Segmentation of towns using call detail records

Romain Guigourès* and Marc Boullé*
*Orange Labs, 2 av. Pierre Marzin, 22300 Lannion, France
romain.guigoures@orange-ftgroup.com, marc.boulle@orange-ftgroup.com

***Abstract*—In this paper, we deal with the segmentation of towns using call detail records. The data can be viewed as a directed bipartite graph wherein the source nodes correspond to the towns, origins of the calls, and the target nodes that are the towns, destinations of the calls. A nonparametric method based on a Bayesian Approach is proposed to determine the finest segmentation of these two sets. Instead of directly clustering the nodes, we propose here to make a coclustering on the edges defined as bidimensionnal items described by two features : the source and target nodes. Once the finest clustering is obtained, the clusters are successively merged on the two sets until only one cluster remains, in such a way that the loss of information is minimal. The initial segmentation is optimally coarsened in order to enable a hierarchical exploratory analysis of the data. Thus, it is possible to get insights either nationalwide or locally. A study of the telephone areas of Belgium by exploring the coclustering structure at different grain levels demonstrates the interest of the method.**

***Keywords*-Community detection; Clustering; Bayesianism; Model Selection; Density estimation**

I. Introduction

Graph partitioning has long been studied in the operational research field. One of the oldest approaches is the minimum-cut method, where the graph is divided into a predetermined number of disjoint subsets, usually of approximately the same size, chosen such that the number of edges between the clusters of nodes is minimized.

With the recent availability of many network data, such as world wide web, social networks, phone call networks, science collaboration graphs [1], there is a renewed interest for the graph partitioning problem, especially for the automatic discovery of community structures in large networks. Many approaches have been studied for the problem of graph clustering, including hierarchical clustering, divisive clustering, spectral methods, random walk [2]. To evaluate the quality of a clustering regardless of the cluster number, the modularity criterion proposed in [3] is now widely accepted in the literature, and has even been treated as an objective function in clustering algorithms [4]. The modularity is a measure ranging from -1 to 1, being all the more high that the clusters have more internal edges than the expected edges number if the connections were made randomly, with the same nodes degrees.

In this paper, we present a way of analyzing and summarizing the structure of large graphs, based on piecewise constant edge density estimation. The approach extends the stochastic block modeling approach [5] in that the modeling method is fully non-parametric with the number of clusters as a free parameter, and exploits a statistical model selection technique and scalable optimization algorithms. Data grid models [6] are applied to graph data, where each edge is considered as a statistical unit with two variables, the source and target nodes. The objective is to find a correlation model between the two variables, owing to a data grid model, which in this case turns to be a coclustering of both the source and target nodes of the graph. The cells resulting from the cross-product of the two clusterings summarize the edge density in the graph. The best correlation model is selected using the MODL (Minimum Optimized Description Length) approach [6], and optimized by the means of combinatorial heuristics with super-linear time complexity.

Then, a post processing technic is introduced consisting in merging successively the clusters in the least costly way, from the finest clustering model down to one single cluster containing all the towns. It appears that the cost of the merge of two clusters is a weighted sum of Kullback-Leibler divergences from the merged clusters to the created cluster which can be interpreted as a dissimilarity measure between the two clusters that have been merged. Thus, the post-processing technique can be considered as an agglomerative hierarchical clustering [7].

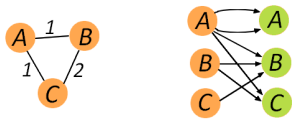
The rest of the paper is organized as follows. In Section 2, we present the MODL approach for data grid models applied to the edge density estimation in graphs and the postprocessing enabling the exploratory analysis. Then in Section 3, experiments on Belgian call detail records illustrate the property of the method. Finally, Section 4 concludes the paper and introduces the future works.

II. The segmentation

A. Graph clustering using MODL

Unlike making clustering on a simple graph like the modularity-based method do, the graph we deal with is directed, bipartite and with multiple edges. The source nodes are the towns, origins of the call, the target nodes the towns, destinations of the calls and the edges the calls. Figure 1 illustrates the different data representations.

The towns are grouped if the distributions of the calls are similar. This means that instead of making groups of towns that frequently call each other, the method brings together the towns that call the same towns and in the same



(a) (b)

	A	B	C
A	2	1	1
B	0	1	1
C	0	1	0

(c)

Figure 1: Representation of the tabular data displayed in Fig.(c) as a simple weighted graph (Fig.(a)) and as a directed bipartite multigraph (Fig.(b))

ratio. The objective of the method is to estimate the density of the edges owing to a coclustering of the sources and target nodes. Figure 2 illustrates such a coclustering with two source clusters and two target clusters. In this example, the probability of edges from Brussels or Liège to Brussels or Namur is 50%.

	Brussels	Namur	Liège	Charleroi
Brussels	1	1	0	0
Liège	1	1	0	0
Charleroi	0	0	1	1
Namur	0	0	1	1

	{Brussels,Namur}	{Liège,Charleroi}
{Brussels,Liège}	4	0
{Charleroi,Namur}	0	4

Figure 2: Example of coclustering

Formally, a model M of edge density estimation is defined by :

- the number of source and target clusters
- the partition of source (resp. target) nodes into source (resp. target) clusters.
- the edges distribution on the coclusters defined as the cross-products of the source and target clusters.
- for each source (resp. target) cluster, the edges distribution on the node of the cluster.

The coarsest model is based on one single cluster of towns, whereas the finest one exploits one cluster per town. Coarse grained models tend to be reliable, whereas fine grained are more informative. The issue is to find a trade-off between the informativeness of the edge density estimation and its reliability, on the basis of the granularity of the coclustering. Applying a Bayesian model selection approach, the best model M^* is defined as being the most probable given the data D, obtained by maximizing a criterion built on prior terms $P(M)$ which give priority to simple models, i.e. with a low number of clusters, and on the likelihood which favours the informative models, i.e. fine grained models :

$$M^* = \underset{M}{\operatorname{argmax}}\, P(M|D) = \underset{M}{\operatorname{argmax}} \left(\frac{P(M)P(D|M)}{P(D)} \right)$$

$$\Rightarrow M^* = \underset{M}{\operatorname{argmin}}(-\log P(M) - \log P(D|M))$$

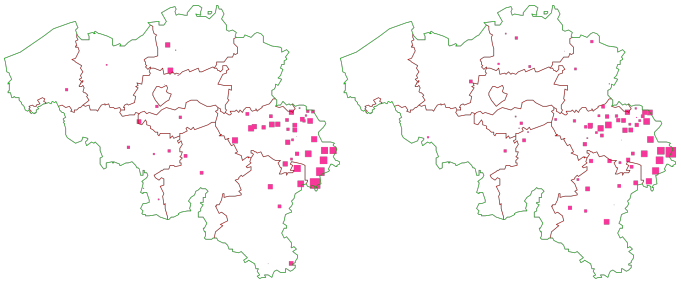
The detailed criterion is left out for brevity. The full criterion is described in [8]. As for the optimization algorithm, we have used the optimization heuristics detailed in [9], which have practical scaling properties, with $\mathrm{O}(m)$ space complexity and $\mathrm{O}(m\sqrt{m}\log m)$ time complexity, with m the number of edges. The main heuristic is a greedy bottom-up heuristic, which starts with a fine grained model, considers all the merges between clusters and performs the best merge if the criterion decreases after the merge. The process is reiterated until no further merge decreases the criterion. This heuristic is enhanced with post-optimization steps (moves of towns across clusters), and embedded into the variable neighborhood search (VNS) meta-heuristic [10], which mainly benefits from multiple runs of the algorithm with different random initial solutions. The optimization algorithms summarized above have been extensively evaluated in [9], using a large variety of artificial datasets, where the true data distribution is known.

B. *Merging the clusters*

In case of large datasets, i.e. with a huge number of edges, the edge density converges to the true edge distribution. This means that, for each town, the distribution of the calls is fine enough to be differentiated. Thus the method yields one town per cluster, that is too fine for an easy interpretation. To overcome this issue, a post-processing technique is proposed. It consists in merging successivly the clusters so as to worsen the least the criterion. By studying in detail the variation of the criterion due to the merge, it appears that the merge of two clusters is all the more likely that the distributions of their in/outcoming calls are similar. Figure 3 illustrates two towns very likely to be merged because of their similar distributions. Technically, this variation is a sum of Kullback-Leibler divergences from the merged clusters to the resulting one, weighted by the size of each of them. In brief, this process is equivalent to making a hierarchical agglomerative classification, whose dissimilarity measure is based on probability divergences.

III. Experiments

Experiments have been conducted on call detail records of the Belgian telecommunications company Mobistar aggregated on 6 months. There are 217 millions calls between 589 towns. Another approach has been applied on the same dataset in [11], which results in 17 clusters. Like in this study



(a) (b)

Figure 3: Similar distributions of the calls of two german-speaking Belgian towns.

our clusters are geographically connected. However our clustering results enable a multiscale exploratory analysis of the Belgian telephone areas.

A. The Finest Clustering

The finest clustering highlights 588 groups over Belgium. Hence, each cluster is made up of one town, except one cluster that groups two towns together. Given the huge number of calls (217 millions for 589 cities), the finest clustering on this dataset has reliably approximated the true distribution. This is shown on Figure 4, the clustering is all the more fine that there are edges.

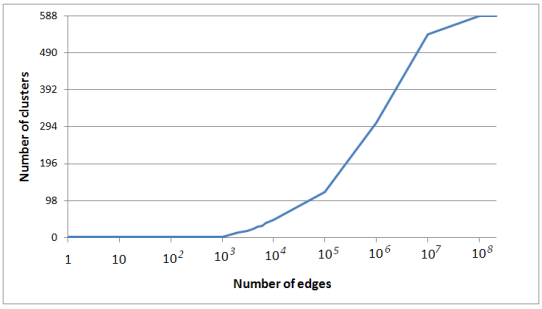


Figure 4: Number of clusters retrieved by the method for a given subset of randomly selected edges

B. Two linguistic communities

In this experiment, the clusters have been merged successivly until obtaining two clusters. This segmentation in two groups highlights the two linguistic communities of Belgium: Flanders and Wallonia. This reveals that the distribution of the calls of a town is denser in the areas with the same linguistic characteristics. The case of Brussels is particular, because the majority of the inhabitants of the city are french-speakers despite it is included into the Flemish territories. That is why the region of Brussels has been clustered into Wallonia.

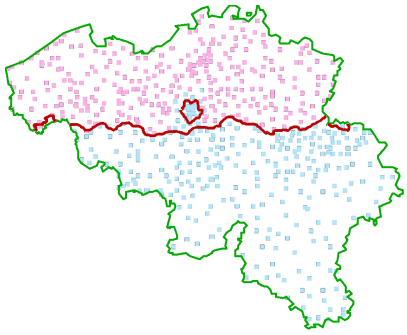


Figure 5: Segmentation of Belgium into two clusters

C. Eleven clusters that do not match with the provinces boundaries

There are eleven provinces in Belgium, five in Flanders and five in Wallonia, the eleventh being the province of Brussels-Capital. In order to compare the delimitation of the telephone areas and the boundaries of the provinces, we have studied the clustering with eleven clusters. The clusters are displayed on a map in Figure 6.

For Antwerp, East and West Flanders, the clusters fit well the provinces territories.

The provinces of Hainaut and Liège are splitted into three clusters. For the first one, it can be explained by the presence of some major cities in the same province (Mons, La Louvière and Charleroi). For the second one, we can notice the sphere of influence of Liège while the area east of the city corresponds to the arrondissement of Verviers where more than 25% of the inhabitants are German-speakers.

There are also clusters that straddles some provinces, like the cluster grouping the arrondissement of Leuven and the province of Limburg or the one grouping the province of Luxembourg and a part of the provinces of Namur and Liège. These telephone areas are consequently vast, that is why a finer and local study would yield enough clusters to make a relevant exploratory analysis.

The case of Brussels highlights the correlation between the calls and the sphere of influence of the city including a little part of the Flemish Brabant and almost all the Walloon Brabant. This can be explained by the current trend of expansion of the suburbs to a southern direction, towards Walloon-Brabant, the inhabitants of Brussels being attracted by more peaceful areas with the same linguistic characteristics [12].

D. A local study of Brussels Region

Brussels is a particular city. The capital of Belgium is very cosmopolitan. In spite of the predominance of the french-speaking community, it is included into the Flemish territories. A study of the calls from the towns of the province of Brussels-Capital to all the Belgian towns allows the segmentation of Brussels and its suburb according to the

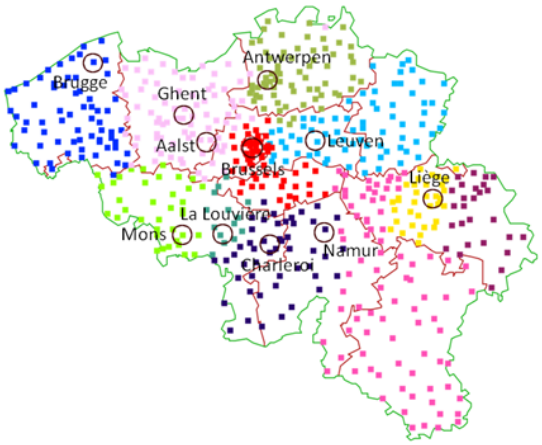


Figure 6: Segmentation of Belgium into eleven clusters

calls the users made all over Belgium. Merging until three clusters (over 19 towns) highlights interesting groups. The first group that is colored in pink in Figure 7 is located on the West side of the downtown and globally corresponds to the disadvantaged neighborhoods of Brussels while the green group highlights the privileged south-east quadrant of Brussels [12]. As for the two towns colored in Orange, Uccle and Ixelles, the higher education institutions are relativly concentrated there.

Figure 7: Segmentation of Brussels-Capital into three clusters

IV. Conclusion

In this paper, we have focused on graph clustering applied to a telephone dataset. The method allows the discovery of structures in graphs. By clustering the source and target nodes while selecting the best model according to a Bayesian approach, the method behaves as a nonparametric estimator of the edge density. In case of large graphs, the best model tends to be too fine grained for an easy interpretation. To overcome this issue, a post-processing technique is proposed. This technique aims at merging successivly the clusters until obtaining a simplified clustering while worsening the least the model.

Experimentations on a Belgian dataset show the variety of the possible analysis. The finest study yields almost one town per cluster. In order to make a global analysis, the model is coarsened by merging the clusters. With two clusters, the linguistic communities are well segmented while the province boundaries do not match with the clusters delimitations when the clusters are merged until eleven groups. Local groupings based on the call made all over Belgium are also possible and illustrated by the example of Brussels and its suburbs.

Because the method is based on a density estimation, the future works will be extended to the dynamic graphs by adding a third temporal variable. This would enable a study of the temporal evolution of social networks and yield the optimal discretization into time slots.

References

[1] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, p. 47, 2002.

[2] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27 – 64, 2007.

[3] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," 2008.

[5] S. Wasserman, G. Robins, and D. Steinley, *Statistical Network Analysis: Models, Issues, and New Directions*, 2007.

[6] M. Boullé, "A bayes optimal approach for partitioning the values of categorical attributes." *Journal of Machine Learning Research*, vol. 6, pp. 1431–1452, 2005.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, New York, 2002.

[8] M. Boullé, "Nonparametric edge density estimation in large graphs," Orange Labs, Tech. Rep., 2011.

[9] ——, *Data grid models for preparation and modeling in supervised learning*. Microtome, 2010.

[10] P. Hansen and N. Mladenovic, "Variable neighborhood search: principles and applications," *European Journal of Operational Research*, vol. 130, pp. 449–467, 2001.

[11] V. D. Blondel, G. Krings, and I. Thomas, "Regions and borders of mobile telephony in belgium and in the brussels metropolitan zone," *the e-journal for academic research on Brussels*, 2010.

[12] C. Kesteloot, C. Vandermotten, and B. Ippersiel, "Dynamic analysis of troubled neighbourhoods in the belgian urban regions," 2007.