RSS flows, world structure & community detection

François Queyroi[*], Laurent Beauguitte[**], Hugues Pecout[***]

Abstract

We investigate the relevance of community detection methods from a thematic and a geographic perspective when applied to a network of co-citations in newspapers international RSS flows. We compare the results of different state-of-the-art algorithms. We find that building a consensus from the possible decompositions found is important in order to capture a macro global organization as well as local phenomena.

Introduction

The aim of the ANR *Corpus-Geomedia* is to collect RSS flows regarding international news issued from a world scale sample of daily newspapers. In this paper, we cross a thematic approach based on states co-citation and an analytic perspective regarding community detection in large networks.

Each RSS is made of several items and, for each item, we identify states quoted. Our hypothesis is that both the hierarchy of states and their co-presence is able to provide

[*] CNRS, UMR Géographie-cités
[**] CNRS, UMR IDEES
[***] CNRS, UMS RIATE

relevant information regarding world structures, notably regarding power and hot spots (Beauguitte *et al.*, 2014).

The network studied is built as follows: when two states are quoted together in at least one item, we consider that it creates a link between them. A link *e* is weighted by *W(e)*: the sum of *2 / (k(i)*(k(i)-1))* over all RSS items citing both states where *k(i)* is the number of countries cited in item *i* (i.e. the more an item refers to different states the less it strengthens the relation between two given states). As we consider a 9 months period, and to take into account the hierarchy of states in international news, each edge is pondered by the number of quotations of each state. The quotation of a state u is equal to *W(u)*: the sum of *1/k(i)* over all RSS items citing state *u*. So the formula which provides the intensity of an edge *e* between two vertex *u* and *v* is the following one: *W(e) /* (W(*u) +* W(*v)*)[1].

Our main interrogation in this communication is to test several algorithms of community detection in order to investigate their relevance on two complementary aspects: firstly the thematic and secondly the geographic relevance of partitions produced. A network community is classically defined as a set of nodes densely inter-connected – the concept of clique being its sociological equivalent (Queyroi *et al.*, 2014).

Four different methods were tested on a sample of 85 RSS flows from March to December 2014: Dominant flows, Louvain, Label Propagation and Markov Cluster Algorithm. We firstly compare the different partitions obtained in order to classify them; then we thematically examine partitions produced; and finally we create and analyze the consensus graph (i.e. graph including state-state links including states being classified together in more than one partition).

Community detection algorithms

Community detection gathers clustering methods aimed at reducing the complexity of real-world networks. It is done by decomposing the vertices of the graph into groups (called *communities*) that are relatively more connected between them than with the rest of the network. This task is achieved by trying to find a compromise between the proportion of *internal* edges (i.e. that link two vertices in a same group) and the number of groups.

Different attempt of formalizing this concept have been done in network science (and for each formalization, different algorithms has been proposed). Here we focus on partitioning methods: each vertex appears in one and only one group. We believe the different algorithms we use in this study (detailed below) represent a comprehensive subset of state-of-the-art methods.

First, we use a modularity maximization algorithm, often called Louvain (Blondel *et al.*, 2008), that is perhaps the most famous and most used in practical applications. It is known to produce large and homogeneous groups (even when it is not relevant from a network or a thematic point of view).

The algorithm Label Propagation (Biemann, 2000) emulates iteratively a process where each state chooses the group with which it has been co-cited the most. If it is an efficient method, this can lead however to non-relevant solutions (*e.g.* a partition with one group) with non-null probability. Louvain and Label Propagation are non-deterministic methods and should therefore be applied multiple time in order to check the robustness of a given solution. We enforce this strategy by keeping only partitions that are significantly different (at a variation of information distance of at least 0.2) after 1000 simulations.

Next, a method based on Dominant Flows (Nystuen and Dacey, 1961) is used (we call it *Dominant Flows Clustering*). It relies on a transformation of the network in a collection of trees (hierarchies). This is achieved here by taking, for each state *a*, the state *b* with whom the state is co-cited most often (if *a* is globally less cited than *b*). Although often used in flows analysis in geography, this method is highly sensitive to noises in edge weights and is clearly not the most robust method tested here.
The algorithm MCL (Markov Clustering for graphs) (Van Dongen, 2000) extracts groups which capture random walks in the network. Indeed, a random walker moving according to the edge weights will most likely stay a long time within the same community. This algorithm therefore strengthens important flows inside communities and weakens the others.

Results

Information about the different partitions is reported in Table 1. Notice that two and three different partitions can be found with Label Propagation (LP) and Louvain respectively although some are more likely than others (see 2nd column). Partitions found with Louvain tends to contains less groups (3rd column). The partition LP1 may correspond to a degenerate case of the algorithm where a big component containing most of the countries is found as shown by its low homogeneity (i.e. a measure between 0 and 1 assessing how balanced the groups are, 4th column). Dominant Flows Clustering (DFC), Markov Clustering (MCL) and LP2 have a smaller proportion of internal edges (5th column taking edge weights into account). This measure is mechanically greater for partitions with larger groups.

Table 1 - Various Statistics on different partitions found

Name Algo.	**% Found**	**Nb Comm.**	**Homogeneity**	**Citations Cover**
DFC	100	64	0.733	0.405
MCL	100	41	0.64	0.487
LP1	15	21	0.383	0.768
LP2	85	41	0.682	0.51
Louvain0	9	11	0.588	0.672
Louvain1	17	12	0.608	0.658
Louvain2	73	11	0.554	0.672

Those observations suggest a difference between the results obtain via Louvain and the others. This is confirm by Table 2 which reports the distance in term of *Variation of Information* (Kraskov *et al.*, 2005) between the different partitions. The decompositions DFC, MCL and LP2 are closer to each other and relatively far from the partitions found using Louvain algorithm.

However, this measure of distance hide the fact that partitions close to Louvain's can be recovered by aggregating groups in DFC, MCL and LP2. Equivalently, DFC, MCL and LP2 can be obtained by slitting communities found by Louvain leaving out less than 10% of the countries (9% for MCL). This suggests that a multi-level organization exists within the network.

Both the geographical and the thematic dimensions can be highlighted but at different level of precision depending on the method used.

Table 2 - Matrix of distance between the different partitions found. Bold represents noticeable low values. A value of 0 indicates that the two partition are identical. A value of 1 indicates they are completely different.

	DFC	MCL	LP2	LP1	Louvain 0	Louvain 1	Louvain2
DFC	0	**0.221**	**0.199**	0.627	0.464	0.481	0.453
MCL		0	**0.24**	0.65	0.427	0.428	0.451
LP2			0	0.606	0.408	0.398	0.401
LP1				0	0.692	0.686	0.702
Louvain 0					0	**0.22**	**0.221**
Louvain 1						0	**0.255**
Louvain 2							0

If Louvain produces giant components organized on a center-periphery basis (Figure 1, left), the three other methods provide clusters organized on a (sub)continental basis (Figure 1, right).

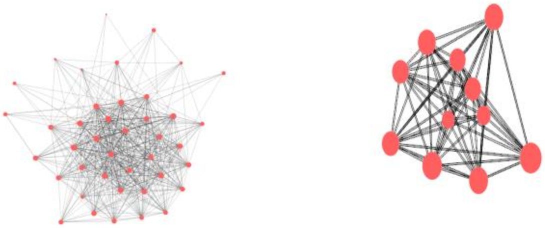

Figure 1 - Example of community found in Louvain 2 (left) and in MCL (right). Vertices sizes correspond to the relative proportion of items where the corresponding state appears in the dataset.

We build a consensus network by removing the state-state link that appears as internal in more than one partition (discarding LP1) pondering Louvain partitions by their chance of occurring. The resulting graph can be found in Figure 2 as a node-link diagram drawn using a force-directed algorithm. This visualization offers a good overview of the groups than are commonly found and the states that are "between" groups and cause the differences observed between partitions of equivalent size.

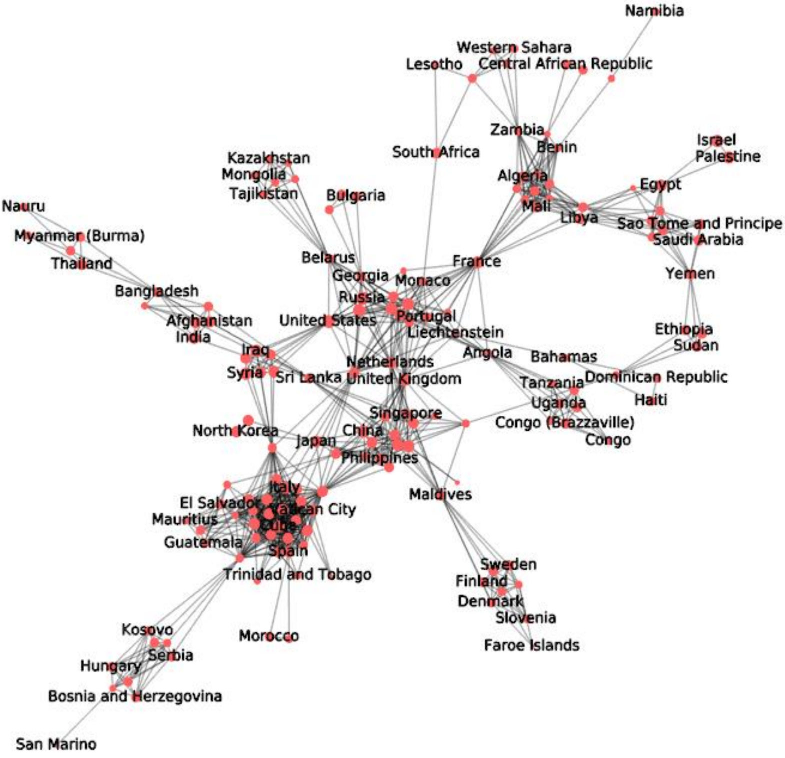

Figure 1 - Largest Component (165 states/209) of the Consensus Graph obtained after removing the edges that are internal in less than ¼ of the partitions (excluding LP1).

As we aggregated 85 RSS flows from daily newspapers on a world scale for a nine months period, interpreting each specific link would be a task of little interest. However, we can detect regional sub-groups more or less homogeneous: for instance, the clique involving ex-Yugoslavian states (Kosovo, Serbia, Bosnia and Herzegovina), an Asian one grouping Singapore, China, Japan and Philippines and several African ones (Tanzania, Uganda, Congo; Ethiopia, Sudan, Yemen; etc.). The presence of antagonist dyads (notably Ukraine-Russia and Israel-Palestine) was expected and these dyads appear whatever the algorithm chosen.

The situation appears less clear for most quoted and powerful states as they can be associated with a vast number of actors according to the news. For example, Europe is split in at least three parts: main political powers close to USA, with France being an articulation point towards African states; a Scandinavian bloc and, Italy and Spain clustered with South American countries (a catholic group?).

Going back to the precise content of RSS flows to determine what news created these links would be necessary to explain some specific configurations (Maldives – Scandinavian countries or Nauru - Thailand for example).

Conclusion

In this paper, we presented a case study of the application of community detection algorithms to a geo-mediatic network. We illustrate the importance of the use of various state-of-the-art algorithms in order to be able to capture the various features contained in the networks.

From a thematic perspective, our work shows that a given method of community detection should always be considered consciously as results present great variations. The consensus graph produced appears as one possible graph of states-interrelations in the World system, but further explorations are needed to validate its robustness, especially with a corpus as unstable as RSS flows: news, per definition, changes everyday, even if some elements regarding its production and its consummation can explain its global structure.

A further step would be to split the all corpus of RSS flows in order to check geographical fluctuation of world structure: for example, do Latin American newspapers give a different image of the world compared to European ones?

[1] Network data are available at https://sites.google.com/site/francoisqueyroi/datasets

References

Beauguitte L., Severo M., Pecout H. (2014), Do international news reflect world structure? A network approach*, First European Colloquium on Social Network Analysis*, Barcelona, July 1-4

Biemann C., (2006), Chinese whispers: An efficient graph clustering algorithm and its application to natural language processing problems, in Radev D., Mihalcea R., *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing,* Stroudsburg (PA), Association for Computational Linguistics, pp. 73-80

Blondel V., Guillaume J.L., Lambiotte R., Lefebvre E. (2008), Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, 10, pp. 10008-10020

Kraskov A., Stögbauer H., Andrzejak R.G., Grassberger P. (2005), Hierarchical clustering using mutual information. *Europhysics Letters*, 70(2), p. 278

Nystuen J. D. and Dacey M. F. (1961), A graph theory interpretation of nodal regions. *Papers of the Regional Science Association*, 7(1), pp. 29-42

Queyroi F., Delest M., Fédou J.M., Melancon G. (2014), Assessing the quality of multilevel graph clustering, *Data Mining and Knowledge Discovery*, 28(4), pp. 1107-1128.

Van Dongen S. (2000), A cluster algorithm for graphs, *Report-Information systems*, 20, pp. 1-40