Classification et visualisation de graphes avec SOMbrero

Madalina Olteanu, Nathalie Villa-Vialaneix

▶ **To cite this version:**

Madalina Olteanu, Nathalie Villa-Vialaneix. Classification et visualisation de graphes avec SOMbrero. Quatrièmes Rencontres R, Jun 2015, Grenoble, France. hal-01171557

HAL Id: hal-01171557
https://hal.science/hal-01171557

Submitted on 5 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Classification et visualisation de graphes avec SOMbrero

Madalina Olteanu[a] and **Nathalie Villa-Vialaneix**[b]

[a]SAMM (Statistique, Analyse et Modélisation Multidisciplinaire), EA 4543
Université Paris 1 Panthéon-Sorbonne, 90 Rue de Tolbiac, 75013 Paris, France
madalina.olteanu@univ-paris1.fr

[b]INRA, UR 0875 MIAT
BP 52627, 31326 Castanet Tolosan Cedex, France
nathalie.villa@toulouse.inra.fr

Mots clefs : fouille de graphes, cartes auto-organisées, classification non supervisée, projection non-linéaire, visualisation.

Récemment, les données structurées et notamment les graphes ont connu un intérêt croissant. Celles-ci ont en effet de multiples applications en sciences humaines et sociales, en biologie ou en informatique. Pour comprendre les structures complexes modélisées par les graphes, une approche courante consiste à combiner classification des sommets du graphe avec visualisation [5] : cela permet de mettre en valeur la structure macroscopique du graphe (classes denses de sommets et leurs relations) avant de se focaliser sur les détails à l'intérieur de telle ou telle classe. En particulier, le graphe est parfois représenté sur la base de la classification, de manière simplifiée : chaque classe est représentée par un méta-sommet, d'aire proportionnelle au nombre de sommets du graphe initial affectés à la classe correspondante, et les relations entre les méta-sommets sont représentées par des arêtes dont l'épaisseur est proportionnelle au nombre de liens entre les sommets des deux classes.

Les cartes auto-organisées (SOM) [3] sont une classe d'algorithmes de classification non-supervisée dont le principe est de projeter de manière non linéaire les données sur une carte de dimension faible (généralement une grille rectangulaire à deux dimensions) tout en préservant la structure topologique des données. Dans cette proposition de communication, nous proposons de présenter le package R **SOMbrero** (version 1.0, mise en ligne en mars 2015 sur le CRAN http://cran.r-project.org/web/packages/SOMbrero). Celui-ci contient une version stochastique de l'algorithme SOM adaptée à des données décrites par des matrices de dissimilarités, appelée SOM relationnel [4]. En particulier, le package permet l'utilisation de l'algorithme pour des graphes à partir de (dis)similarités calculées entre sommets (comme par exemple, la longueur du plus court chemin entre deux sommets) : celui-ci peut être mis en œuvre avec la fonction générique `trainSOM(..., type="relational")` qui produit un objet de type `somRes` contenant les résultats de l'algorithme SOM.

Cette proposition se focalisera donc sur l'utilisation de **SOMbrero** pour l'exploration de graphes et la recherche d'une représentation simplifiée comme décrite plus haut. En particulier, le package contient des fonctions permettant l'analyse et la représentation des résultats de l'algorithme SOM relationnel lorsque les données sous-jacentes sont modélisées par un graphe : 1) la fonction `projectIGraph.somRes` qui permet d'obtenir le graphe projeté sur la carte sous la forme d'un objet `igraph` (du package **igraph** [1]) : ce graphe a pour sommets les unités de la carte (localisées sur une grille rectangulaire à deux dimension) et pour arêtes le nombre de liens entre ces classes ; 2) la fonction `plot(..., type="graph", ...)` qui permet la représentation sur la grille du graphe projeté précédent (comme dans la Figure 1 (c)) ; 3) en combinant la

projection des sommets du graphe sur la grille avec une super-classification des prototypes (*ie* des représentants des unités), **SOMbrero** permet une représentation du graphe à l'échelle de super-classes positionnées au centre de gravité des unités de la super-classe. Cette fonctionnalité est disponible au travers des fonctions `superClass` et des fonctions prévues pour l'analyse des résultats de la super-classification, `projectIGraph.somSC` et `plot(..., type="projgraph", ...)`. La représentation obtenue est celle illustrée dans la Figure 1 (d). Nous illustrerons ses fonctionnalités au travers du jeu de données contenu dans le package et qui fait l'objet d'une vignette détaillée incluse dans celui-ci pour présenter l'algorithme SOM relationel : celui-ci est un graphe issu du roman de Victor Hugo « Les misérables » [2] (voir Figure 1).

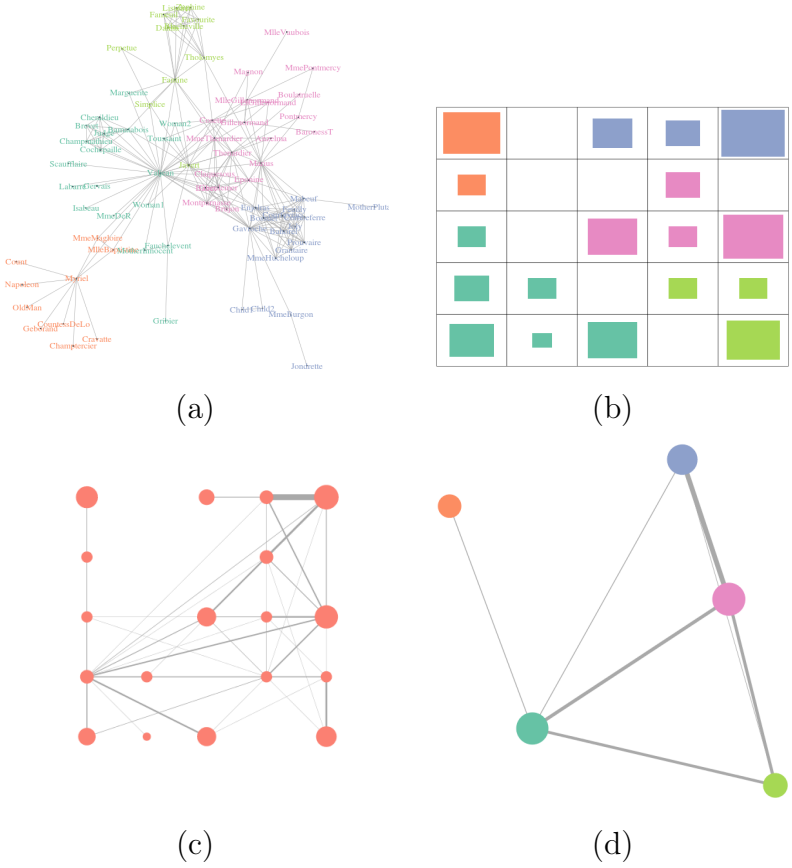

FIGURE 1 – (a) Le graphe initial issu du roman « Les Misérables », (b) La « hitmap » (dans (a) et (b), les couleurs correspondent aux super-classes), (c) Le graphe projeté sur la grille, (d) Le graphe projeté en utilisant les super-classes.

Références

[1] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems, 2006.

[2] D.E. Knuth. *The Stanford GraphBase : A Platform for Combinatorial Computing*. Addison-Wesley, Reading, MA, 1993.

[3] T. Kohonen. *Self-Organizing Maps, 3rd Edition*, volume 30. Springer, Berlin, Heidelberg, New York, 2001.

[4] M. Olteanu and N. Villa-Vialaneix. On-line relational and multiple relational SOM. *Neurocomputing*, 147 :15–30, 2015.

[5] F. Rossi and N. Villa-Vialaneix. Optimizing an organized modularity measure for topographic graph clustering : a deterministic annealing approach. *Neurocomputing*, 73(7-9) :1142–1163, 2010.