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

François Queyroi*, Yves Chiricota**

*Université de Bordeaux, CNRS, LaBRI, France
francois.queyroi@labri.fr
**Université du Québec à Chicoutimi, Canada
Yves_Chiricota@uqac.ca

Abstract. The division of a national territory is a mandatory process to analyse socio-economic dynamics. Commuting is then an important dimension to build such classification and weighted network analysis is adapted to study this phenomenon. We present in this paper a procedure to help users to identify hierarchical partitions of cities that capture commuters flows density. We enforce our method on a 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 at different levels using the variation of a quality measure function.

1 Introduction

The definition of good spatial units is important for regional planning and geo-statistical analysis. Spatial network analysis based on different kinds of human interactions has been used in this context. A good example is the study of a telecommunication network in Great Britain by Ratti et al. (2010). Another interesting approach is the study of commuting (Gargiulo et al. (2011); Rouwendal and Nijkamp (2004)) which 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 assessed without taking commuters flows into account. In this context, graph based methods have been used to visualize and study these flows (Patuelli et al. (2007)).

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 follows : 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.

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

We are interested in finding *clusters*, which correspond to subsets 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 disjoint subsets 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. We want to find alternative classifications of cities that can be used by regional planners. This problem will therefore not be solved using overlapping clustering. We hope the situation where a city can not be assigned to only one group shall be captured by the way groups are organized hierarchically. This approach also matches with the official classification used by the French institute of statistics and economical studies (INSEE) which is described latter.

Numerous network clustering procedures or algorithms have been developed in the last decades (Fortunato (2010)). De Montis et al. (2011) used the modularity maximization based algorithm of Blondel et al. (2008) to test if new provinces of Sardinia (Italian island) correspond to labor basins found using the algorithm. Another procedure due to Lancichinetti et al. (2011) was applied to the United-Kingdom commuters' network.

In this paper, we propose a procedure to that allow user to extract hierarchical partitions from a weighted network. We illustrate the relevance of this approach by looking at the commuters networks induced by four French regions. The rest of this paper is organized as follows. In section 2, we describe the official definition of urban areas used by the 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 describing 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 as long as a discussion of our results when compared to existing methods and algorithms. The visualizations we present are built using Tulip, a network analysis framework (Auber et al. (2012)).

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 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 an 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 classification 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 scales. One can assume that flows are dense inside metropolitan regions and even denser inside metropolitan areas while being sparse between these regions.

Two ideas underlie the way the ZAUER classification is built : firstly 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).

1. http://www.insee.fr/en/methodes/

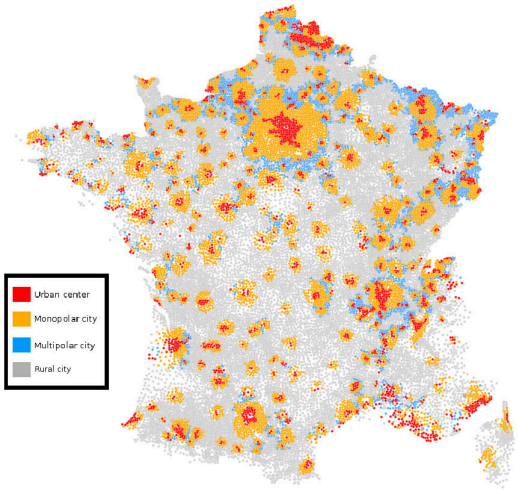


FIG. 2: French cities according to the 1999 ZAUER classification

3 Highlighting dense areas

In order to identify close cities using commuters flows, we want to provide visualizations of areas where commuting is important and (visually) identify clusters of cities. To do so, we start by defining a metric capturing interesting topological features of this network. To simplify our problem note that the orientation of the edges is not very relevant. Indeed, we are looking for areas where the commuting phenomenon is important (*i.e.* densely connected subsets of nodes) not including cities that are the origin or the destination of only a few workers. (*i.e.* weakly connected nodes). We thus replace each double way arc by a single edge weighted by amount of workers travelling between these two destinations.

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 regions. We can quantify this by calculating a *strength metric* (Auber et al. (2003); Chiricota et al. (2003)) on edges of the network taking the number of commuters into account. This metric, described here, is denoted $J(u, v)$ for a link (u, v) in the network.

Let u, v (see Figure 3 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 travelling in the direct neighborhood of both city u and v is given by :

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

and let

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

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

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

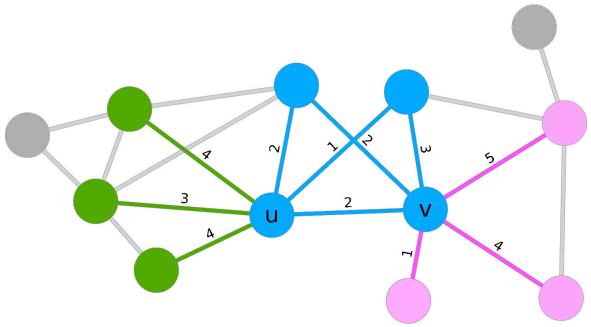
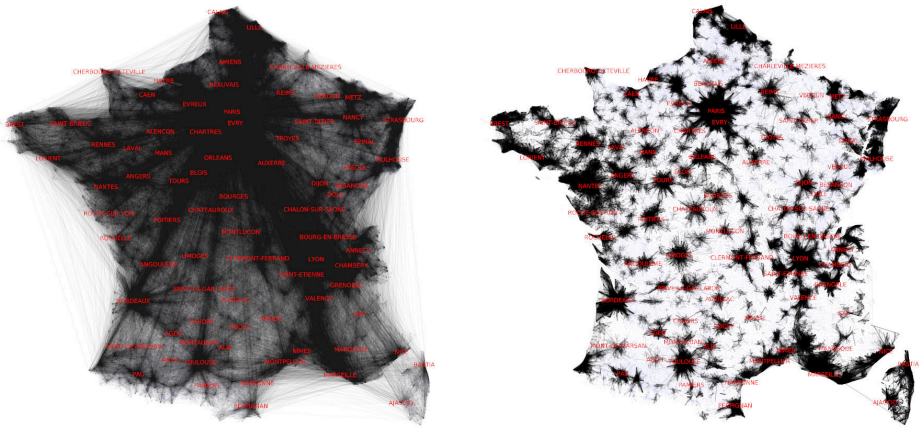


FIG. 3: An example of small network. Edges labels indicate their value. The blue part represents the common neighborhood of both entities u and v ($N_u \cap N_v$), the green is the exclusive neighborhood of u ($N_u \setminus N_v$) and the pink that for v ($N_v \setminus N_u$). We have $I(u, v) = 10$, $E(u) = 11$ and $E(v) = 10$ then we have $J(u, v) \approx 0.19$

This metric is close to the Jaccard index (Hamers et al. (1989)) between the neighborhood of u and v taking the weight of relations into account. A value close to 1 indicates that the relation between the two cities occurs most likely within a dense area. On the other hand a value close to 0 indicates that the relation could be either a bridge between two communities or an exchange of workers between two isolated cities. A trivial algorithm to compute J is to compare u and v neighborhoods for each edge (u, v). The time complexity is $\mathcal{O}(mn^2)$ where m (resp. n) is the number of edges (resp. vertices) in the network.

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 4(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 4(b) 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 areas close to big cities. These areas seem to also coincide with peripheral regions around big cities. Observe that these groups can be of different sizes and can be linked together with edges of significant weight, revealing the presence of hierarchies in the network.

Looking at the strength metric mapping in Figure 4(b) and the ZAUER classification in Figure 2 one can easily see that urban areas match with regions of the graph where the strength



(a) Simple graph layout

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

FIG. 4: Representation of the 1999 French commuters network. Departments biggest city is labeled in red.

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 edges with high value 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 in filtering out the low valued edges in relation to the metric and assuming 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 components resulting of the removal of the low valued edges. This procedure is known as *single-linkage clustering* (Fortunato (2010)). To enforce this method we need to define what is a strong or a weak edge according to our measure. A convenient approach consists in using 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).

Figure 5 illustrates the procedure. With a threshold equal 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.

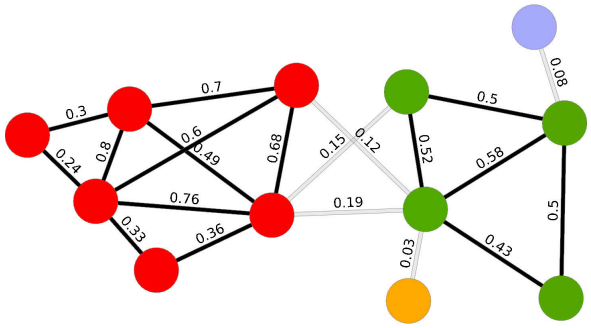


FIG. 5: Illustration of *single-linkage clustering* based on the example introduced in Figure 3.

This method allows 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. In the example of Figure 5, taking another threshold $t_2 = 0.4$ yields to a hierarchical clustering by splitting the red coloured nodes into three subclusters. 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. The quality measure used here is the MQ measure first introduced by Mancoridis et al. (1998) and further analysed by Delest et al. (2011). This measure is based on the difference between internal and external connectivity ratio and is bounded by $[-1, 1]$. The MQ value is close to 1 when clusters are densely connected while the connections with the rest of the network are sparse. 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

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

Note that even if we filter edges below the threshold value to determine the clustering, the measure MQ is computed for the whole network including filtered edges. Consider the small network illustrating the construction of the J-metric. The Figure 5 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 coloured nodes, we have $W_{in}(C_{red}) = 5.26$ and $W_{out}(C_{red}) = 0.46$. Finally, we get $MQ \approx 0.3$.

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 lower impact on the MQ value than a cluster composed of hundred of cities. The method described here is then well adapted to the fact that some rural cities cannot reasonably be assigned to larger and denser group (see the ZAUER classification). 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 Visualization-based procedure

As said in Section 1, mapping of a colour scale on edges 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 kinds of evolution of the clustering:

- **Steady state:** Most of the time no variation occurs if no other edge is 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 do not have 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 loss of quality.

The behaviours listed above may be combined to give a visual representation which helps to understand the clustering process. Looking at the MQ curve for the example network (Figure 6), 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). Even if the best MQ value is obtained after the removal of two nodes from the red cluster, the major change happen here when the red and green cluster appear.

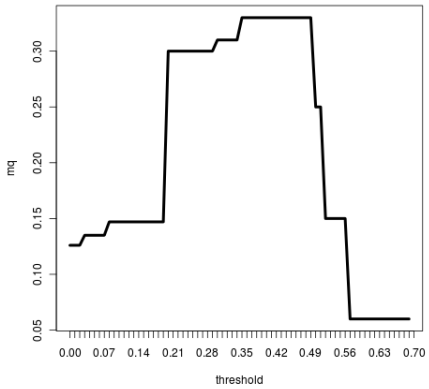


FIG. 6: Evolution of the MQ (y-axis) measure according to the threshold value (x-axis) used to clusterize the example network in Figure 5.

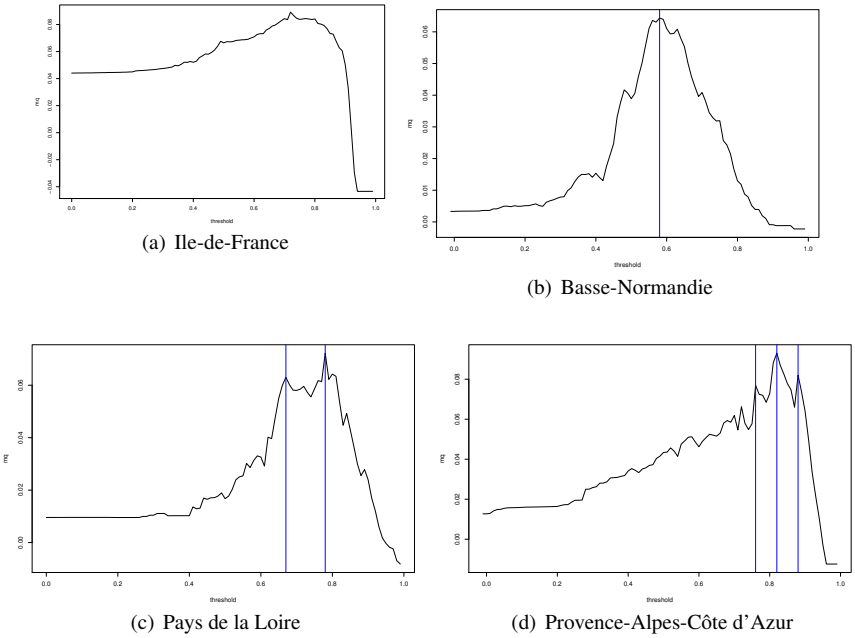
The plot given by the pairs (threshold,MQ) value is used here to extract hierarchical clusterings. A hierarchical organization of the network can then be inferred using the evolution of MQ by looking for local maxima (huge variations in the measure followed by a null or negative gains) which are relatively close to the global maximum (to guarantee a certain robustness of each level of the hierarchical clustering). Instead of using heuristics we can rely on the human eyes for several reasons:

- The user knows the number of level he/she wants (the INSEE uses a two level classification).
- Several alternatives partitioning can be found for the same network.
- The analysis of the evolution of MQ can be coupled with a filtering of edges in the graph layout.
- MQ curves of various networks can be compared to find similar connectivity patterns.

For a given threshold t_1, the partition and the corresponding MQ value can be computed in linear time. Plotting the MQ curve for l different threshold value would take $\mathcal{O}(lm)$ in this case. In practice, the procedure can be implemented efficiently by updating the partition and the MQ value. Indeed, when an edge is removed either the partition does not change or a cluster is split into two sub-clusters. The MQ value can be updated by removing the gain of the previous cluster and adding the gains of the two sub-clusters.

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 4(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 phenomenon.



(a) Ile-de-France

(b) Basse-Normandie

(c) Pays de la Loire

(d) Provence-Alpes-Côte d'Azur

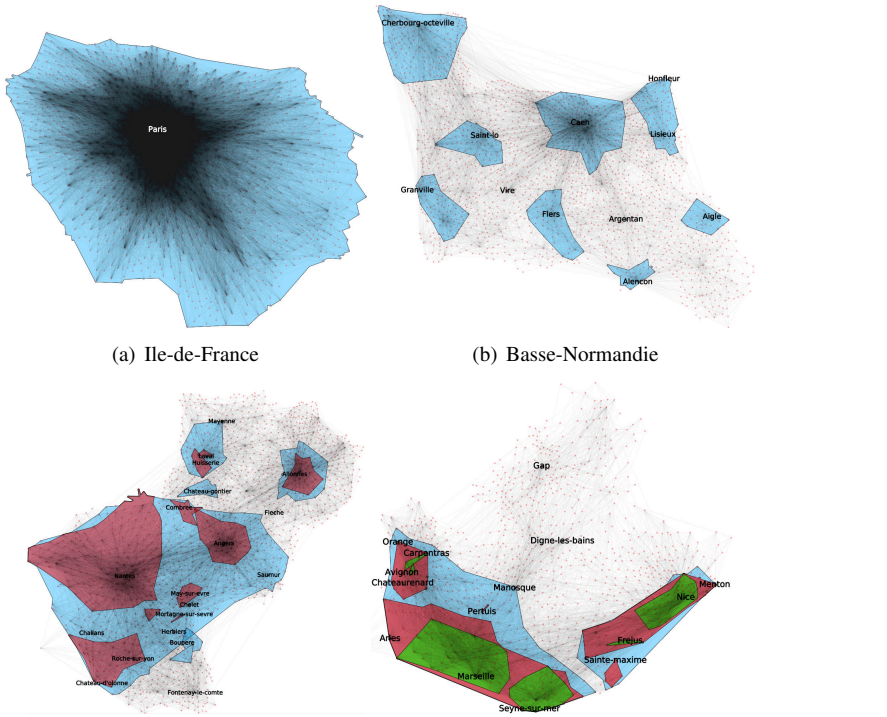
FIG. 7: Evolution of MQ values for several French regions in the 1999 commuters' flow network. The blue vertical lines indicate the threshold selected for each region.

In Figure 7 we present MQ curves in relation to these regions. The result of our procedure for each region is shown in Figure 8.

The Figure 7(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 7(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 8(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 8(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 from the ZAUER classification. However, we see that we can use a third level to



(a) Ile-de-France (b) Basse-Normandie

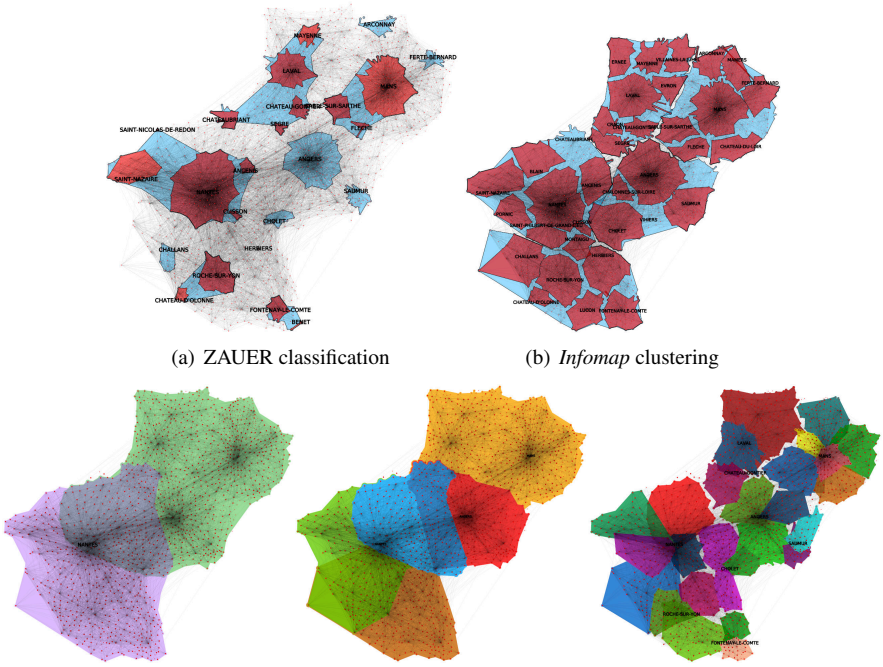
(c) Pays de la Loire (d) Provence-Alpes-Côte d'Azur

FIG. 8: Representation of the hierarchical clusterings found using the threshold values chosen in Figure 7. Only the groups of cities which contain more than 5000 workers are shown. The groups are displayed using concave 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 group is shown.

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 Vendée) can be found. It can be explained by the fact that road and rail infrastructure is very developed between these zones.

We shall now compare our results to different classifications. The Figure 9 provided the results using the ZAUER classification (Fig. 9(a)) detailed in Section 2. We used two recent hierarchical clustering algorithms: the *Infomap* algorithm (Rosvall and Bergstrom (2011)) (Fig. 9(b)) and the *Oslom* algorithm (Lancichinetti et al. (2011)) (Fig. 9(b)). The clustering provided by the later is actually overlapping *i.e.* a city can be part of several clusters.

As said previously, our results differ from the ZAUER classification. They are however closer than those obtained using other methods. Both *Infomap* and *Oslom* algorithms provide



(a) ZAUER classification (b) *Infomap* clustering

(c) *Oslom* clustering (from left to right: first, second and third level).

FIG. 9: Different clusterings for the region Pays-de-la-Loire. Only the groups corresponding to more than 5000 workers are displayed.

clusters of uniform sizes at the different levels. This configuration is very different from the ZAUER classification where large groups can be found around big cities. Moreover, relatively large groups can be found in rural areas where the number of commuters is low.

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 returning one solution. With our method however, the user (geographer or sociologist) is able to visually mine the network that he/she studies and explore various solutions. Combining edges filtering with the evolution of a suited quality measure provides an efficient method for the detection of dense clusters and hierarchies inside a network. It may also be used to classify networks base on the shape of the quality curve.

We enforced this procedure to study French commuters flows. It appears that hierarchies of cities can be found for several regions. The results provide a different kind of information than the ZAUER classification but are more consistent than those obtained using other hierarchical algorithms. Note also that the ZAUER classification takes into account the geographical

distance between cities while we only used the amount of workers travelling between cities as input. People working for the INSEE in charge of the ZAUER classification contacted us and were interested by this work. Moreover, geographers found our method and results relevant for the study of commuters flow networks [2]. We also claim that our procedure can be used to study and partition other kind of networks (weighted or not) such as social networks. However, the method proposed here focused on finding non-overlapping partitions.

An interesting lead to validate the choice of the threshold values is to use a hierarchical quality measure introduced in Delest et al. (2011). We should be able to tell whether the local maxima in the evolution of MQ corresponds to a good hierarchical clustering. Regarding our case study, we plan to use our method to extract hierarchical decompositions over the years using the previous national census data. We suspect that this approach can help the understanding of the dynamic of such networks. However, additional visualisation tools may be needed. For example, we should be able to highlight significant changes in the structure of the network represented by a hierarchical clustering.

References

Auber, D., D. Archambault, R. Bourqui, A. Lambert, M. Mathiaut, P. Mary, M. Delest, J. Dubois, G. Mélançon, et al. (2012). The tulip 3 framework: A scalable software library for information visualization applications based on relational data.

Auber, D., Y. Chiricota, F. Jourdan, and G. Melançon (2003). Multiscale visualization of small world networks.

Blondel, V., J. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 2008*, P10008.

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

De Montis, A., S. Caschili, and A. Chessa (2011). Commuter networks and community detection: a method for planning sub regional areas. *Arxiv preprint arXiv:1103.2467*. Anglais

Delest, M., G. Melançon, F. Queyroi, and J.-M. Fédou (2011). Assessing the Quality of Multilevel Graph Clustering. Technical report, LaBRI.

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

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

Hamers, L., Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, and A. Vanhoutte (1989). Similarity measures in scientometric research: the jaccard index versus salton's cosine formula. *Information Processing & Management 25*(3), 315–318.

Lancichinetti, A., F. Radicchi, and J. Ramasco (2011). Finding statistically significant communities in networks. *PloS one 6*(4), e18961.

2. this work was presented during the 2011 Spangeo (*Spatial Networks In Geography*) group meeting.

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

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

Ratti, C., S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. Strogatz (2010). Redrawing the map of great britain from a network of human interactions. *PLoS One 5*(12), e14248.

Rosvall, M. and C. Bergstrom (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one 6*(4), e18209.

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