Uncovering Community Structures in Inter-industry Labor Mobility Networks in Argentina

Viktoriya Semeshenko[1] and Sergio Andrés De Raco[2]

[1]Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires, Argentina. CONICET-Universidad de Buenos Aires. Instituto Interdisciplinario de Economía Política de Buenos Aires. Buenos Aires, Argentina

[2]Universidad de Buenos Aires. Facultad de Ciencias Económicas. Buenos Aires, Argentina. Universidad de Buenos Aires. Instituto Interdisciplinario de Economía Política de Buenos Aires. Buenos Aires, Argentina

Abstract

In this paper we systematically explore high granularity economic activity data of labor flows using a network filtering reduction technique, focusing on a meso-structure analysis of relevant groups of economic activities. Using administrative data of interindustry labor flows for 1996-2020, we built the networks, extracted their representative backbones and applied a community detection algorithm. The results unmask inter-industry connectivity with persistent structures and well-identified communities organized in a core-periphery structure.

Keywords: Labor Mobility; Network Analysis; Disaggregated Data; Disparity Filter; Communities

1 Introduction

Aggregate sectoral economic activity information is usually analyzed by economists on a reduced set of categories (at sectoral letter and/or two digits), familiar to analysts and public policy decision makers, which can provide comprehensible knowledge in terms of traditional concepts, and be handled by also traditional methods. However, often at these levels of detail the analysis is limited, and a restricted analysis of selected subset of productive sectors of interest with more disaggregated data is done additionally on-demand.

The data from administrative records on labor identify the activity of employers at a high level of detail, which offers a big challenge to work with on systematic basis and is left unused. In particular, analyzing sectoral data at higher levels of detail, in this case at four digits, presents challenges regarding the appropriate methods of analysis as well as the posterior interpretation of the productive sectors and the economic and public policy relevance of the information at this level.

Network framework is most suitable to use for labor mobility analysis, because these data can be mapped into a network of bipartite interactions between workers and industries[1], and thus, the structural characteristics of these interactions can be revealed through the employment between industries at such levels of details, as well as the extent to which they persist over time. Additionally, relevant groupings of exchange flows between industries can be identified and the information of the productive framework can be extracted.

In this paper, we extract information from disaggregated data of intersectoral labor flows at a practically unexplored level of detail, and make a systematic analysis of economic activities at this high level of granularity. In previous works we have shown that the intersectoral relationships which arise from interindustrial employment transitions reveal a history of high and persistent connectivity that evidences a dense structure of interactions, organized in a single connected component with short paths, exhibiting the "small world" properties [1], [2]. Sectoral organization structures of the core and

[1]Hereafter we will use indifferently "economic activities"/"sectors"/"industries".

periphery have been identified, which provide information on functional aspects of the interactions between sectors mediated by employment in line with the productive organization, which makes it possible to detect two large nuclei of high interactions and other subgroups of relatively stable peripheral sectors over time. In terms of the evolution of the system, there are changes in the organization of the interconnection structure in the medium term in connectivity ”regimes” of labor mobility, in correlation with aggregate economic activity [3].

Doing analysis at this high level of desaggregation implies working with a large number of sectors increasing the level of complexity that another analysis method needs to extract knowledge and understand the operation of the system at that level of granularity. The labor flow networks have shown to be dense, in contrast with other social and financial networks that are sparse. These features together with fact that these networks are directed and weighted, make the extraction of truly relevant connections forming the network backbone a very difficult task. Many different network reducing techniques exists, each having its pros and cons [4], [5].

Traditionally, in these type of flow networks economists tend to use global thresholds to focus on ”representative” flows. This approach effectively reduces the complexity of the data, but may oversee the information contained in some small flows that in relative terms could be of high significance. In particular, it is plausible to think that new flow patterns may begin small and thus, if they were under the cutoff threshold and by so excluded, they may pass unnoticed. Also, small flows may have special significance under the public policy lens.

In this paper, we are using the Disparity Filter for weighted networks [4], a method that operates at all scales defined by the weighted network structure, and it is able to extract the network backbone by considering the relevant edges keeping at the same time almost all of the weights (flows). Through another type of analysis more focused on identifying groups with greater interaction between their members, community detection algorithms, we seek to identify groups of sectors among which are observed greater number of job transitions than between members of the same group and the rest of the network. In other words, we are looking to meso-structures, by using Leiden algorithm [6]. Analysing communities in the different periods makes it possible to distinguish groups that associate sectors according to the classification code (v.g.: activities with a different letter) although close in terms of productive activities in the broadest sense of productive organization. We are interested to know its persistence and its evolution over time in order to identify those sectors that determine the main groupings and sectors that can be associated with different communities over time.

As the main results we highlight the following findings: a) the backbones of labor flow networks are shown to be representative and informative; and, b) there is a persistent structure of communities with two well identified cores.

The paper is organized as follows: the data and methods used for the analysis are presented in section 2. The results are presented in section 3. In section 4, we present conclusions and we add supplemental information in the appendix A.

2 Data and Methods

Data. We used data from the Argentinian Integrated Pension System (SIPA), provided by the Ministry of Labor, Employment and Social Security (MTEySS), at four digits of detail of the ISIC Rev 3 of industrial activity classification code, for the periods between 1996 and 2020. The MTEySS provides data on private sector employment received from affidavits submitted by employers to the AFIP[2]. These administrative records express the relationship between the employer and its payroll and contain information on the economic and social characteristics of both actors. This information is processed and interannual transition matrices are built, recording the individual switches in employment between the different economic activities.

In this work, only the flows of inter-industrial employment transactions are used, which represent the aggregate of individuals who are employed in one period and in next one are employed in another company belonging to a different activity sector. So far, the

[2]The Administración Federal de Ingresos Públicos (Spanish: Federal Administration of Public Income), usually shortened as AFIP.

set of productive activities includes approximately 300 sectors or branches of activity at 4 digits organized in 14 sections (i.e.: letters).

Methods. We use two methods for the analysis of the extracted networks: the disparity filter for directed weighted networks [4], and the Leiden algorithm to uncover the meso-structures and identify the principal communities structures [6].

Disparity Filter. The networks of labor flow that emerge from these data are complex weighted networks, present dense structures that are not sparse. It has been shown previously, that the structure presents a single large connected component, of a reduced diameter (3 steps) and exhibits small world properties with core-periphery characteristics [1, 2]. The structure of the original graph of flows has been useful in the representation of the interactions at the industry level. However, analysing dense networks like these makes it difficult to extract the relevant information from these connections without further looking for a reduced representation of these networks preserving the key features that highlight the relevant connections and hierarchies. The idea here is to use the filtering methodology as a network reduction algorithm in order to extract the backbone structure of relevant connections of directed weighted networks, and thus, to obtain sparse graphs for further analysis.

The Disparity Filter (DF) algorithm is a network reduction technique to identify the 'backbone' structure of a weighted network without destroying its multi-scale nature. The algorithm operates like a statistical significance test for the weighted edges of a network: given a distribution of edge weights, the algorithm singles out those forming the 'backbone' of the network, using the equivalent of a statistical significance threshold to identify them. In Figure 1 we plot the histogram of weights of the original network versus the weights that have been retained by the backbone after the filtering. The method locally identifies statistically relevant weight (links) heterogeneities, filters out the backbone of dominant connections that are not compatible with the significance level α and preserves the structural properties that are important.

We explored the effects of selecting different values for α with respect to fundamental

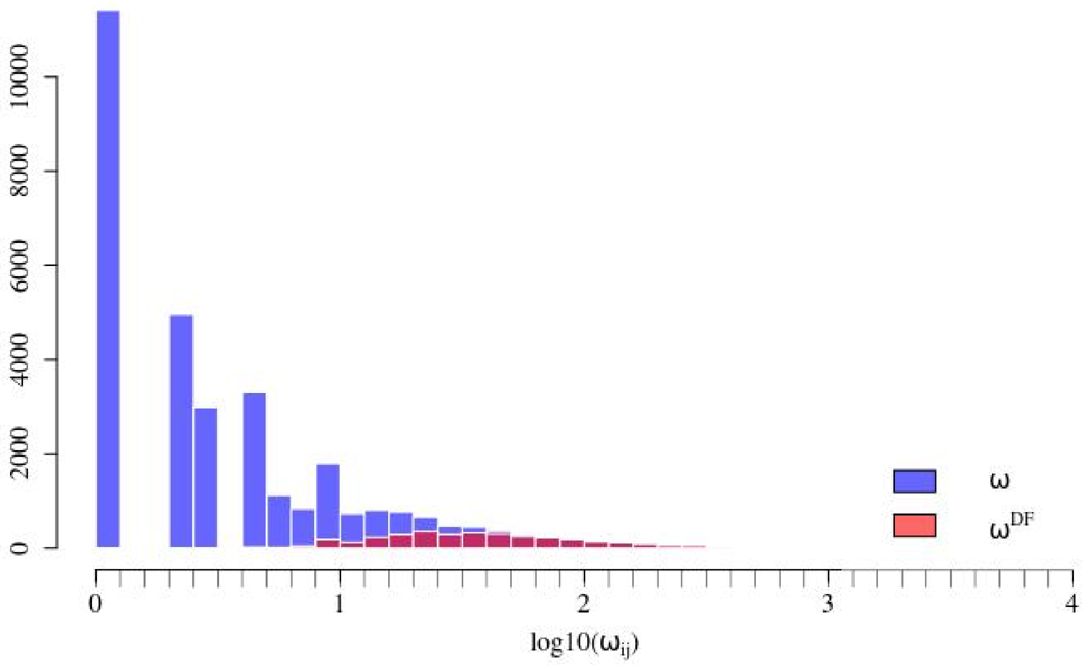

Figure 1: The histogram of weights (flows) of the original network and of the backbone ($\alpha = 0.25$), for period 2006. As can be seen, the weights ω, a large proportion of connections share small flows. While global threshold (artificial cutoff) would drastically remove all the information (flows) below the cutoff scale, the DF technique used here preserves the links with weights which are statistically significant. Here, backbone retains transitions of 4 (lower bound) and more individuals, see ω^{DF}.

elements of the network as the shares of the number of nodes, N_B/N, the number of edges, E_B/E, and the accumulated weights, W_B/W), and the average (local) clustering coefficient, $\langle C_{Df} \rangle$), shown in Figure 2. We selected $\alpha = 0.25$ to carry our analysis, such that it guarantees the networks with almost all the nodes, with few nodes left as isolates, and despite the fact that the DF reduces the number of edges significantly (almost to 24%), it keeps at the same time almost all the weights that interest us. The distribution of weights does not lose its statistical significance. Looking at the topological properties of the resulting networks, clustering coefficient, we observe that at this value of α it is still high and remains so until the filtering becomes more restrictive (smaller values of α) and starts to decrease.

The backbone extracted by the DF is shown on Figure 3 for 2006, which is a year that shows a predominant network structure (see [1]). The resulting graph has 0.095 density of connections and it is sparse. The network consists of a single component

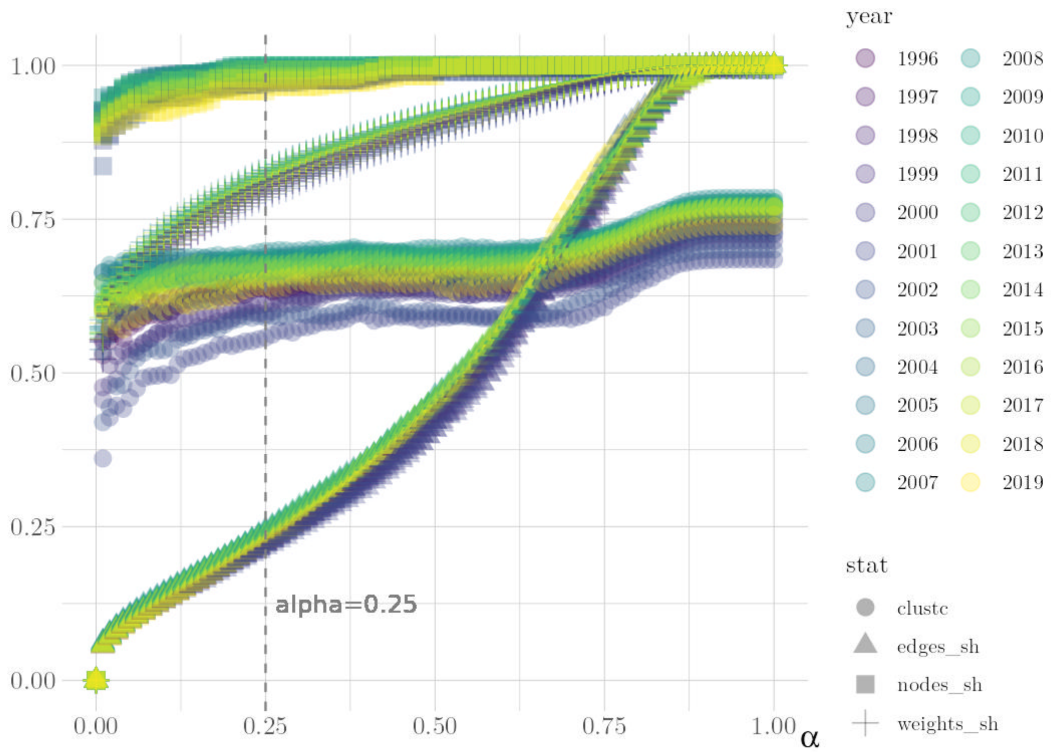

Figure 2: Shares of Nodes kept in the backbone N_B/N, Edges E_B/E, total Weights W_B/W, and the average (local) clustering coefficient $\langle C_{Df} \rangle$) retained by the filter, for different values of significance level α, for all 24 years. The value $\alpha = 0.25$ has been selected for further analysis. For this value of α, we get the reduction of the connections to 24% preserving almost all the nodes (with 4 isolates on average left), more than 75% of weights with all of the respecting edges included in the tail of ω, see Fig. 1, are included in the extracted backbone.

preserving almost all the nodes, 285 out of 287 nodes (only two isolates are left), 76693 edges and 309922 flows are preserved. Taking into account Figure 2, we can conclude that backbones obtained for this values of α for all periods are good for considering for further analysis because they retain a large proportion of nodes and weights, similar clustering of the original network, and with a very small number of connections compared with the original network.

We can get better details of the isolates left by the DF in each period, in Figure 4. We observe that there are some sectors that tend to be outside the backbone throughout the period (see for example, the sectors in yellow color, which is repeated almost 14 out of the 24 periods analysed. The observed evolution of isolates can be directly linked to the labor transitions dynamics in these periods, in particular, to the periods of relative macroeconomic stability with structural changes in labor flows: 1996-2002 (a declining period associated with the last part of the validity of the Plan of Convertibility and the

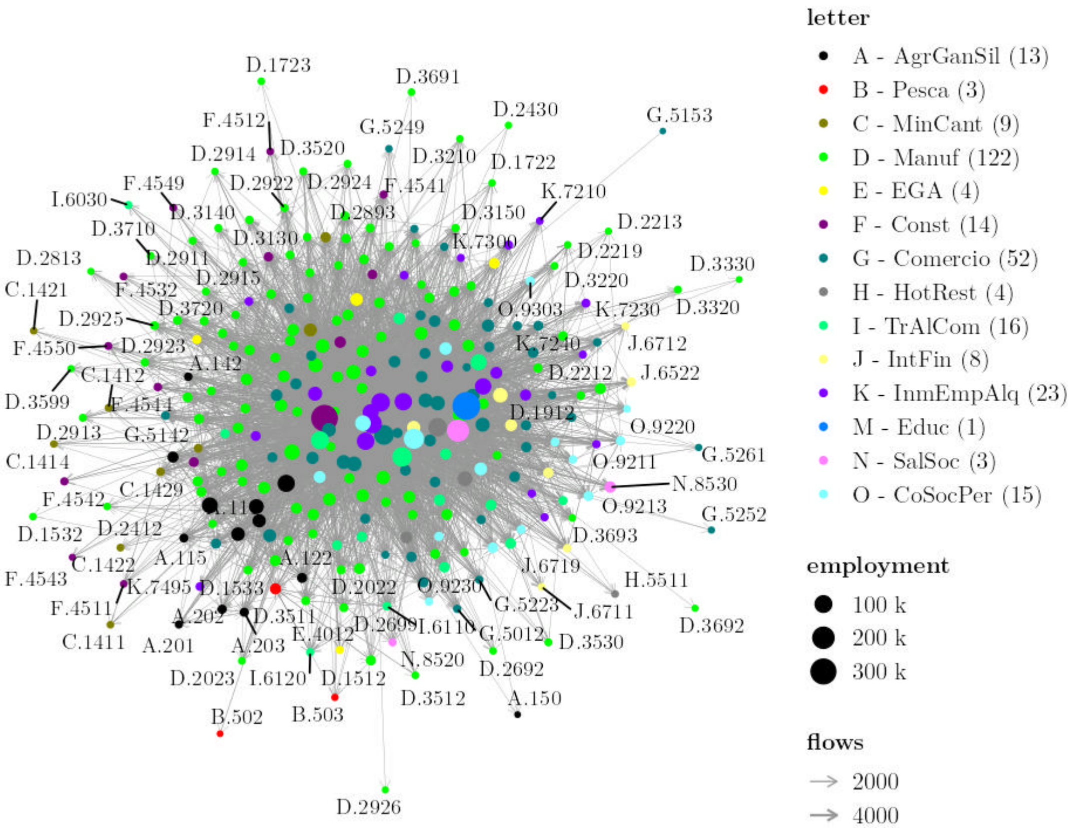

Figure 3: The DF backbone extracted for α = 0.25, for 2006. This backbone includes 24% of the links of the original network while keeping 83% of the weights. The graph presents a large connected component with 285 nodes of 287 in the unfiltered network. Density 0.095 with 76693 edges and 309922 flows. Clustering Coefficient = 0.341. Diameter = 0.343. Assortativity = -0.343.

subsequent crisis); 2002-2011 (recovery and post-crisis growth); 2012-2017 (stagnation); and 2018-2020 (one of sharp decline associated with the change of government and the beginning of the COVID-19 pandemic) [7].

Leiden algorithm. Another relevant aspect in the analysis of interindustry flow networks is the need to inquire about groups of industries that exchange employment more frequently with each other than with the rest of the industries. This means to identify groups of sectors among which are observed greater number of job transitions than between members of the same group and the rest of the network. This refers to the exploration of the meso-structure in the network, identifying clusters and detecting communities. On the other hand, it is also useful to know its persistence and its evolution over time in order to identify those sectors that determine the main groupings and sectors associated with different communities over time.

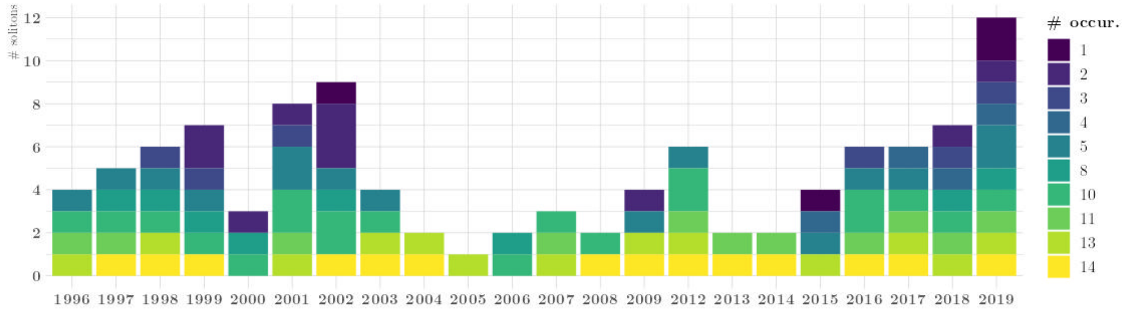

Figure 4: The occurrence of isolates left after filtering by DF in each period, for $\alpha = 0.25$. Colors for different accumulated occurrences. Details of these sectors at 4-digits and letter level, together with their employment reference, are available in the Appendix A.

We are using a node partitioning, the Leiden algorithm [6], applied to directed weighted flow networks. This algorithm improves, extends and refines the Louvain method [8] based on the three-stage optimization of modularity, a quality measure that calculates the density of ties within communities relative to the density of ties between communities. To guarantee convergence and stability of community partitions, we run $n = 1000$ iterations of the algorithm on each backbone.

3 Results

After having extracted 24 backbones (directed and weighted networks) and run the community detection algorithm, we obtained consistent and relatively stable community structures depicted in Figure 5 (top). The main structure decompose these networks in, at most, eleven communities of economic activity sectors sorted in decreasing order of stability in time. A greater number of communities appear in years where the declining macroeconomic situation affected the performance of the labor market, for example during the period 1998-2004 (change of the Convertibility regime), and 2015 and 2019 (years in which the government shifted to opposition parties conduction). By definition, the communities of the filtered networks (backbones) exclude isolated nodes which, in turn, present a positive relationship with the economic cycle similar to that described for the communities (see Figure 4). Interestingly, the first two ($C1 - C2$) group between 40-65% of sectors analysed and capture more than 40% of total employment on average, while the first four communities ($C1 - C4$) include between 66-97% of sectors and accumulate 77% of total employment on average for almost all years in the period of analysis (see Figure 5, bottom).

To assess the *persistence* of the extracted communities we summarise each sector community label assignment during all periods in a matrix of sectors and communities. We analyse these matrix of counts using a heatmap sorted with hierarchical clustering algorithms with complete linkage, depicted in Figure 6, cutting the tree into $k = 6$ groups of sectors naturally showing more persistence in defined community labels, which we hereafter refer to as $S1 - S6$. We can appreciate an interesting structure capturing a core-periphery sectoral organization. There are two cores involving communities $C1$ and $C2$, and a periphery composed by communities $C3$ and $C4$. Communities $C5$ and $C6$ are less persistent and $C7$ to $C11$ tend to be small, as they are a sort of "residual" groups in the community detection phase, and show poor identification and impermanence for sectors members. Leveraging on the proposed clustering classification $S1 - S6$, Figure 7 shows the column dendrogram in Figure 6 and an alternative representation of the tree with a reference to the sectors included.

The first and more relevant community $C1$, the "industrial" core capturing 18% of total employment on average, is the most stable one (in the sense of sector members with high persistence). As can be seen in Figure 6, $C1$ is composed mainly by activities belonging to cluster $S1^*$, which include mining and extractive activities, manufacturing of metals, machinery and tools, water and electricity infrastructure and services, construction, and associated transport and commerce services, and cluster $S3^*$, including manufacturing of motor vehicles, furniture, plastic, paper, and other machinery. The $C2$ community, the "services" core, is composed mainly by activities belonging to cluster $S2^*$, including legal, merchandising, news and related services, telecommunication infrastructure and services, financial and computing services (v.g.: Information Technology services), publication and printing services, radio and television services, and travel services, and cluster $S6^*$, including insurance, education, health, social work, aerial transportation, motor vehicles sale services, as well as pharmaceutical manufacturing and associated manufacturing and services. As for $C3$ and $C4$ communities can be classified as a "close periphery", they include activities with relative persistence (8 of 24 periods) that come mainly from cluster $S4^*$ such as tourism, entertainment, retail, security, cleaning, real estate services, as well as agricultural and forestry sector activities, with associated services. Communities $C5$ and $C6$ can be categorised as a "less-close periphery" and include sectors with less persistence mainly from cluster $S5^*$, such as: textile, shoe and footwear manufacturing and associated commercialization and repair services, cargo and aquatic transportation, and phishing, and from cluster $S6^*$ optics and furniture manufacturing and associated services. Regarding communities $C7$ to $C11$, they do not have high persistent members and can be classified as a "distant periphery".

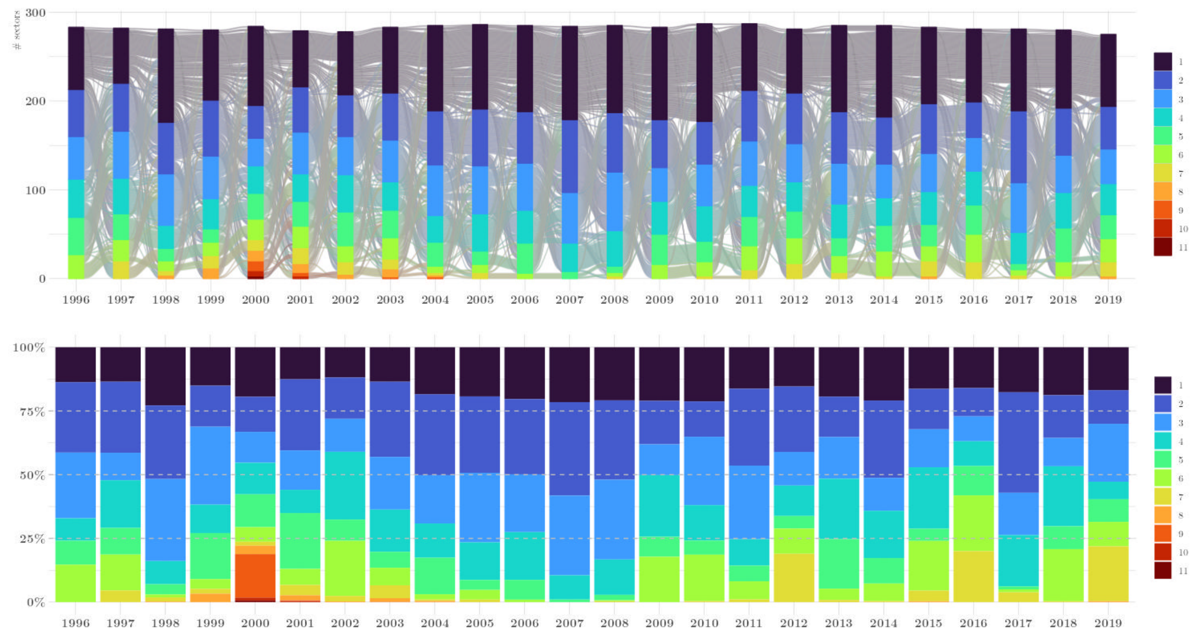

Figure 5: Communities. Color for each detected community. Top: Sankey diagram for flow dynamics on backbones of the 24 periods showing communities detected by Leiden algorithm after $n = 1000$ iterations (sectors on left axis). Bottom: Employment share of sectors by community membership (shares on left axis).

4 Discussions

In this paper we presented an approach for the systematic analysis of high granularity economic activity data applied to labor flows explored at an under-explored level of detail, including network filtering using significativity criteria and, most importantly, a meso-structure analysis of relevant groups of economic activities that brings new insights

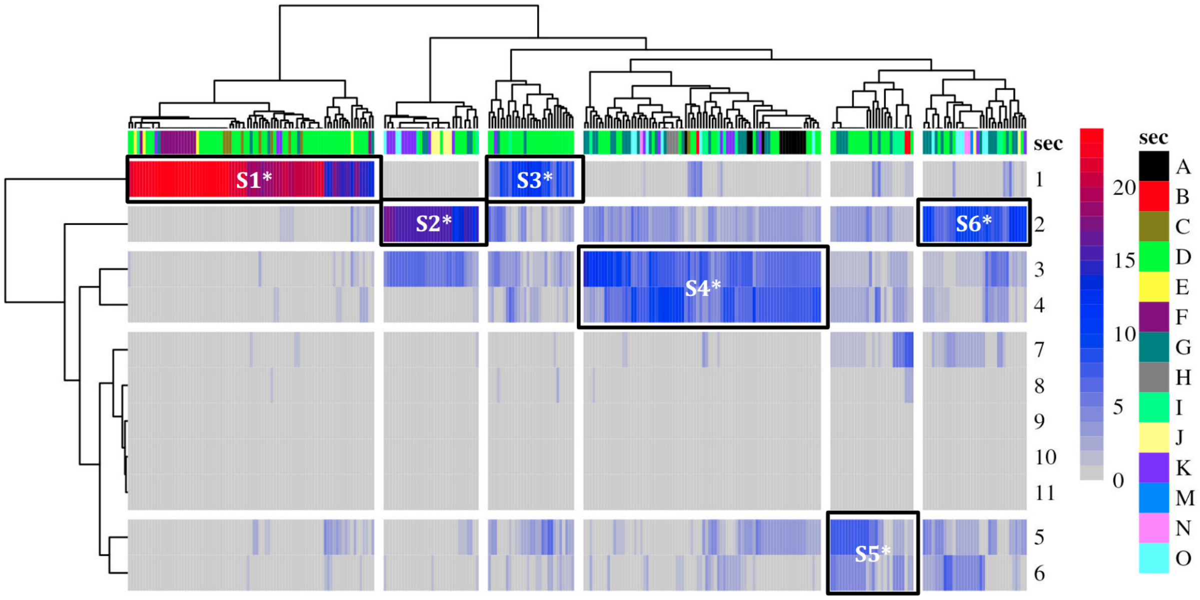

Figure 6: Communities persistence. Intensity scale red-blue-gray (max. 24, min. 0 periods). Sectors in columns, communities in rows. "SEC" color annotation refers to the letter ISIC Rev 3 classification code of the 4 digits sectors used. The heatmap is sorted with hierarchical clustering algorithms with complete linkage, cutting the tree into 6 groups of sectors naturally showing more persistence in defined community labels (rows), each group referred as $S1 - S6$. Selected sectors with major contribution marked with $S1^* - S6^*$.

on the productive framework analysis. The results of this analysis reveal inter-industry connectivity with persistent structures and well-identified communities.

Using administrative data of interindustry labor flows for 1996-2020, we built the networks and extracted their representative backbones and applied a community detection algorithm. We found persistent communities organized in a core-periphery structure. We identified two persistent cores, the first one comprised mostly of industrial activities and associated services and the second one comprised of specialized services and some other manufacturing activities. Additionally, we identified four communities functioning as a "close periphery" of the cores, less persistent and grouping diverse subsets of activities, and another kind of communities that capture the less connected activities or "distant periphery".

These findings are promising in two ways. Firstly, the results regarding the persistence of connectivity structure of the productive activities through labor flows bring new information to contrast with findings at an aggregated levels, in labor economics as well as in macroeconomics and real sector analysis studies. Secondly, the methodological contribution in terms of providing a systematic analysis framework for disaggregated economic

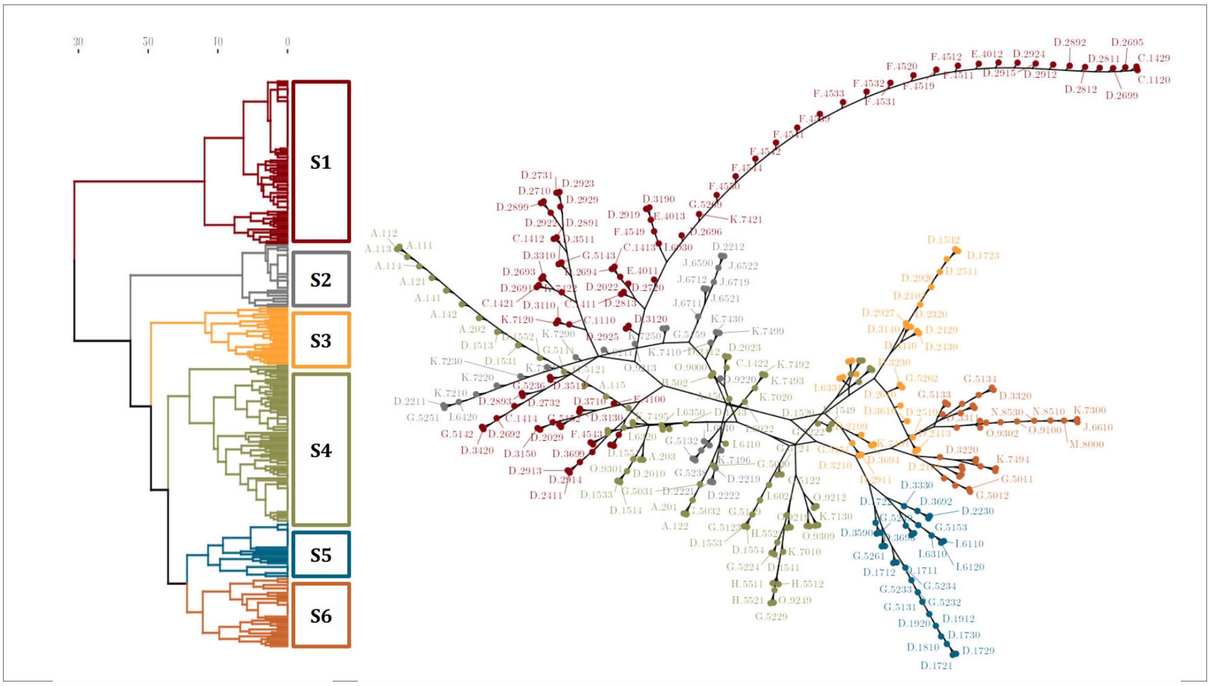

Figure 7: Sector persistence hierarchical clustering. Colors by sector clusters. Detail of column clustering tree in Figure 6 with an alternative representation showing sectors identification codes.

data. We believe the potential of these methods to analyze economic data is huge.

The complex network approach is appropriate and useful for the analysis of data at a very high level of detail, in particular labor flow data. It allows to explore the data in a persistent and systematic way, and to inquire on dimensions and scope that traditional analysis does not reach or match.

References

[1] S. A. De Raco and V. Semeshenko, "Labor mobility and industrial space in Argentina," *Journal of Dynamics & Games*, vol. 6, no. 2, p. 107, 2019.

[2] S. A. De Raco and V. Semeshenko, "The network structure of inter-industry labor mobility in Argentina," in *6th Regulating for Decent Work Conference*, (Geneva), ILO, 2019.

[3] A. De Raco, Sergio and V. Semeshenko, "Identificación de diferencias y similitudes estructurales de las redes interindustriales del empleo de argentina," *to appear in Proceedings 52 JAIIO-AGRANDA*, 2023.

[4] A. M. Serrano, M. Boguña, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *PNAS*, vol. 106, pp. 6483–6488, 2009.

[5] I. S, L. R, K. N, M. R, I. M, and A. U., "Coarse-graining and self-dissimilarity of complex networks," *Phys. Rev E*, vol. 71, p. 016127, 2005.

[6] V. Traag, L. Waltman, and N. van Eck, "From louvain to leiden: guaranteeing well-connected communities," *Sci Rep*, vol. 9, p. 5233, 2019.

[7] V. Semeshenko and S. A. De Raco, "Analysis of the evolution of labor market flows in Argentina," *Proceedings 50 JAIIO-AGRANDA*, pp. 20–24, 2021.

[8] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. P10008, p. 12, 2008.

A Appendix

Information of Isolates			
Letter	Node	Economics Activity at 4 digits	Employment
A	150	Hunting, trapping and game propagation including related service activities	85
B	502	Fish hatchery, fish farm and other aquatic fruit operation (aquaculture)	97
C	1411	Extraction of ornamental rocks	657
D	1532	Manufacture of starches and starch products	666
D	1722	Manufacture of carpets and rugs	599
D	1723	Manufacture of cordage, rope, twine and netting	730
D	2213	Publishing of recorded media	291
D	2230	Reproduction of recorded media	82
D	2430	Manufacture of man-made fibres	376
D	2692	Manufacture of refractory ceramic products	713
D	2813	Manufacture of steam generators, except central heating hot water boilers	242
D	2914	Manufacture of ovens, furnaces and furnace burners	428
D	2926	Manufacture of machinery for textile, apparel and leather production	166
D	2927	Manufacture of weapons and ammunition	268
D	3330	Manufacture of watches and clocks	136
D	3692	Manufacture of musical instruments	200
F	4543	Glazing on site	173
G	5153	Wholesale of vehicles, equipment and machines for rail, air and navigation transport	70
G	5252	Retail sale via stalls and markets	294
G	5261	Shoe repair and leather goods	344
J	6620	Retirement and Pension Fund Administration (AFJP)	9885
		SUBTOTAL	16502

Table 1: Disaggregated information on the isolates that have been persistent during 24 periods, after have been extracted by disparity filter, for $\alpha = 0.25$. In total of 22 isolates, at the letter level and 4 digits with the total employment absorbed by each economic activity.