Minimal Flows Analysis (MFA). In *Input-Output Analysis: Frontiers and Extensions*, pp. 245–267. Palgrave.

Sherman, J. & Morrison, W.J. (1950). Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix. *Ann. Math. Statist.*, 21(1):124–127.
URL: *http://projecteuclid.org/euclid.aoms/1177729893*

Slater, P.B. (1977). The Determination of Groups of Functionally Integrated Industries in the United States Using a 1967 Interindustry Flow Table. *Empirical Economics*, 2(1):1–9.

——— (1978). The Network Structure of the United States Input-Output Table. *Empirical Economics*, 3(1):49–70.

Sraffa, P. (1960). *Production of Commodities by Means of Commodities*. Cambridge University Press, Cambridge.

Weber, C. & Schnabl, H. (1998). Environmentally Important Inter sectoral Flows: Insights from Main Contributions Identification and Minimal Flow Analysis. *Economic Systems Research*, 10(4):337–356.

Weil, R.L.J. (1968). The Decomposition of Economic Production Systems. *Econometrica*, 36(2):260–278.

Yan, C.S. & Ames, E. (1965). Economic Interrelatedness. *The Review of Economic Studies*, 32(4):299–310.