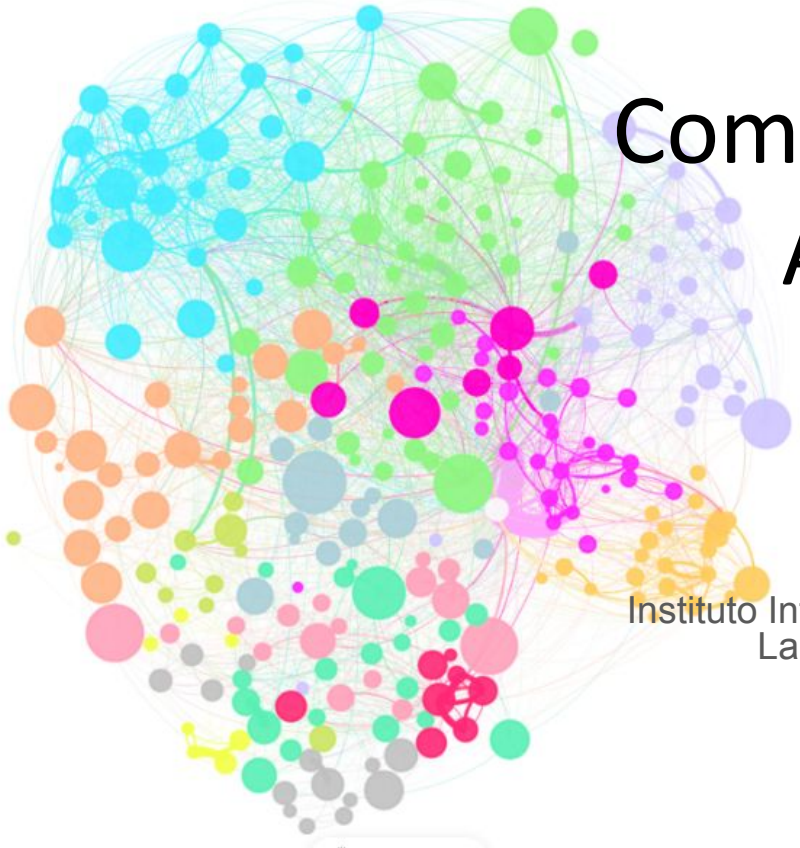
Instituto Interdisciplinario de Economía Política de Buenos Aires (IIEP-BAIRES)

NetLab
Laboratorio de Redes y Sistemas Complejos

CONICET UBA
I I E P

ICTP SAIFR



Complex Economic Networks: Analysis, Applications and Data

Viktoriya Semeshenko

Instituto Interdisciplinario de Economía Política (IIEP-UBA-CONICET)
Laboratorio de Redes y Sistemas Complejas (**Netlab**)

vika.semeshenko@gmail.com
@vsemesh @Netlab_IIEP

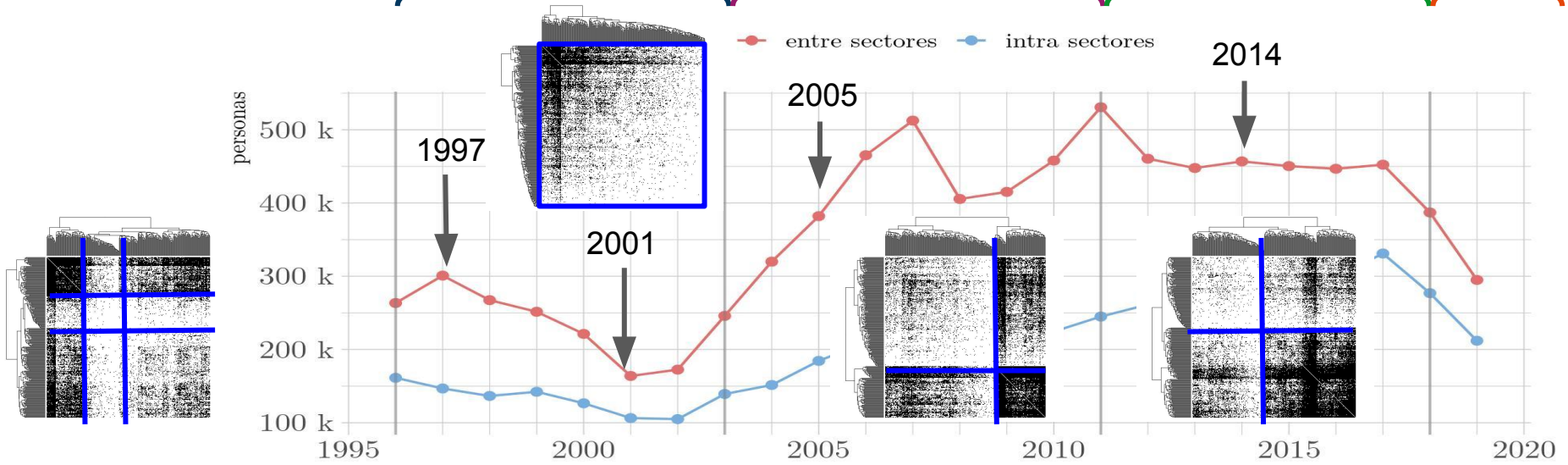
Class 4: Distances and the Networks. Communities in Networks

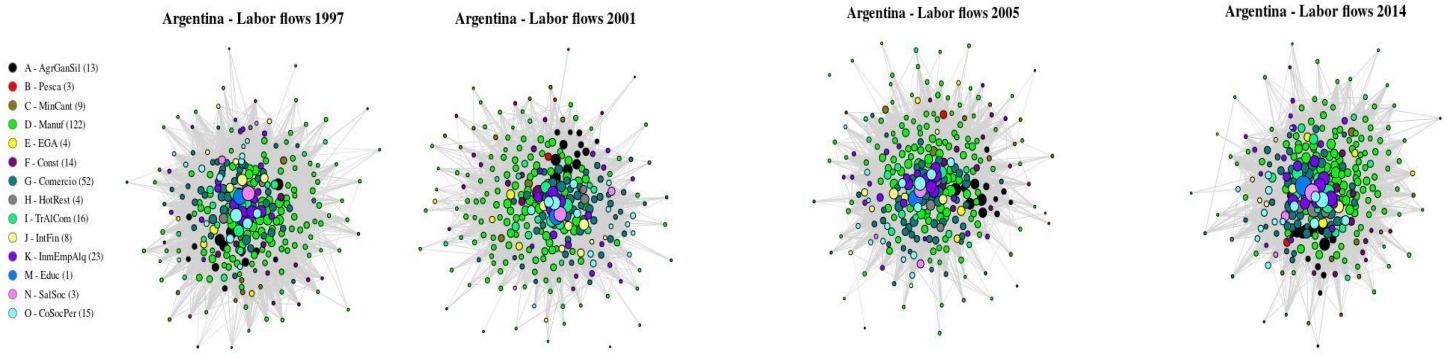
Sources used in the course:

- Barabási: Albert-László Barabási. Network Science
- Menczer et al: Filippo Menczer, Santo Fortunato, Clayton A. Davis, A First Course in Network Science. Cambridge University Press 2020

What kind of analysis can be done to compare redes?

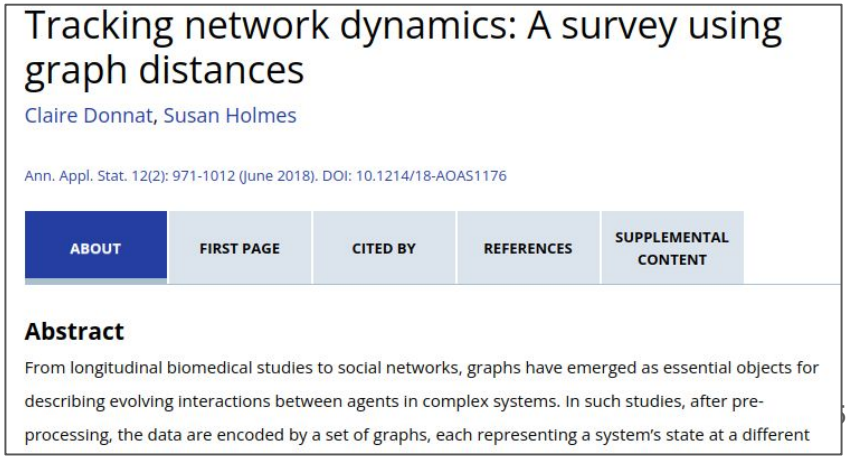
Convertibility ◉ 9602 | **post-Convertibility** ◉ 0310 | **local stagnation** ◉ 1117 | **Recession** ◉ 1820





Donnat and Holmes 2018

1. Developed an approach to compare networks of nodes aligned at different moments in time
2. Using the fundamental matrices of the graphs associated with these networks, different metrics and dissimilarities are used for their comparison through the analysis at different scales of their topological structures:
   - *local*
   - *intermediate (meso)*
   - *global*



Tracking network dynamics: A survey using graph distances

Claire Donnat, Susan Holmes

Ann. Appl. Stat. 12(2): 971-1012 (June 2018). DOI: 10.1214/18-AOAS1176

ABOUT | FIRST PAGE | CITED BY | REFERENCES | SUPPLEMENTAL CONTENT

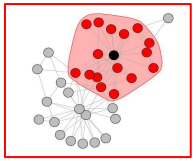
Abstract

From longitudinal biomedical studies to social networks, graphs have emerged as essential objects for describing evolving interactions between agents in complex systems. In such studies, after pre-processing, the data are encoded by a set of graphs, each representing a system's state at a different

Distances that allow networks to be evaluated at different scales of connectivity

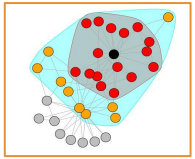
- Growing Connectivity Analysis
  - **Jaccard**: total change of connections and/ nodes (N1)
  - **Polinomial 2**: expanded neighborhood with interactions (N2)
  - **Espectral**: propagation of information from each node to the rest of the graph (frequencies)
- What do these distances allow to discover?
  - Interconnection stability regimes (*d* min)
  - Anomalies and transitory periods (*d* max)

Jaccard distance



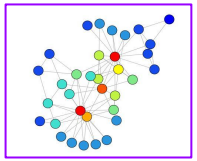
- Allows to identify changes in the networks at the level of local interconnections of each node and is appropriate to evaluate the rate of rearrangement of links
- The coefficient can be interpreted as the proportion of links, eliminated or incorporated, with respect to the total number of links that appear in both networks
- If its value is close to 1, it indicates a complete remodeling of the graph structure between two periods

Polinomial distance



- Based on the use of the power polynomial of the adjacency matrix that is related to the local topology of the graph through the coefficients that represent the paths from node *i* to node *j* through k steps
- It allows to quantify differences between two graphs of the connectivity structure that each node presents with the neighborhood of nodes with which it has direct exchanges and of the connectivity of these with the rest of the network

Spectral distance



- Used to characterize the state of the graph through its eigenvalues allowing to detect changes in global and intermediate structures
- But why should eigenvalues characterize the state of a graph better?
- The eigenvalues of a graph characterize its topological structure, and in particular the way that information localized at a particular node can be propagated over the graph. As such, they are related to the stability of the complex system that the graph represents.
- When carrying out the spectral analysis (eigen-analysis) on the fundamental matrices, information is extracted on the stability of the dynamics of the complex system that is represented, and on its evolution by monitoring the changes in the eigenvalues of these matrices.
- The spectral distance is defined as the distance between functions of the spectrum of the eigenvalues of the matrices of two graphs

Distances

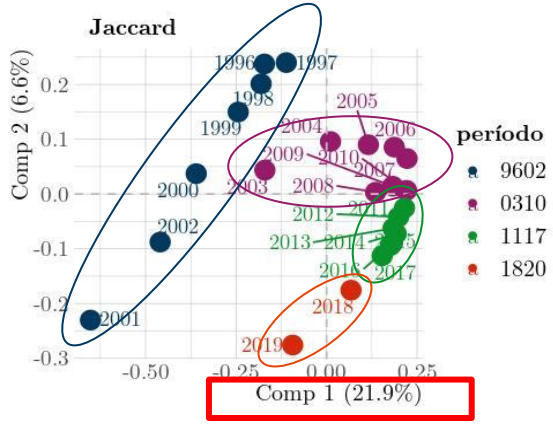
- Allow to identify
  - Stability regimes
  - Anomalies and Transitions
- Utility for employment transitions
  - Characterization and monitoring of intersectoral connectivity
  - Basis for identifying “connectivity regimes”

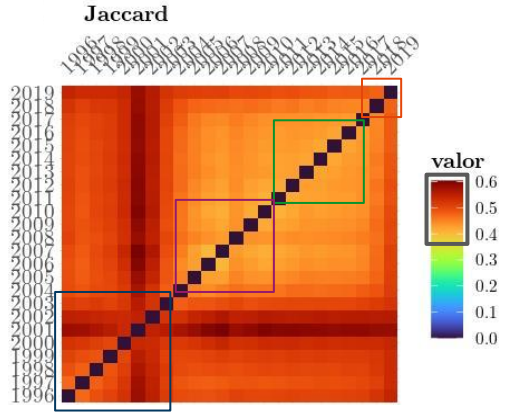
- In order to evaluate the results of the distances, the statistical method of metric multidimensional scaling (MSD) is used. It is a classic nonlinear dimensionality reduction technique similar to PCA, based on the decomposition of eigenvalues of a distance matrix, which allows mapping the proximity between different networks in the components that capture the greatest variability of the analyzed matrix.

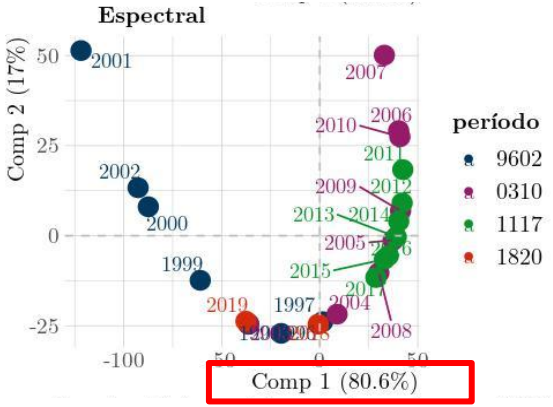
Two large periods are observed, between 1996-2000 (convertibility), and 2004-2018, which present relative similarity within them in consecutive years (orange), and differ between periods (orange dark), separated by years of crisis 2001-2003 and 2019, where large differences appear (almost red).

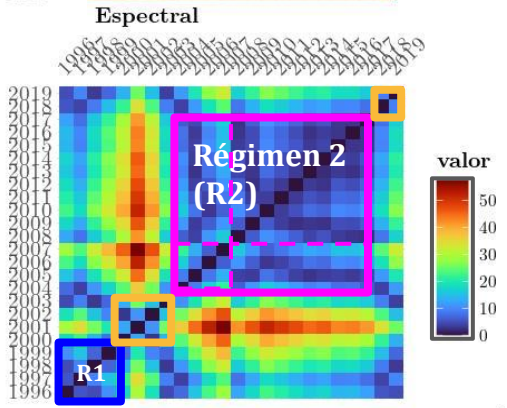
The distance matrix is mapped into the first two components that capture the greatest variability of this matrix, which between the two accumulate 28.5% of the total variability.
It is observed how it orders the years of different groups and separates the transition years.

Spectral: Two large groups of periods (1996-2005) and (2007-2018) or labor mobility regimes are identified, while 2006 is identified as a period different from all the others and 2019 as different from the years of the post-convertibility and more similar to the labor mobility regime of convertibility.





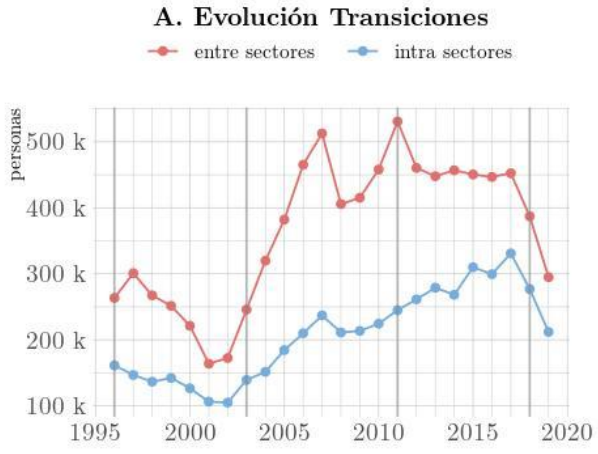


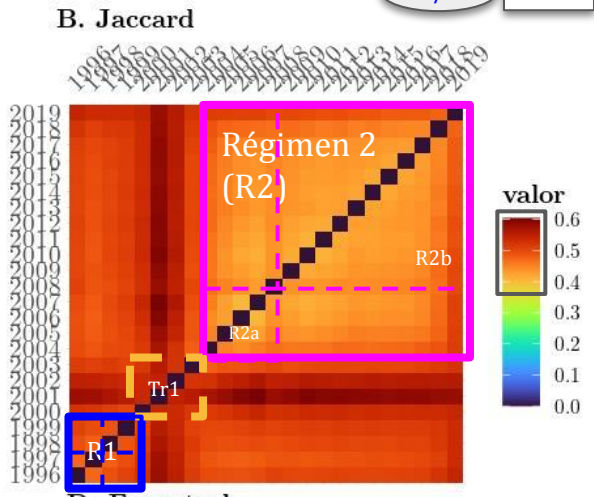


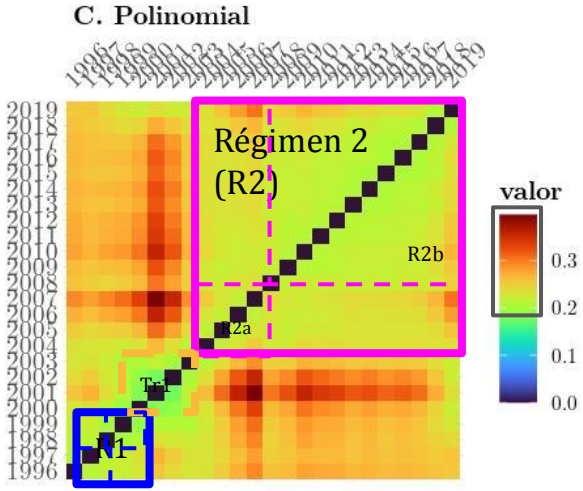
Fuente: Elaboración propia en base a SIPA

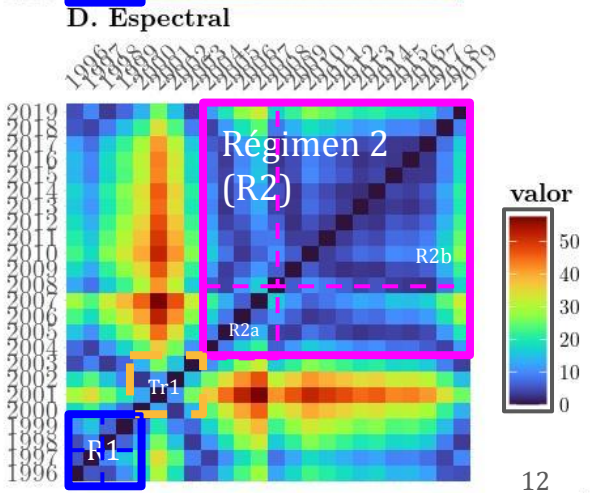
Polynomial: A sub-regime appears with some variability between 2004-2007 and a more stable one between 2008-2019, and the maximum distance is observed between 2001 and 2007.

In the MDS projection, the first two components accumulate 48.7% of the total variability of the matrix.
Again, the Convertibility period including the transition years 2003-2004 and 2019 are ordered in its negative values, and to the right in the positive values the remaining years.



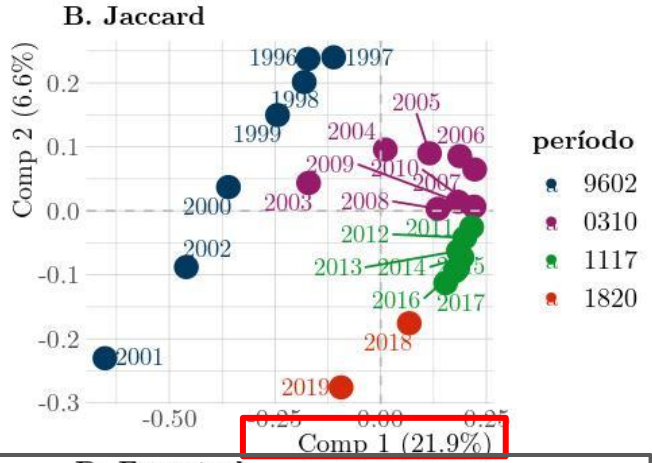


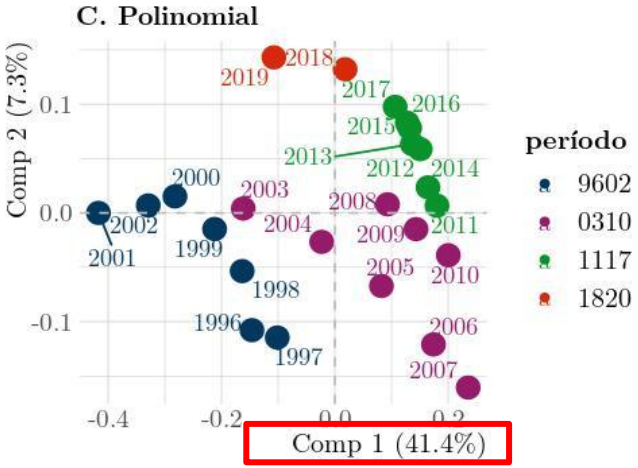


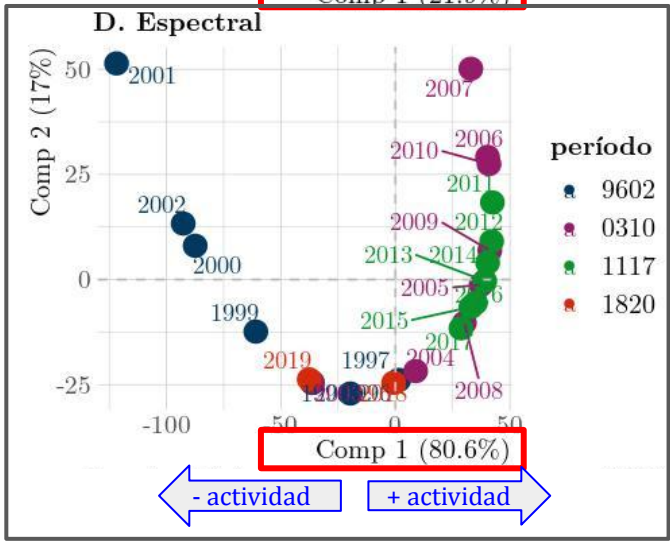


Fuente: Elaboración propia en base a SIPA

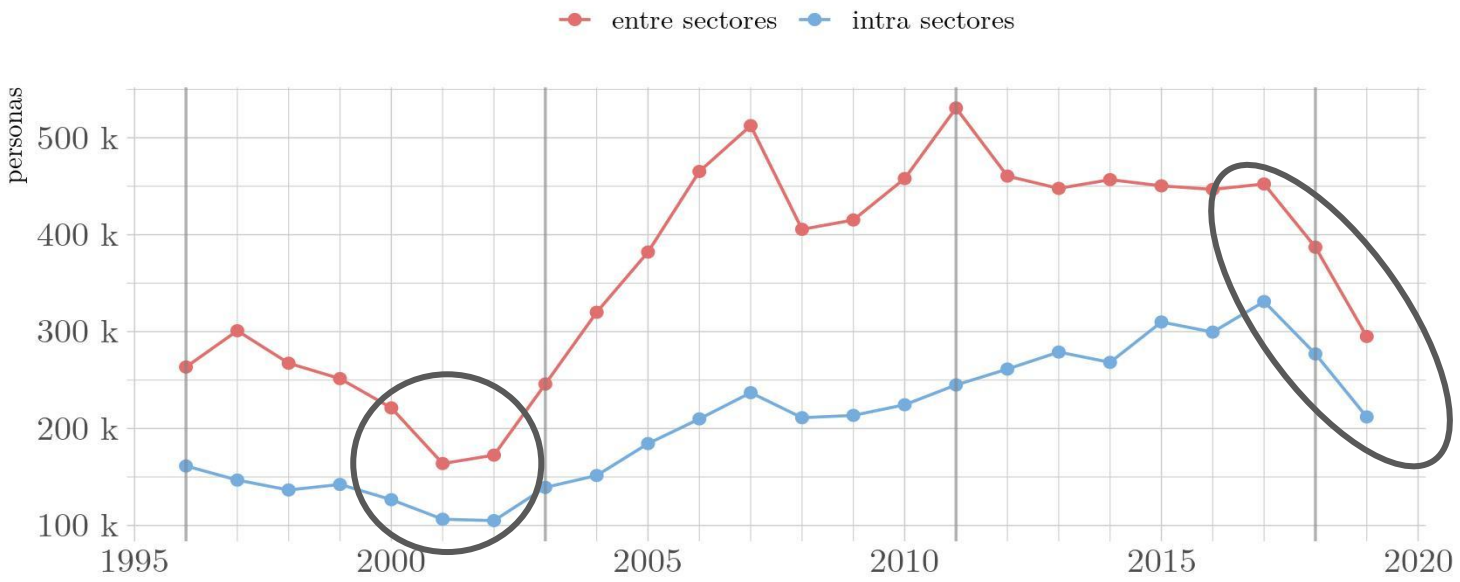
Using the distances, it is possible to confirm the differences in the structures of the employment flow networks at different levels of analysis of their connectivity, and the periods appear to represent different regimes of employment mobility.







Result of the distances analysis



Communities

- A relevant aspect when analyzing employment flows is the need to inquire about groups of industries that exchange jobs more frequently with each other than with the rest of the industries. In terms of networks this task refers to the exploration of the meso-structure in the network under cluster analysis.

- Within the families of ***community*** detection algorithms, it is possible to classify community detection methods into two categories, depending on whether nodes are assigned to disjoint groups (partitions), or nodes are allowed to be assigned to multiple membership groups (overlapped communities).

- On the other hand, it is relevant to distinguish whether the analysis is performed on directed or undirected graphs, given that the technique differs.

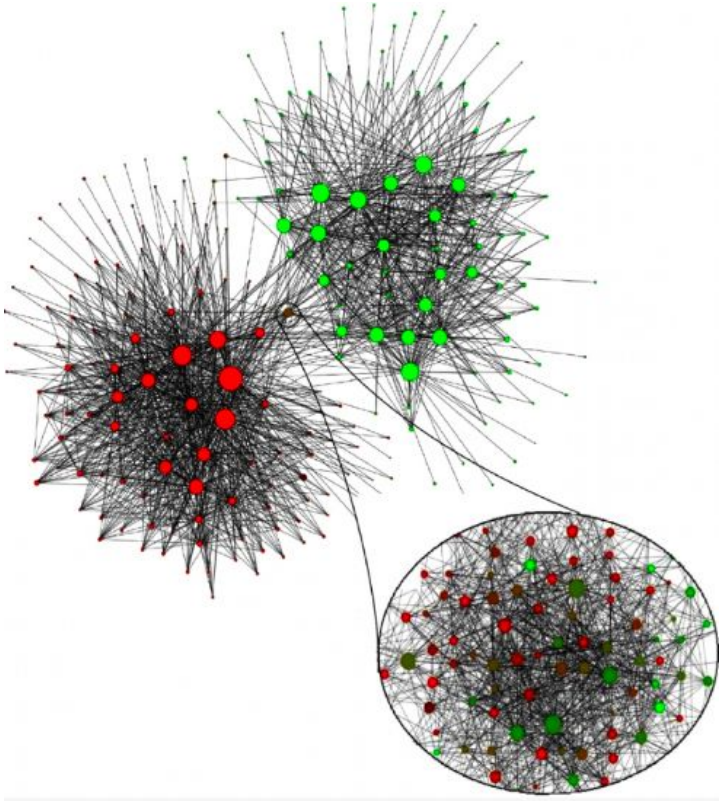
Ref: Santo Fortunato, Community detection in graphs

Communities: some intro from Social Networks

Belgium appears to be the model bicultural society: 59% of its citizens are Flemish, speaking Dutch and 40% are Walloons who speak French. As multiethnic countries break up all over the world, we must ask: How did this country foster the peaceful coexistence of these two ethnic groups since 1830? Is Belgium a densely knitted society, where it does not matter if one is Flemish or Walloon? Or we have two nations within the same borders, that learned to minimize contact with each other?

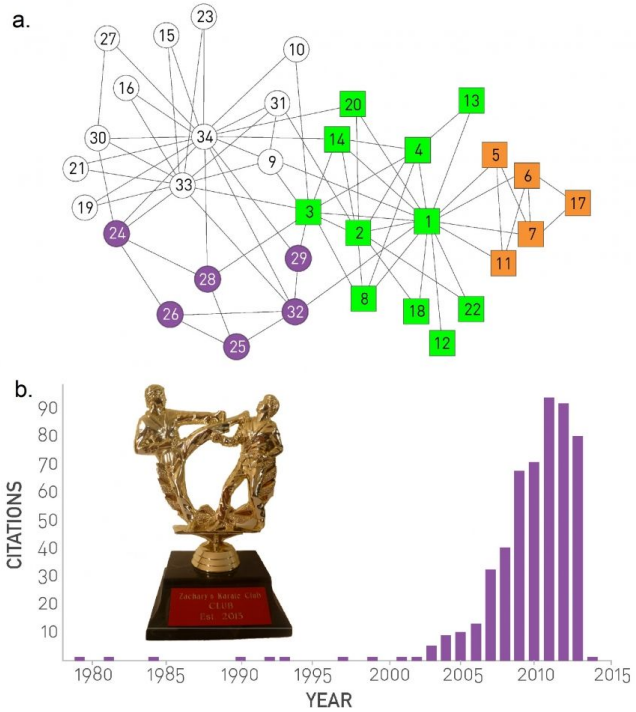
The answer was provided by Vincent Blondel and his students in 2007, who developed an algorithm to identify the country's community structure. They started from the mobile call network, placing individuals next to whom they regularly called on their mobile phone. The algorithm revealed that Belgium's social network is broken into two large clusters of communities and that individuals in one of these clusters rarely talk with individuals from the other cluster. The origin of this separation became obvious once they assigned to each node the language spoken by each individual, learning that one cluster consisted almost exclusively of French speakers and the other collected the Dutch speakers.

The color of each community on a red–green scale represents the language spoken in the particular community, red for French and green for Dutch.
The community that connects the two main clusters consists of several smaller communities with less obvious language separation, capturing the culturally mixed Brussels, the country's capital.



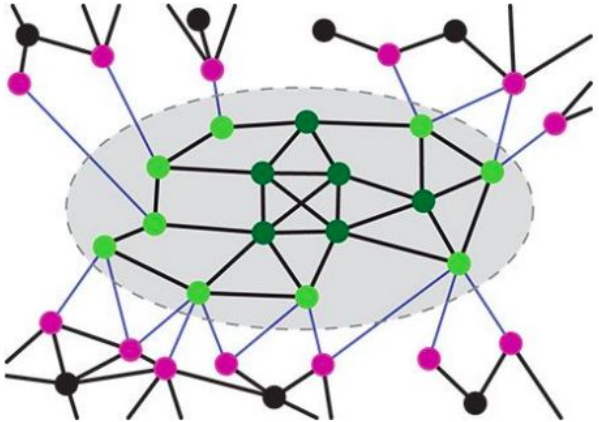
Zachary's Karate Club

- A social network that has received particular attention in the context of community detection, capturing the links between 34 members of a karate club
- Given the club's small size, each club member knew everyone else. To uncover the true relationships between club members, sociologist Wayne Zachary documented 78 pairwise links between members who regularly interacted outside the club.
- The interest in the dataset is driven by a singular event: A conflict between the club's president and the instructor split the club into two. About half of the members followed the instructor and the other half the president, a breakup that unveiled the ground truth, representing club's underlying community structure (Image a). Today community finding algorithms are often tested based on their ability to infer these two communities from the structure of the network before the split.



Communities: basic definitions

- What do we really mean by a community?
- How many communities are in a network?
- How many different ways can we partition a network into communities?



Schematic picture of a community (inside the gray oval) and of its immediate neighbors.

Communities: basic definitions

- The numbers of nodes and internal links in community C are N_c and L^c, respectively.
- The internal degree k_i^{int} and the external degree k_i^{ext} of a node *i* with respect to a community C are the number of links connecting *i* to nodes in C and to the rest of the network.
- The degree of *i* is $k_i = k_i^{int} + k_i^{ext}$
- If $k_i^{ext} = 0$ and $k_i^{int} > 0$ then *i* has neighbours only within C and is internal node.
- If $k_i^{ext} > 0$ and $k_i^{int} > 0$ for a node $i \in C$, then *i* has neighbours both inside and outside C and is a boundary node of C
- If $k_i^{int} = 0$, then the node is disjoint from C and has no neighbours inside C

Communities: basic definitions

- The internal link density is given by $\delta_C^{int} = \frac{L_C}{\binom{N_C}{2}} = \frac{2L_C}{N_C(N_C-1)}$
- The community degree, or volume, is the sum of the degrees of the nodes in C: $k_C = \sum_{i \in C} k_i$

Hierarchical Clustering

- To uncover the community structure of large real networks we need algorithms whose running time grows polynomially with N. Hierarchical clustering helps us achieve this goal
- The starting point of hierarchical clustering is a **similarity** matrix, whose elements x_{ij} indicate the distance of node *i* from node *j*
- In community identification the similarity is extracted from the relative position of nodes *i* and *j* within the network
- Once we have x_{ij}, **hierarchical clustering** iteratively identifies groups of nodes with high similarity
- There are two different procedures to achieve this:
  - **agglomerative algorithms** merge nodes with high similarity into the same community
  - **divisive algorithms** isolate communities by removing low similarity links that tend to connect communities
  - Both procedures generate a hierarchical tree, called a **dendrogram**, that predicts the possible community partitions

Hierarchical Clustering: Decide Group Similarity

- As nodes are merged into small groups, we must measure how similar two groups are
- Three approaches, called **single**, **complete** and **average cluster** similarity, are frequently used to calculate the groups similarity from the node-similarity matrix x_{ij}

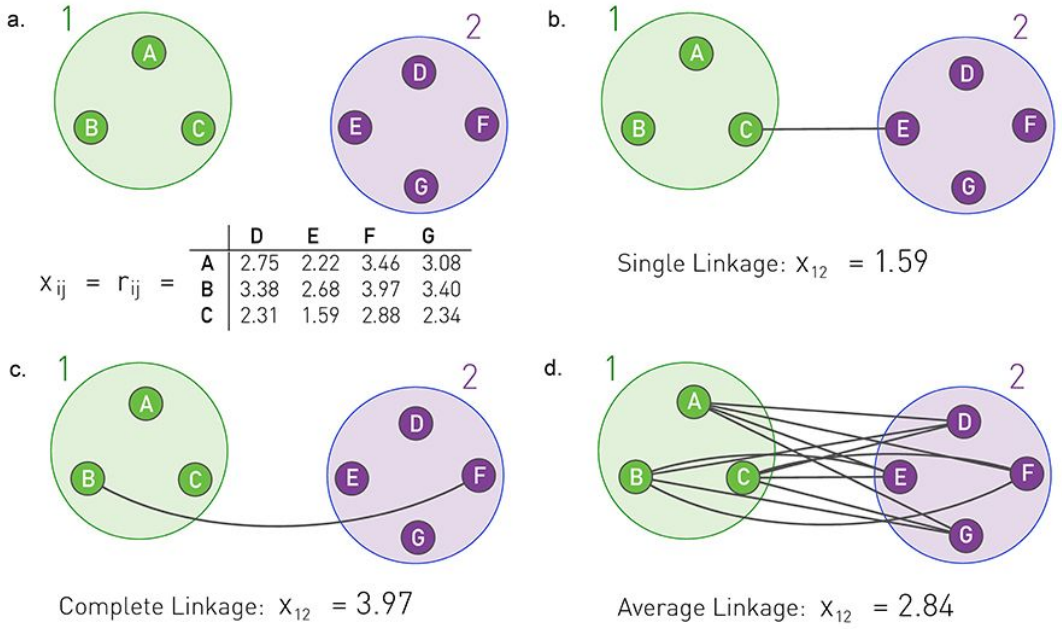
Hierarchical Clustering: Decide Group Similarity

Cluster Similarity

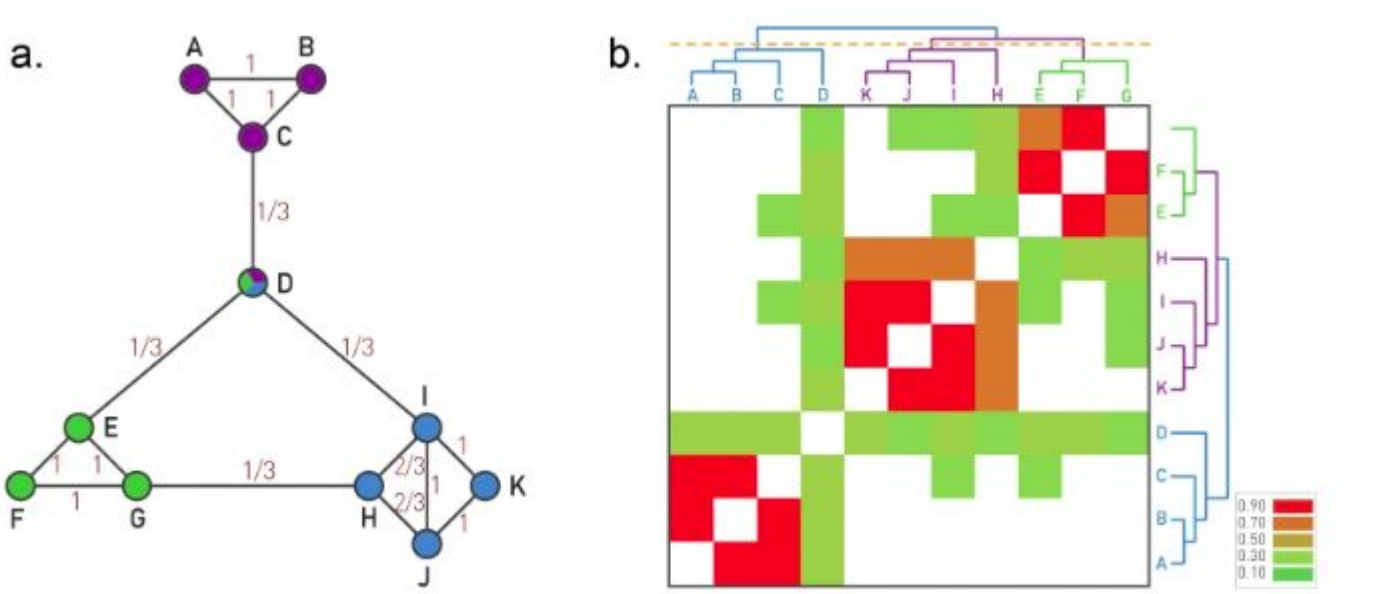
- We illustrate this procedure for a set of points whose similarity x_{ij} is the physical distance r_{ji} between them or derivada x_{ij}
- Seven nodes forming two distinct groups. The table shows the distance r_{ij} between each node pair, acting as the similarity

Single Linkage Clustering

- The similarity between communities 1 and 2 is the smallest of all x_{ij} , where i and j are in different groups. Hence the similarity is x_{12}=1.59, corresponding to the distance between nodes C and E.



Hierarchical Clustering: dendrograms



Communities in labor networks

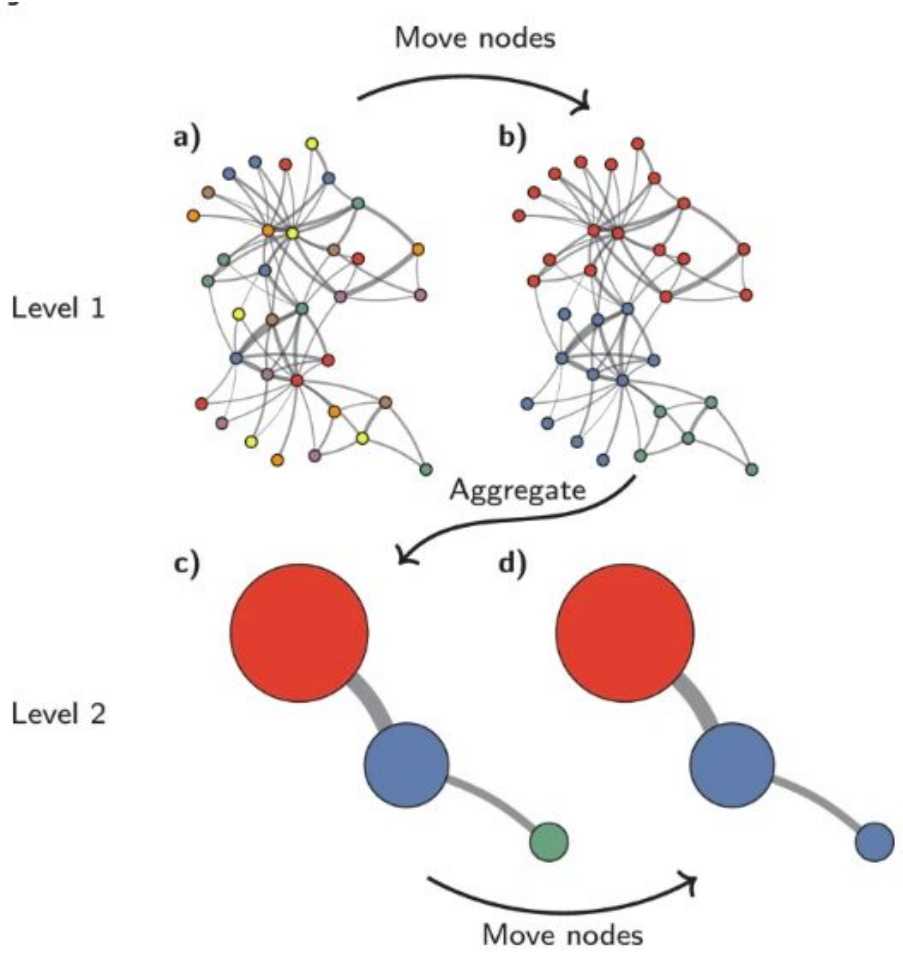
- Community detection algorithms like **Leiden** identify structures based on the *modularity* indicator as a quality criterion.
- This indicator measures the strength of a partition of the network into modules or groups by means of the density of links within the groups and the low density of links between groups.
- In networks of employment flows, this is equivalent to identifying groups of sectors among which a greater number of employment transitions are observed than between members of the same group and the rest of the network.
- This feature is useful in itself to identify sectors that may be sharing jobs based on the skills of the workers changing jobs.

Louvain Algorithm

- The algorithm optimises modularity in two elementary phases
  1. local moving of nodes
  2. aggregation of the network
- In the local moving phase, individual nodes are moved to the community that yields the largest increase in the modularity
- In the aggregation phase, an aggregate network is created based on the partition obtained in the local moving phase
- Each community in this partition becomes a node in the aggregate network
- The two phases are repeated until the quality function cannot be increased further

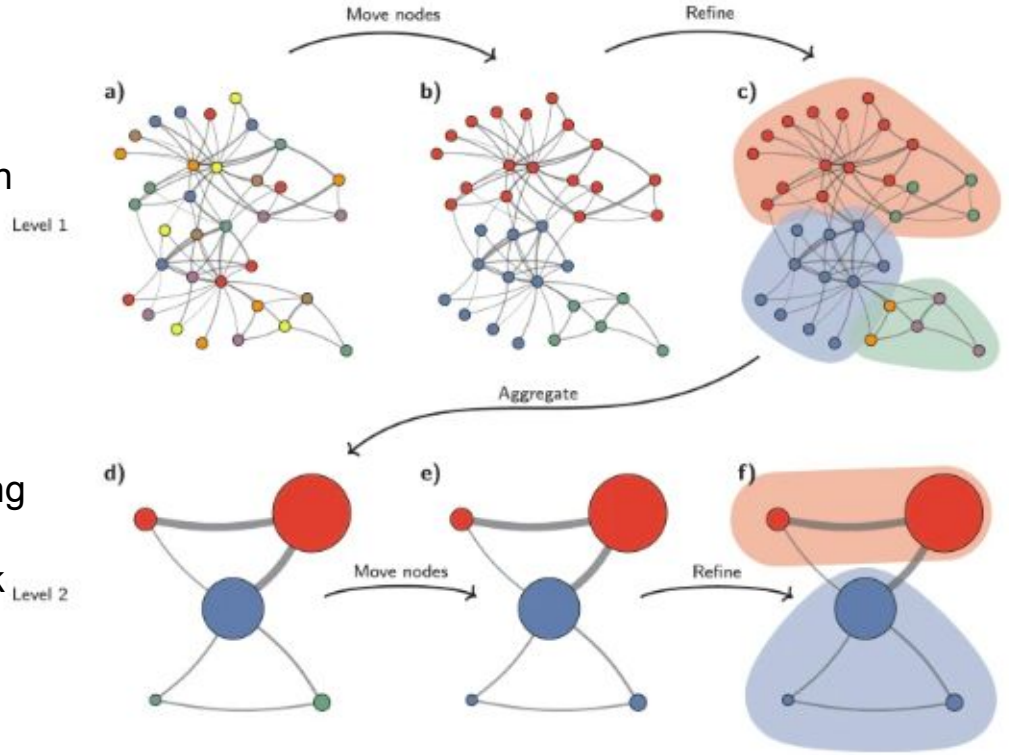
Louvain Algorithm

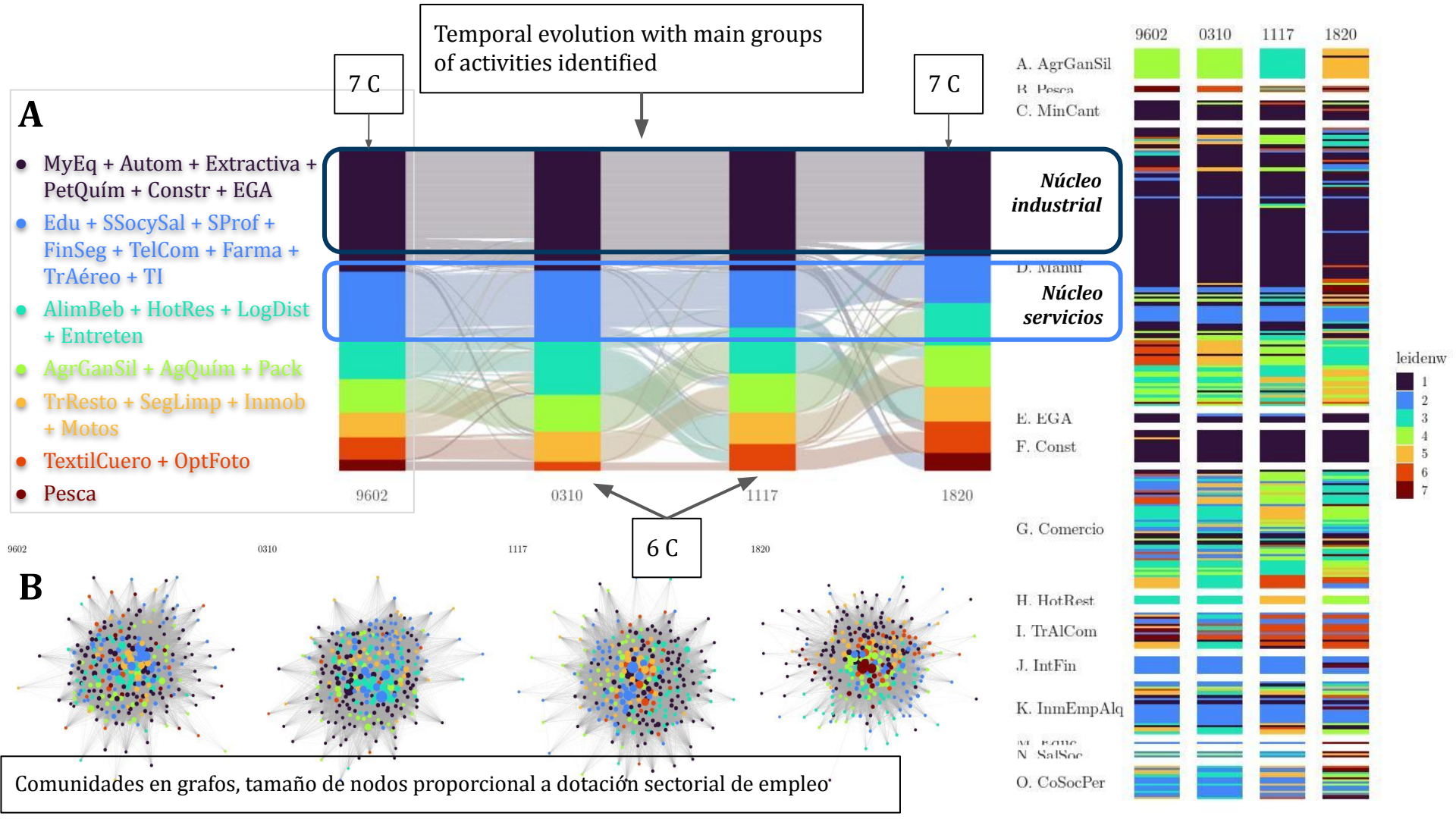
The Louvain algorithm starts from a singleton partition in which each node is in its own community (a).
The algorithm moves individual nodes from one community to another to find a partition (b).
Based on this partition, an aggregate network is created (c).
The algorithm then moves individual nodes in the aggregate network (d).
These steps are repeated until the quality cannot be increased further.



Leiden algorithm

The Leiden algorithm starts from a singleton partition (a).
The algorithm moves individual nodes from one community to another to find a partition (b), which is then refined (c). An aggregate network (d) is created based on the refined partition, using the non-refined partition to create an initial partition for the aggregate network. For example, the red community in (b) is refined into two subcommunities in (c), which after aggregation become two separate nodes in (d), both belonging to the same community. The algorithm then moves individual nodes in the aggregate network (e). In this case, refinement does not change the partition (f). These steps are repeated until no further improvements can be made.




Temporal evolution with main groups of activities identified
7 C
7 C
6 C
A
MyEq + Autom + Extractiva + PetQuím + Constr + EGA
Edu + SSocySal + SProf + FinSeg + TelCom + Farma + TrAéreo + TI
AlimBeb + HotRes + LogDist + Entreten
AgrGanSil + AgQuím + Pack
TrResto + SegLimp + Inmob + Motos
TextilCuero + OptFoto
Pesca
9602
0310
1117
1820
Núcleo industrial
Núcleo servicios
A. AgrGanSil
B. Pesca
C. MinCant
D. Manuf
E. EGA
F. Const
G. Comercio
H. HotRest
I. TrAlCom
J. IntFin
K. InmEmpAlq
N. SalSoc
O. CoSocPer
leidenw
1
2
3
4
5
6
7
B
Comunidades en grafos, tamaño de nodos proporcional a dotación sectorial de empleo

Inside the obtained communities

The number of sectors by their letter classification grouped in each community and period

Per./Com.	Sector (#)														Total (%)
	A	B	C	D	E	F	G	H	I	J	K	M	N	O	Total	
9602																
1			9	71	4	13	7		1		4				**109**	38,0
2				13			13		4	8	11	1	2	8	**60**	20,9
3				11			16	4			2			3	**36**	12,5
4	13			12			3				1		1		**30**	10,5
5				2		1	6		5		4			4	**22**	7,7
6				12			7				1				**20**	7,0
7		3		1					6						**10**	3,5
Total	**13**	**3**	**9**	**122**	**4**	**14**	**52**	**4**	**16**	**8**	**23**	**1**	**3**	**15**	**287**	**100,0**
0310																
1			8	72	3	14	6		1		4				**108**	37,6
2				12	1		11		4	8	12	1	2	10	**61**	21,3
3				7			22	4	7		5			5	**50**	17,4
4	13		1	14			3				1		1		**33**	11,5
5				16			10				1				**27**	9,4
6		3		1					4						**8**	2,8
Total	**13**	**3**	**9**	**122**	**4**	**14**	**52**	**4**	**16**	**8**	**23**	**1**	**3**	**15**	**287**	**100,0**
1117																
1			8	71	3	14	5		2		5				**108**	37,6
2				11	1		5		4	8	10	1	2	9	**51**	17,8
3	13	1		17			8				1		1		**41**	14,3
4				16			18				1				**35**	12,2
5				5			9	4			5			5	**28**	9,8
6		2	1	2			7		10		1			1	**24**	8,4
Total	**13**	**3**	**9**	**122**	**4**	**14**	**52**	**4**	**16**	**8**	**23**	**1**	**3**	**15**	**287**	**100,0**
1820																
1	1	1	7	54	4	14	7		1	1	4			1	**95**	33,1
2				13			5		4	6	10			4	**42**	14,6
3				19			17				1			1	**38**	13,2
4			1	9			14	4			5			4	**37**	12,9
5	12			14			3				1		1		**31**	10,8
6		2	1	8			4		11		1			1	**28**	9,8
7				5			2			1	1	1	2	4	**16**	5,6
Total	**13**	**3**	**9**	**122**	**4**	**14**	**52**	**4**	**16**	**8**	**23**	**1**	**3**	**15**	**287**	**100,0**

A relevant aspect to consider refers to the non-uniform distribution of sectors according to their grouping by classification letter.

D, the most numerous: 122 sectors is disaggregated into a greater number of sectors and is represented in more communities

less disaggregated tend to be concentrated in one community

Cuadro 6: Participación del empleo sectorial por comunidad (Leiden) y período.

Per./Com.	Sector (%)														
	A	B	C	D	E	F	G	H	I	J	K	M	N	O	Total
9602															
1			0,9	10,5	1,1	5,8	1,5		0,0		1,2				**21,0**
2				2,2			3,7		2,1	3,5	4,3	6,6	4,6	4,1	**31,1**
3				5,0			6,9	3,1			2,5			0,5	**18,0**
4	6,5			2,3			1,0				0,0		0,0		**9,9**
5				0,1		0,0	1,4		6,1		5,0			2,1	**14,7**
6				2,8			1,3				0,1				**4,2**
7		0,3		0,2					0,7						**1,2**
Total	**6,5**	**0,3**	**0,9**	**23,1**	**1,1**	**5,8**	**15,8**	**3,1**	**8,9**	**3,5**	**13,1**	**6,6**	**4,6**	**6,7**	**100,0**
0310															
1			1,0	9,4	0,7	6,7	0,9		0,0		1,2				**19,9**
2				2,1	0,2		3,3		2,0	2,5	5,8	7,1	4,2	4,3	**31,5**
3				2,9			9,0	3,7	6,1		6,9			2,3	**30,9**
4	6,2		0,0	3,6			1,0				0,1		0,0		**10,8**
5				2,9			3,0				0,0				**5,9**
6		0,3		0,2					0,5						**0,9**
Total	**6,2**	**0,3**	**1,0**	**21,0**	**0,9**	**6,7**	**17,1**	**3,7**	**8,5**	**2,5**	**14,0**	**7,1**	**4,3**	**6,5**	**100,0**
1117															
1			1,1	9,4	0,7	6,9	0,9		0,6		1,2				**20,8**
2				1,7	0,2		1,9		1,9	2,3	5,8	7,2	4,8	4,3	**30,2**
3	5,4	0,0		4,1			2,8				0,1		0,0		**12,5**
4				2,4			4,9				0,0				**7,4**
5				2,3			4,7	4,3			6,3			1,6	**19,0**
6		0,2	0,0	0,2			2,9		6,4		0,1			0,4	**10,1**
Total	**5,4**	**0,2**	**1,1**	**20,0**	**1,0**	**6,9**	**18,0**	**4,3**	**8,9**	**2,3**	**13,4**	**7,2**	**4,8**	**6,3**	**100,0**
1820															
1	0,0	0,0	1,1	5,9	1,1	6,7	1,8		0,0	0,0	0,9			0,5	**18,0**
2				1,1			1,2		1,9	1,6	5,9			0,4	**12,1**
3				2,5			4,6				0,0			0,1	**7,2**
4			0,0	3,9			6,9	4,4			6,0			0,7	**22,0**
5	5,5			2,6			0,9				0,0		0,0		**9,1**
6		0,2	0,1	1,2			1,2		6,9		0,2			0,6	**10,4**
7				1,3			1,6			0,7	0,1	8,0	5,4	4,1	**21,2**
Total	**5,5**	**0,2**	**1,2**	**18,4**	**1,1**	**6,7**	**18,2**	**4,4**	**8,9**	**2,3**	**13,3**	**8,0**	**5,5**	**6,3**	**100,0**

The analysis shows:

- structures with a certain degree of permanence in time and also differences that show their evolution are identified.
- period 1820 is identified differently

- two stable communities are clearly distinguished, each one forming a group of sectors with high interactions and function as centers, an industrial nucleus and a services nucleus, which together participate in more than half of the flows of employment and accumulate more than half of the total endowment of private employment.

Collaborators of the group

Sergio De Raco
Economist

Daniel Heymann
Economist, physicist

Instituto Interdisciplinario de Economía Política de Buenos Aires (IIEP-BAIRES)

NetLab
Laboratorio de Redes y Sistemas Complejos

✉ : netlab.iiep@gmail.com
: @NetLab_IIEP
: http://netlab.webiiep.econ.uba.ar/

c’est tout!
Thank you very much!