Méthodes de classification organisée pour la recherche de communautés dans les réseaux sociaux

Nathalie Villa[1,2] & Fabrice Rossi[3]

1– IUT STID, Carcassonne, Université de Perpignan
Domaine Universitaire d'Auriac, F-11000 Carcassonne – France
2– Institut de Mathématiques de Toulouse, Université de Toulouse - UMR CNRS 5219
118 route de Narbonne, F-31062 Toulouse cedex 9 – France
3– Institut TELECOM, TELECOM ParisTech, LTCI - UMR CNRS 5141
46, rue Barrault, 75013 Paris – France

Mots clés : Analyse des données - data mining, Données économiques et sociales
Keywords: Data mining, economics and social data

1 Résumé

Les graphes sont l'outil mathématique privilégié pour modéliser des systèmes dans lesquels les individus sont décrits par leurs interactions deux à deux. On les retrouve de manière naturelle dans l'étude des réseaux sociaux, des réseaux d'interactions biologiques, de l'internet... Ces graphes peuvent atteindre des tailles importantes et, au-delà d'une centaine de sommets, il devient difficile de comprendre leurs structures et de les visualiser de manière lisible. La recherche de groupes de sommets fortement liés dans un grand graphe et l'étude des relations existant entre ces groupes permet de donner une représentation simplifiée de la structure de grands graphes : une telle représentation est importante pour l'utilisateur final, sociologue, biologiste, historien ..., car elle permet de pouvoir appréhender de manière très intuitive le réseau social ou biologique modélisé [9].

Une solution classique consiste à s'appuyer sur une méthode de classification des sommets d'un graphe (voir [13] pour une revue complète des méthodes de classification de sommets d'un graphe). Les classes, que l'on peut voir comme un graphe simplifié, sont ensuite représentées par des méthodes classiques de visualisation de graphes [3]. Ici, nous nous proposons de présenter des approches plus directes, introduites dans [14, 2, 12].

Nous présenterons en fait deux types d'approche. La première est un algorithme de cartes auto-organisatrices (cartes de Kohonen, [4]) adapté à des données non vectorielles par le biais de noyaux. Cette approche, qui n'est pas spécifique à la classification de sommets de graphe, utilise un plongement implicite du graphe dans un espace de Hilbert à noyau reproduisant. Plusieurs versions de cet algorithme ont été développées [8, 1, 6, 15]. Dans [14], nous avons proposé une version "batch" de cette approche et l'utilisation d'un noyau spécifique aux sommets d'un graphe, le noyau de la chaleur [5].

Plus récemment [12], nous avons proposé une approche plus spécifique à l'organisation de sommets d'un graphe. Cette autre approche permet également de projeter les sommets du graphe sur une structure de type carte auto-organisatrice mais elle est basée sur l'extension d'un critère de qualité spécifique à la classification de sommets de graphes : la modularité [10]. Nous proposons d'optimiser directement une version "organisée" de la modularité par un algorithme de recuit déterministe [11, 7].

Enfin, nous présenterons des exemples de l'utilisation de ces méthodes d'organisation de sommets d'un graphe, exemples issus de réseaux sociaux réels : un réseau social venu du Moyen-Âge (Figure 1 gauche, représenté à l'aide du logiciel Tulip[1]) et un réseau de collaborations scientifiques (Figure 2 gauche). Les approches présentées permettent d'aboutir à une vision simplifiée du graphe comme le montrent respectivement, la Figure 1 droite, pour l'algorithme de carte de Kohonen à noyau et la Figure 2 droite, pour l'algorithme d'optimisation de la modularité organisée. En guise de conclusion, nous discuterons de quelques éléments de comparaison entre les deux méthodes.

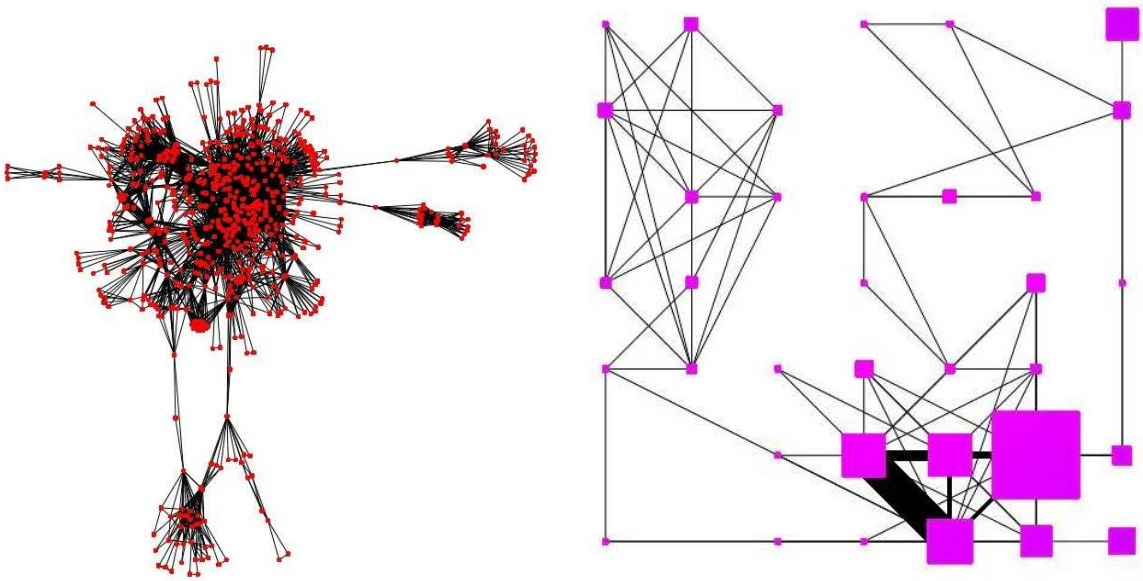

Figure 1: Un réseau social moyennageux représenté à l'aide du logiciel Tulip (à gauche) et sa simplification par carte de Kohonen à noyau (à droite)

2 Abstract

Graphs are natural mathematical tools to model real systems defined only by interactions between individuals: social networks, biological interaction networks, world wide web...

[1]Logiciel libre de visualisation de graphes, disponible à `http://tulip.labri.fr/`

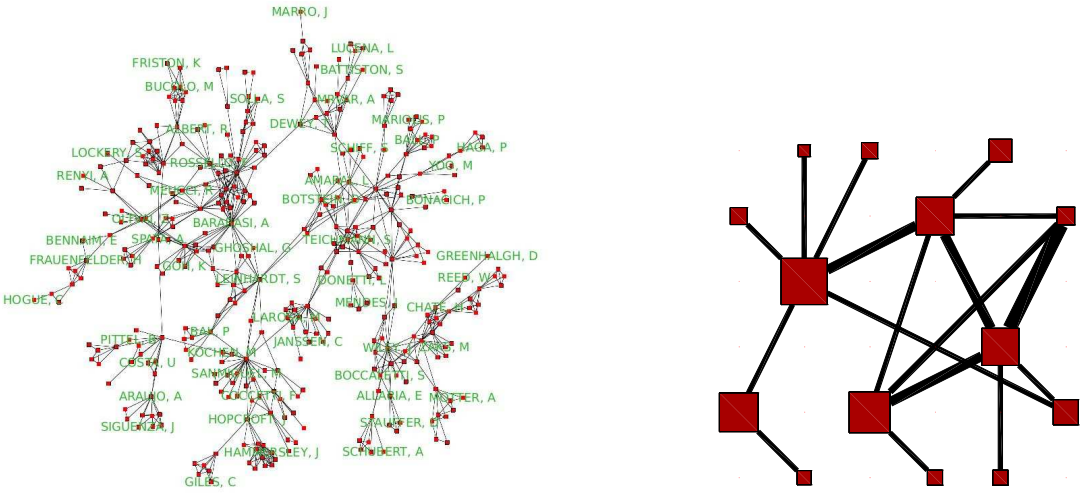

Figure 2: Un réseau de collaboration scientifique (à gauche) et sa simplification par optimisation de la modularité organisée (à droite)

They sometimes have a large number of vertices and up to one hundred vertices, understanding their structure or visualizing them in a meaningfull way are difficult problems. A solution to overcome this difficulty consists in finding sets of vertices that are highly connected to each others and to study the relations existing between these main groups. Such approaches can help the user (sociologist, biologist, historian, ...) as they provide very intuitive simplified representations of the studied network [9].

To that purpose, a possible solution is provided by graph clustering methods (see [13] for a complete review). The obtained clusters formed a simplified graph that can be rendered by usual graph visualization algorithms [3]. Here, we propose more direct approaches described in [14, 2, 12].

More precisely, we will present two algorithms. The first one, which is not specific to graph clustering, is an adaptation of self-organizing map algorithms (Kohonen algorithm, [4]) to nonvectorial data described by kernels. This approach uses an implicit mapping of the graph in a reproducing kernel Hilbert space. Several versions of this algorithm have be proposed in [8, 1, 6, 15]. In [14], we develop a "batch" version of this approach and we also suggest to adapt it for clustering the vertices of a graph by the use of a kernel that is specific to graph vertices: the heat kernel [5].

More recently [12], we proposed another approach which is more specific to organize the vertices of a graph. This method is also based on a prior structure similar to Kohonen maps and consists in extending a popular quality criterion for graph clustering (the modularity, [10]) in order to include organization soft constraints. We explain how to directly optimize this organized modularity by a deterministic annealing scheme [11, 7].

We will illustrate our method by real world examples in the field of social networks.

The first example (Figure 1 left, represented with Tulip software[2]) is a medieval social network and the second one (Figure 2 left) is a scientific collaboration network. The presented approaches lead to simplified representations of these networks: Figure 1, right, has been obtained by the use of kernel batch SOM and Figure 2, right, has been obtained by the optimization of the organized modularity. Finally, We will compare the results obtained by both approaches.

References

[1] P. Andras. Kernel-Kohonen networks. *International Journal of Neural Systems*, 12:117–135, 2002.

[2] R. Boulet, B. Jouve, F. Rossi, and N. Villa. Batch kernel SOM and related laplacian methods for social network analysis. *Neurocomputing*, 71(7-9):1257–1273, 2008.

[3] G. Di Battista, P. Eades, R. Tamassia, and I. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.

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

[5] R. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[6] K. Lau, H. Yin, and S. Hubbard. Kernel self-organising maps for classification. *Neurocomputing*, 69:2033–2040, 2006.

[7] S. Lehmann and L. Hansen. Deterministic modularity optimization. *The European Physical Journal B*, 60(1):83–88, 2007.

[8] D. Mac Donald and C. Fyfe. The kernel self organising map. In *Proceedings of 4th International Conference on knowledge-based intelligence engineering systems and applied technologies*, pages 317–320, 2000.

[9] M. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review, E*, 74(036104), 2006.

[10] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review, E*, 69:026113, 2004.

[2]Free software for graph visualization available at `http://tulip.labri.fr/`

[11] K. Rose. Deterministic annealing for clustering, compression, classification, regression and related optmization problems. *Proceedings of the IEEE*, 86(11):2210–2239, November 1998.

[12] F. Rossi and N. Villa. Topologically ordered graph clustering via deterministic annealing. In *Proceedings of ESANN 2009*, Bruges, Belgium, 2008. To appear.

[13] S. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, August 2007.

[14] N. Villa and F. Rossi. A comparison between dissimilarity SOM and kernel SOM for clustering the vertices of a graph. In *Proceedings of the 6th Workshop on Self-Organizing Maps (WSOM 07)*, Bielefield, Germany, September 2007.

[15] H. Yin. On the equivalence between kernel self-organising maps and self-organising map mixture density networks. *Neural Networks*, 19:780–784, 2006.