Visualization-based communities discovering in commuting networks : a case study

François Queyroi, Yves Chiricota

Visualization-based communities discovering in commuting networks : a case study

François Queyroi
LaBRI - Université de Bordeaux, France queyroi@labri.fr

Yves Chiricota
Université du Québec à Chicoutimi, Canada Yves_Chiricota@uqac.ca

Abstract

The division of a national territory is a mandatory process to analyze socio-economic dynamics. Commuting is then an important dimension to build such classification and weighted network analysis is adapted to study this phenomena. We present in this paper a procedure to identify groups of cities where commuters flow are relatively dense through a case study: a huge network which represents commuting in France (based on the 1999 national census). Our approach is based on a common technique improved by visual tools: highlight dense areas using a strength metric and extract clusters using the variation of a quality measure function.
Keywords: network analysis, graph visualization, clustering, commuting

1 Introduction

The definition of good spatial units is important for regional planing and geo-statistical analysis. Commuting had become a relevant dimension in numerous fields[10]. Commuting can be defined as the regular travel between place of residence and place of work. It is obviously related to the development of suburbs and commuter towns. A "Regionalization" of urban areas could not today be reasonably assess without taking commuters' flows into account. In this context, graph based methods have been used to visualize and study these flows[9].

The work we present here is based on the result of the 1999 french national census on all the national territory without overseas departments. According to this census there were about 3 millions commuters in France who correspond to 12% of the total labor force.

The network induced from these data contains about 36500 cities divided in 96 departments and 22 administrative regions (see Figure 1 for a map). The relations between the cities (network's nodes) are built as follow : two cities A and B are linked by an arc (oriented edge) if there is at least one person living in A and working in B. This arc is then weighted by the number of commuters going from A to B.

We are interested at finding *clusters*, which correspond to subset of nodes (cities). In the case of this work, a clustering corresponds to a *partition* of the set of nodes. That is, a collection of mutually disjoints subset such that their union gives the initial set of nodes. When nodes inside a cluster are again divided into subclusters the resulting configuration is denoted *hierarchical clustering*. Note that a possible hierarchical clustering is the division of french cities into administrative regions which are divided into departments.

Numerous network clustering procedures or algo-

Figure 1: Administrative division of France into Departments and Regions without overseas departments (*source : Le Robert - 1995*)

rithms have been developed in the last decades[6]. Gargiulo *et al.*[7] used a modularity maximization based algorithm[3] to test if new provinces of Sardinia (Italian island) correspond to labor basins found using the algorithm.

In section 2, we describe the official definition of urban areas used by French institute of statistics and economical studies (INSEE) which is based on commuters' flows. In section 3 we present a graph metric allowing to visually highlight dense areas. A classic procedure to calculate clusters according to this metric is introduced in section 4. By precisely describe how this method works, the section 5 presents an interactive and visual way to detect multi-scale clustering. The results are detailed in section 6.
The visualizations we present are built using Tulip, a network analysis framework[2].

2 Official INSEE Classification

The work we present here is based on the result of the 1999 french national census on all the national territory without overseas departments. The French institute of statistics and economical studies (INSEE) uses commuters' flows to define a partition of cities into **metropolitan areas** and **metropolitan regions** along with a classification into **urban cores**, **monopolar cities**, **multipolar cities** and **rural cities** which are parts of the ZAUER classification[1]. These concepts were developed after the 1999 french national census. This classification is mostly used in analysis of demographic evolution and then plays a important role in regional planning. We shall explain here its construction.

The base component of the **metropolitan area** is the **urban core** which is a group of close cities providing at least five thousand jobs such that any city inside this group does not belong to any other metropolitan area. The metropolitan area is then constructed iteratively by merging cities having at least 40% of their labor force commuting inside the area. These cities are designed as **monopolar**.
After that cities having 40% of their labor force commuting to multiple metropolitan areas are designed as **multipolar**. The metropolitan areas linked by multipolar cities form **metropolitan regions**.
A city which does not belong to any metropolitan area or region is designed as **rural**. The ZAUER actually provides a finner classification of rural areas but we shall here focus on urban areas where the commuting is stronger.

The ZAUER classification is illustrated in Figure 2. Note that a hierarchical clustering can be induced by the INSEE classification because metropolitan areas are included in metropolitan regions. Then the flows of workers can be analyzed at different scale. One can assume that flows are dense inside metropolitan regions and even denser inside metropolitan areas while being sparse between these regions.

[1] `http://www.insee.fr/en/methodes/`

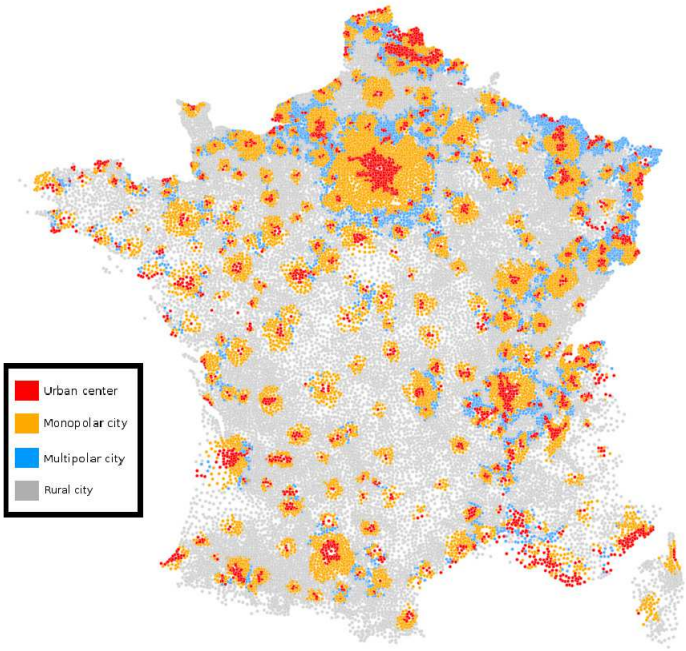


Figure 2: French cities according to the 1999 ZAUER classification

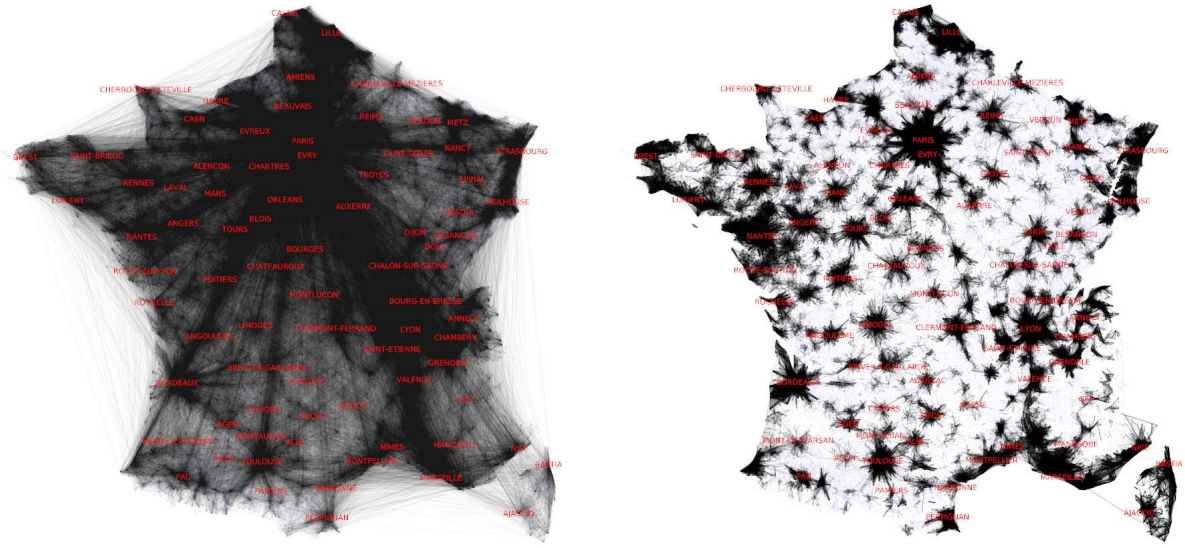
Two ideas underlie the way the ZAUER classification is built : first cities belonging to the same group are close one to each other. This corresponds to the fact that commuters destination is not far from their living place. Secondly areas where commuters' flows are stronger (here metropolitan areas) are often smaller than administrative departments or regions due to the fact that some cities can not reasonably be assigned to urban group (rural cities).

3 Highlighting dense areas

In order to identify close cities using commuters flows, we want to provide visualizations of areas where commuting is important. In terms of network analysis, we want to (visually) identify clusters of city. To do so, we start by define a metric capturing interesting topological features of this network.
To simplify our problem note that the orientation of the edges is not very relevant because we are looking for areas traveled by many workers and which are the origin or the destination of only a few workers. We thus replace each double way arc by a single edge weighted by amount of workers traveling between these two destinations. In terms of graph theory, we say that we make the graph simple.

To highlight dense areas we can begin be identifying the relations that do not likely belong to such areas. The strong metric values correspond to links in dense region. We can quantify this by calculating a *strength metric*[4, 1] on edges of the network taking the number of commuters into account. The precise definition of this metric is given in the box **"Evaluate edge strength"** in page 11.

This metric is very close to the Jaccard similarity between the neighborhood of u and v taking the weight of relations into account. A value close to



(a) Simple graph layout

(b) Linear mapping between strength metric values and a transparent to opaque scale

Figure 3: Representation of the 1999 French commuters network. Departments biggest city is labeled in red.

one indicates that the relation between the two cities occurs most likely within a dense area. On the other hand a value close to zero indicates that the relation could be either a bridge between two communities or an exchange of workers between two isolated cities.

A simple way to visualize the distribution of the metric over the network is the linear mapping between metric values and a color scale applied on edges in the layout (see Figure 3(b)). The idea is to filter out the low values. In this image we removed edges having a value below 0.5.

The image displayed in Figure 3(b) clearly contains some interesting features. First note that relations with a high metric value are not uniformly spread over the network but are most of the time gathered in small regions especially within area close to big cities. These areas seem also coincide with peripheral regions around big cities. Observe that these groups of highly valuated edges can be of different size. Some of these groups are linked by high valued edges, revealing the presence of hierarchies in the network.

Looking at the strength metric mapping in Figure 3(b) and the ZAUER classification in Figure 2 one can easily see that urban area matches with regions of the graph where the strength metric is high. Those simple observations validate our approach. The visualization of dense areas may however differ on some part of the network. For instance in the West of France, we can see high valuated edge crossing wide areas over the coasts. However, in the ZAUER classification, these regions contain many rural cities and dense regions are concentrated around big cities.

4 Clusters calculation

An intuitive way to retrieve clusters of cities inside the network consists to filter out the low valued edges in relation to the metric and assume that

two nodes are in the same cluster when they are connected by an edge having a high strength value. In terms of graph theory, the clustering is given by the connected component resulting of the removal of the low valued edges. This procedure is known as *single-linkage clustering*[6].

However, to enforce this method we need to define what is a strong or a weak edge according to our measure. A convenient approach consists to use a threshold: an edge is considered weak if its strength metric is below this threshold (the edge is discarded) and strong otherwise (the edge is kept). The procedure is illustrated in the box **"Single Linkage clustering"** in page 12.

This method allow hierarchical clustering. Indeed, let t_1 and t_2 be two thresholds such as $t_1 < t_2$, the clustering corresponding to t_2 can be obtained by applying the single-linkage procedure to each group of the clustering corresponding to t_1. The single-linkage clustering is then well adapted to our study because we suppose that hierarchies may exist in the network formed by commuters' flows.

In order to evaluate the groups of cities found with a given threshold, we use a quality measure. They are often used in graph clustering algorithm to compare methods or choose between different results (see [6] for a survey on graph clustering techniques).

The quality measure used here is the MQ measure first introduced in [8] and further analyzed in [5] (see its definition in the box **"Evaluate clustering quality"** in page 13). This measure is based on the difference between internal and external connectivity ratio. The MQ value is close to 1 when clusters are densely connected while the connections with the rest of the network are sparse. An important feature of this measure is that the size of clusters is taken into account. It means that a cluster consisting of only few cities has a lesser impact on the MQ value than a cluster composed of hundred of cities.

A threshold value is associated with the corresponding MQ value. Most of the time the quality measure is used to decide the best threshold (we seek the threshold corresponding to the maximum MQ value). However doing so risks to discard interesting phenomena such as the presence of hierarchies inside the network. This idea is developed in the next section.

5 Visual based procedure

As said in Section 2, mapping of a color scale on edge according to strength metric is effective at highlighting dense areas. It is however hard to determine the thresholds to use in order to identify a hierarchical clustering of a network. We explain in this section how one can use the evolution of MQ to detect this kind of features and turn clustering of the network into a data exploration process.

Each variation of the quality measure MQ corresponds to different kind of evolution of the clustering:

- **Steady state:** Most of the time no variation occurs if no other edges are discarded at this step, the clustering then stays the same.
- **Slow increase/decrease:** This situation occurs when several small clusters are disconnected from a larger component, making this component slightly denser/sparser. And because the disconnected small clusters have not a huge weight in the MQ value, the increase/decrease of the measure is not very high.
- **Step upward/downward:** At some value of the threshold a big component can be cut into smaller clusters which are big and dense enough to lead to a huge gain of the MQ measure. Alternatively and most of the time for a large threshold value, dense clusters may be totally disconnected leading to an important lost of quality.

The behaviors listed above combined to give a visual representation which helps to understand the clustering process. The hierarchical organization of a network can then be inferred using the evolution of MQ by looking for local maxima (huge variations in the

measure followed by a null or negative gains) which are relatively close to global maximum (to guarantee a certain robustness of each level of the hierarchical clustering). Instead of using heuristic we can rely on the human eyes for several reasons:

- The analysis of the evolution of MQ can be coupled with a filtering of edges in the graph layout.
- The user can compare MQ curves of various network and detected similar connectivity patterns.

6 Results

We apply the procedure described above on french administrative regions. Several reasons justify this choice. First, people living in a region and working in another only represent 5% of the total number of commuters. Secondly, cities that send more workers outside the region than inside are most of the time located near the borders separating these regions. Finally when looking at Figure 3(b) we can see that high valued edges barely cross regions' borders. In this section we detail the results for four regions, each of them illustrates a different phenomena.

In Figure 4 we present MQ curves in relation to these regions. The result of our procedure for each region is shown in Figure 5.

The Figure 4(a) corresponds to the region Ile-de-France having Paris as capital. Looking at the evolution of MQ for this sub-network it is hard to detect any significant increase. Indeed, increasing the threshold value disconnects cities that are less connected (often at the border of the region).

The situation is very different for the region Basse-Normandie (Figure 4(b)) : the positive variation of MQ is stronger and leads to a single threshold which also corresponds to the maximal value. No significant hierarchical configuration can really explain the commuters' flow occurring in this region. Looking at the representation in Figure 5(b), we note that the clusters correspond most of the time to the suburbs of the biggest cities. Note also that these groups are very distant and separated by singletons that are actually rural cities.

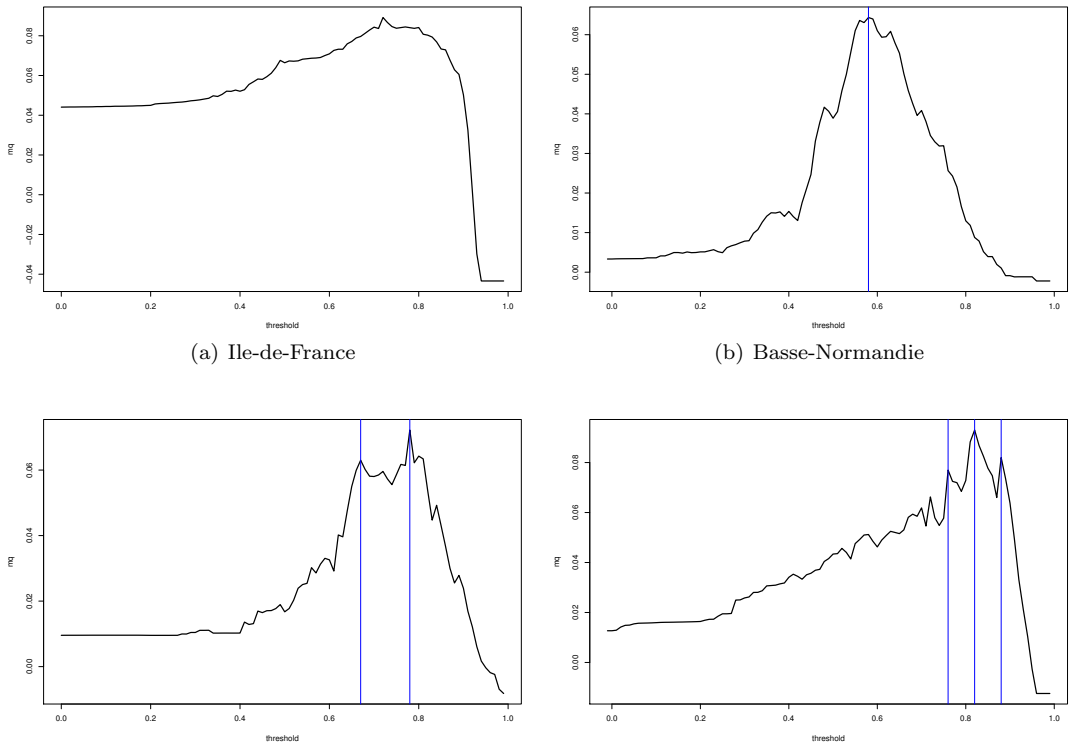
The analysis of the MQ curves for the region Pays-de-la-Loire and Provence-Alpes-Côte d'Azur reveals that these regions contain areas we can hierarchically decompose. The Figure 5(d) shows that the dense groups are located in the south (near the Mediterranean sea) and include some important cities (such as Marseille or Toulon). This clustering does not differ so much than the ZAUER classification. However, we see that we can use a third level to disconnect smaller and very dense areas.
The region Pays-de-la-Loire mostly consists of isolated metropolitan areas in the ZAUER classification. However we found that a larger group which contains three important metropolitan areas (Around Nantes, Angers and the North of the Vendee) can be found. It can be explained by the fact that road and rail infrastructure is very developed between these zones.

The Figure 6 shows the results for all the french territory. Note that groups size is larger when they contain one or more big cities. We can also see that wider groups can be detected over the coasts. A hierarchical organization of the commuters' flows is mostly found for the regions located in the West or in the South.

7 Conclusion and future works

We introduced in this paper a procedure to detect multilevel clustering in commuters network. In the literature, graph clustering algorithms are most of the time black box tools. With our method however, the user (geographer or sociologist) is able to visually mine the network that he/she studies. Combining edges filtering with the evolution of a suited quality measure helps the detection of dense clusters and hierarchies inside a network.

We enforced this procedure to study French com-



(a) Ile-de-France

(b) Basse-Normandie

(c) Pays de la Loire

(d) Provence-Alpes-Côte d'Azur

Figure 4: Evolution of MQ values for several French regions in the 1999 commuters' flow network.

muters' flows. It appears that hierarchies of cities can be found for several regions. The results provide a different kind of information that the ZAUER classification. However, geographers evaluation is needed to interpret these results.

An interesting lead to validate the choice of the threshold values is to use a multilevel quality measure introduced in [5]. We should be able to tell if whether or not the local maxima in the evolution of MQ correspond to the best hierarchical clustering.

In this paper we do not focus on the best way to visualize groups. The cities are geolocalized and we assume that using nearly-convex hulls which overlap in case of hierarchical decomposition of the area is a good approach. But if we try to track the evolution of the dense areas over the years (using the previous national census' data for example) we think that tools such as morphing of concave hulls is a more effective approach.

References

[1] D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon. Multiscale visualization of small world networks. 2003.

[2] David Auber. Tulip - a huge graph visualization framework. In Petra Mutzel and Mickael Jnger, editors, *Graph Drawing Software*, Mathematics and Visualization Series. Springer Verlag, 2003.

[3] V.D. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008:P10008, 2008.

[4] Y. Chiricota, F. Jourdan, and G. Melançon. Software components capture using graph clustering. In *11th IEEE International Workshop on Program Comprehension*, 2003.

[5] Maylis Delest, Guy Melançon, François Queyroi, and Jean-Marc Fédou. Assessing the Quality of Multilevel Graph Clustering. Technical report, LaBRI, 2011.

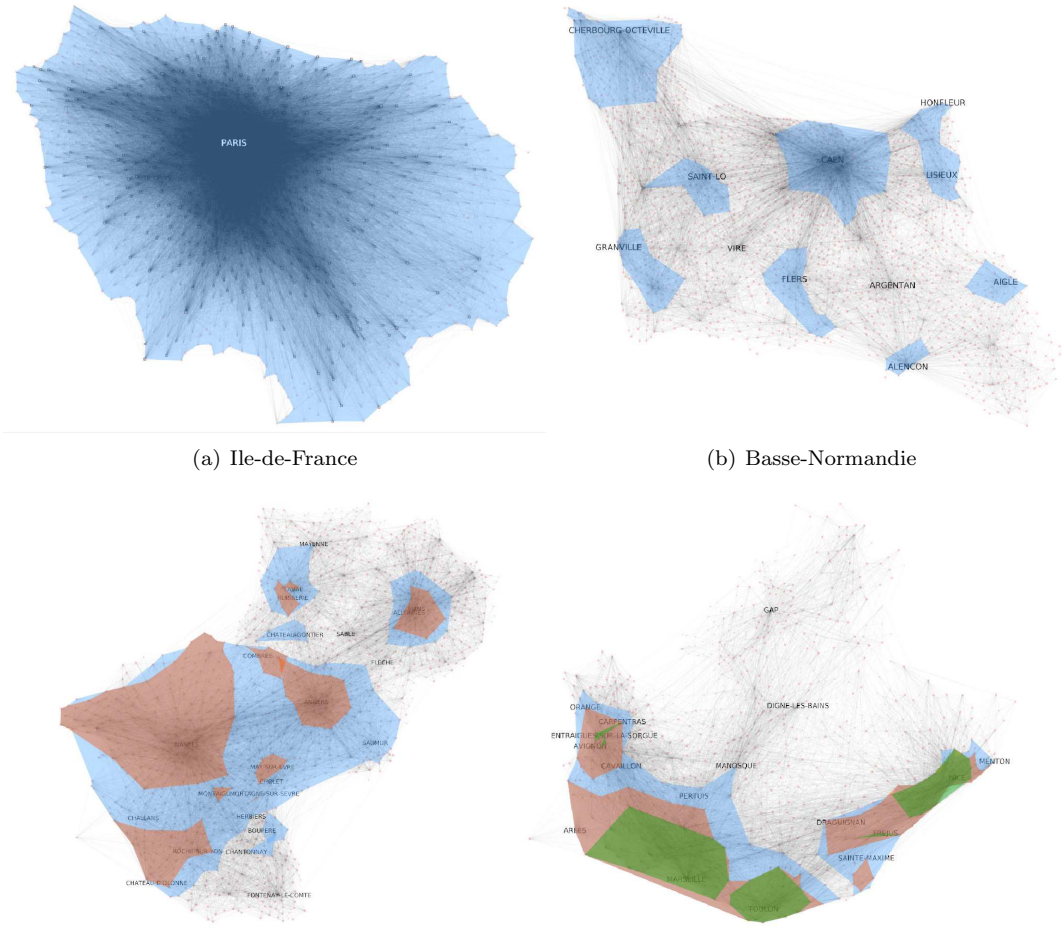
[6] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[7] F. Gargiulo, M. Lenormand, S. Huet, and O.B. Espinosa. A commuting network model: going to the bulk. *Arxiv preprint arXiv:1102.5647*, 2011.

[8] Spiros Mancoridis, Brian S. Mitchell, C. Rorres, Y. Chen, and E. Gansner. Using automatic clustering to produce high-level system organizations of source code. In *IEEE International Workshop on Program Understanding (IWPC'98)*, 1998.

[9] R. Patuelli, A. Reggiani, S.P. Gorman, P. Nijkamp, and F.J. Bade. Network analysis of commuting flows: A comparative static approach to German data. *Networks and Spatial Economics*, 7(4):315–331, 2007.

[10] J. Rouwendal and P. Nijkamp. Living in Two Worlds: A Review of Home-to-Work Decisions. *Growth and Change*, 35(3):287–303, 2004.



(a) Ile-de-France

(b) Basse-Normandie

(c) Pays de la Loire

(d) Provence-Alpes-Côte d'Azur

Figure 5: Representation of the hierarchical clusterings found using the threshold values chosen in Figure 4. Only the groups of cities which contains more than 5000 workers are shown. The groups are displayed using nearly-convex hulls. The color of the hulls corresponds to the depth of the cluster inside the hierarchy (first level: blue, second: brown, third: green). Finally, the name of the biggest city of each groups is shown.

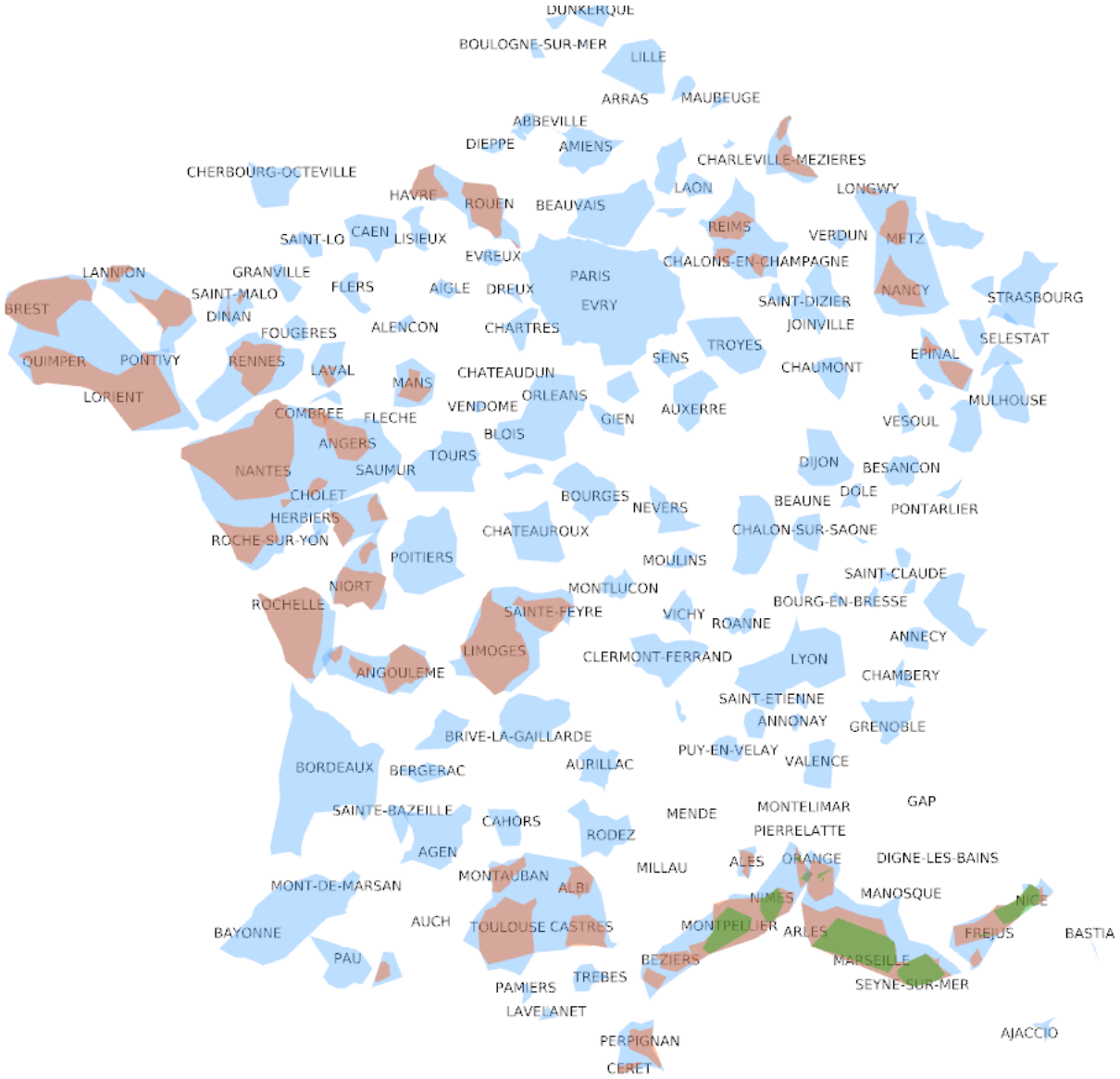


Figure 6: Clustering of the 1999 commuters' flows network drawn with nearly-convex hulls. Only groups of cities containing at least 5000 workers are shown. Labels corresponds to the biggest city in each areas in terms of labor forces.

Technical notes

Evaluate edge strength

Let u, v (see Figure 7 for a small example) be the two cities and $t(u, v)$ be the number of commuters between u and v. We also define the direct neighborhood of u as N_v which is a set of cities w such as $t(u, w) > 0$ (including v). The number of commuters traveling in the direct neighborhood of both city u and v is given by :

$$I(u, v) = \sum_{w \in N_u \cap N_v} (t(u, w) + t(v, w)) - t(u, v)$$

and let

$$E(u) = \sum_{w \in N_u \setminus N_v} t(u, w)$$

be the number of commuters traveling in neighborhood of the city u but not in the neighborhood of v. Our strength metric (denoted J) is then

$$J(u, v) = \frac{I(u, v)}{2(E(u) + E(v)) + I(u, v)}$$

In the network drawn in Figure 7, we have $I(u, v) = 10$, $E(u) = 11$ and $E(v) = 10$ then we have $J(u, v) \approx 0.19$.

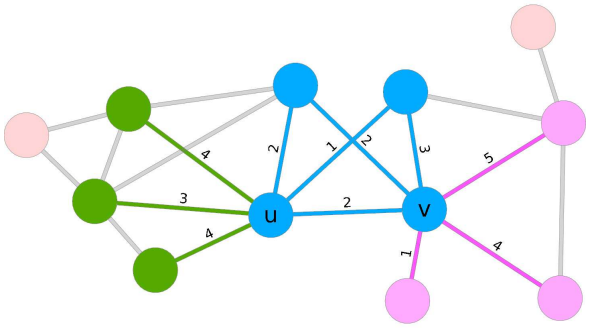


Figure 7: A example of small network. Edges labels indicate their value. The blue part represents the common neighborhood of both entities u and v ($N_u \cap N_v$), the green is the exclusive neighborhood of u ($N_u \setminus N_v$) and the violet the v one ($N_v \setminus N_u$).

Single Linkage Clustering
The procedure runs as follow:

1. Compute the strength metric for each edge of the graph
2. Remove edges below a given threshold
3. Take connected components of the resulting graph as a clustering

Figure 8 illustrates the procedure. With a threshold equals to 0.2, the weak edges (in grey) are removed. This new network has four connected components (red, green, blue and orange) which form a clustering containing two singletons (blue and orange). Taking a threshold value equal to 0 does not disconnect the network while taking 0.8 or more results in a clustering containing only singletons.

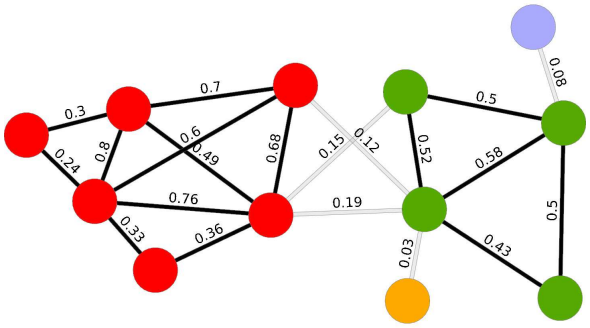


Figure 8: Illustration of *single-linkage clustering* based on the example introduced in Figure 7.

A hierarchical clustering can be obtained as follow: taking $t_1 = 0.2$ leads to the clustering shown in Figure 8. Taking another threshold $t_2 = 0.3$ yields to a hierarchical clustering by splitting the red colored nodes into three subclusters.

Evaluate clustering quality
Let C be a clustering of cities *i.e.* C_i corresponds to a group of cities, its size is denoted $|C_i|$. Set

$$W_{in}(C_i) = \sum_{u \neq v \in C_i} J(u, v)$$

the sum of the J-metric for edges *within* the cluster C_i and

$$W_{out}(C_i) = \sum_{u \in C_i} \sum_{v \in V \setminus \{C_i\}} J(u, v)$$

the sum of the J-metric for edges *outside*. The network contains a total of n cities. The MQ quality measure is then

$$MQ = \frac{1}{n} \sum_{i=1}^{k} \left(\frac{2W_{in}(C_i)}{|C_i| - 1} - \frac{W_{out}(C_i)}{n - |C_i|} \right)$$

This measure is bounded by $[-1, 1]$.

Consider the small network illustrating the construction of the J-metric. The Figure 8 provides an example of clustering for this graph using an arbitrary threshold value. The resulting clustering denoted C is composed of four clusters. For example taking C_{red} the cluster corresponding to the red colored nodes, we have $W_{in}(C_{red}) = 5.26$ and $W_{out}(C_{red}) = 0.46$. Finally we get $MQ \approx 0.3$.

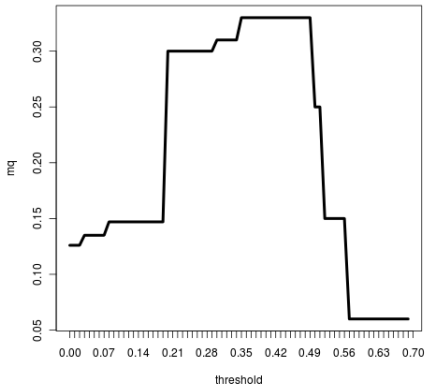


Figure 9: Evolution of the MQ (y-axis) measure according to the threshold value (x-axis) used to clusterize the example network in Figure 8.

Looking at the MQ curve for the example network (Figure 9), we can visually identify a phase of slow increase (orange then blue clusters are disconnected), a huge variation (red and green clusters are separated), another phase of slow increase (two red nodes are disconnected) then MQ rapidly falls (the dense red and green clusters are disconnected).