IDENTIFICATION OF INTERNATIONAL MEDIA EVENTS BY SPATIAL AND TEMPORAL AGGREGATION OF NEWSPAPERS RSS FLOWS
Application to the case of the Syrian Civil War between May 2011 and December 2012

GIRAUD Timothée
CNRS - UMS 2414 RIATE
Timothee.Giraud@ums-riate.fr

GRASLAND Claude
Université Paris Diderot – UMR 8504 Géographie-cités
Claude.Grasland@parisgeo.cnrs.fr

LAMARCHE-PERRIN Robin
Université de Grenoble - UMR 5217 LIG
Robin.Lamarche-perrin@imag.fr

DEMAZEAU Yves
CNRS - UMR 5217 LIG
Yves.Demazeau@imag.fr

VINCENT Jean-Marc
Université Joseph Fourier - UMR 5217 LIG
Jean-Marc.Vincent@imag.fr

SUMMARY
The research project GEOMEDIA (ANR Corpus, 2013-2015) elaborates an international observatory of mediatized events, based on the collection of RSS flows feeded by 100 newspapers in French and English languages. The aim of this presentation is (1) to describe the complexity of the information contained in RSS flows according to space, time and media dimensions; (2) to derive basic solutions for the identification of international events on the basis of *time* aggregation procedures; (3) to analyze the *spatial* interactions between countries through an analysis of co-quotations in RSS flows; (4) to check the existence of interactions between time and space dimensions.

KEYWORDS
Space-Time Process, Data Aggregation, Newspaper, International Event, Syria

INTRODUCTION

The analysis of international events diffusion through media is a recurrent topic of research (Galtung & Ruge, 1965) that has benefit recently from the availability of very large database related to online media (newspapers, blogs, tweets…). Geography, Computer Science and Media Studies offer different but complementary points of view on a common field of research that we propose to call "*Geomediatic Analysis*". Contrarily to studies that explore the diffusion of news through abstract networks (e.g. Gomez-Rodriguez & al. 2013), our GEOMEDIA project tries to connect the analysis of media linkage to political theory and more precisely to theoretical studies on international relations (Battistella, 2003). Focusing on the analysis of international news (i.e. information published by a newspaper of a given country about other countries) we try to evaluate both direct flows (frequency of apparition of the country j in the news published by a media of the country i) and indirect association (frequency of association of countries j and k in the same news). The GEOMEDIA project is at a very early stage so that this article only investigates first directions of research combining methods developed in computer science and spatial analysis.

The empirical analysis focuses on the RSS flows of "international" or "world" news sent by four newspapers located in different countries: *Le Monde* (France), *The Times of India* (India), *The Washington Post* (USA) and *the Financial Time* (UK). More precisely, we analyze the news related to Syria between May 2011 and December 2012 to identify periods of more or less important international interest for this country. We also evaluate which countries are also mentioned in the news related to Syrian crisis in relation with local events (e.g. refugees in neighboring countries) and global diplomatic agenda (e.g. veto of Russia or China at UN against military intervention in Syria).

The paper is organized in four parts:

The first part presents the characteristics of the data under investigation and discusses the choice of a weighting criterion in the analysis of country co-quotations.

In the second part, we discuss the problem of international *events* identification – and more generally of international *agendas*. We propose a solution based on a procedure of time optimal aggregation grounded on information theory (*Lamarche-Perrin R., Vincent J.-M. and Demazeau Y., 2013*). This procedure is applied to the probability of apparition of "Syria" in the items sent by the RSS flows of the newspapers.

In the third part, we analyse the network of country co-quotations. Following a methodology previously applied to the vision of euro crisis by media (*Grasland C., Giraud T., Severo M., 2012),* we propose to visualize how different media have associated different countries in the same news. Focusing on the example of Syria, we evaluate the degree of autonomy of the network associated to these countries through a specific method of graph analysis – the Dominant flows.

The final and concluding part crosses the two previous dimensions and explores the changing links of country co-quotations according to the level of media focus on Syria.

1) HOW TO EXTRACT THE INTERNATIONAL INFORMATION CONTAINED IN RSS FLOWS?

We consider each international RSS flows of a newspaper as a **sensor** that publishes information regarding world events every day. This information is composed of small packages called **items** that contain two short strings of characters giving a title and a small description of the reported event. We extract from each item a **list of states** that have been recognized by the application of an ontology based on the name of the country, its inhabitants and its adjectives (*Table 1*).

Table 1: Examples of extraction of international information from items

Example 1 : item with 1 country quotation (Afghanistan)
RSS Flow : New York Times- World
Time : Monday 10th June 2013 à 06:22
Title : *Kabul Airport Attacked, **Afghans** Say*
Summary : *Gunmen and suicide bombers attacked Kabul International Airport before dawn on Monday, **Afghan** officials said.*

Example 2 : item with 3 country quotations (USA, Syria, Lebanon)
RSS Flow : Times of India- World
Time : Monday 10th June 2013 à 02:52
Title : ***US** close to deciding on arming **Syrian** rebels: Report*
Summary : *As many as 5,000 Hezbollah fighters are now in **Syria**, officials believe, helping the regime press on with its campaign after capturing the town of Qusair near the **Lebanese** border last week*

Example 3 : item with 5 country quotations (Germany, Switzerland, Austria, Hungary, Czech R.)
RSS Flow : Le Monde - International
Time : Friday 7th June 2013 à 21:01
Title : *Les inondations en Europe centrale coûteront plusieurs milliards d'euros*
Summary : *Les inondations qui frappent de vastes zones **d'Allemagne**, **Autriche**, **Hongrie**, **République tchèque** et **Suisse** devraient coûter des milliards d'euros en cultures gâchées, usines à l'arrêt, bâtiments ou infrastructures endommagés.*

The fact that the number of countries identified in items is not homogeneous introduces a difficulty in the measure of quotations weight. There are basically two possibilities: (1) we consider the number of countries' quotations, i.e. we assume that it is equivalent for a country to be mentioned alone or together with other countries in the same item; (2) we consider that the amount of information is inversely proportional to the number of countries mentioned in the item. We decided to apply the second approach because we consider that each item represent an atom of information with equivalent weight, whatever the number of coutries mentioned. In our data model it is possible to analyze not only the weighted frequency of countries but also the weighted frequency of linkage between countries when they are mentioned together in a given item. In this model, an item where k countries are mentioned will be transformed into k*k linkage between countries, each of them with a weight equal to $(1/k^2)$.

Table 2: Table of weight for countries mentioned in the items of Table 1

Flow (k)	Time (t)	Country (i)	Country (j)	Weight (Fijkt)
NYT-World	10/06/2013	AFG	AFG	1.000
TOI-World	10/06/2013	LIB	LIB	0.111
TOI-World	10/06/2013	LIB	SYR	0.111
TOI-World	10/06/2013	LIB	USA	0.111
TOI-World	10/06/2013	SYR	LIB	0.111
TOI-World	10/06/2013	SYR	SYR	0.111
TOI-World	10/06/2013	SYR	USA	0.111
TOI-World	10/06/2013	USA	LIB	0.111
TOI-World	10/06/2013	USA	LIB	0.111
TOI-World	10/06/2013	USA	USA	0.111

LM-Intern	07/06/2013	AUT	AUT	0.040
LM-Intern	07/06/2013	AUT	CHE	0.040
LM-Intern	07/06/2013	AUT	CZE	0.040
LM-Intern	07/06/2013	AUT	DEU	0.040
LM-Intern	07/06/2013	AUT	HUN	0.040
…	…	…	…	…
LM-Intern	07/06/2013	HUN	HUN	0.040

This format enables different types of aggregation concerning either isolated countries (i) or couples of countries (i,j). For example, Figure 1 presents the variation through time of the media weight of 6 countries in the international RSS flows of 4 newspapers (*Financial Times, Times of India, New York Times, Le Monde*) between May 2011 and December 2012.

Figure 1

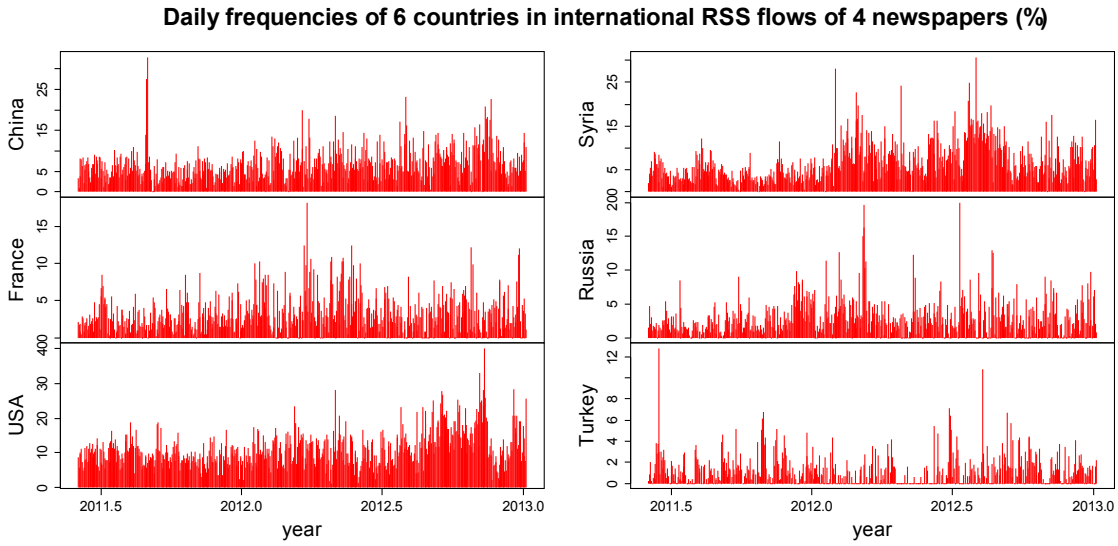
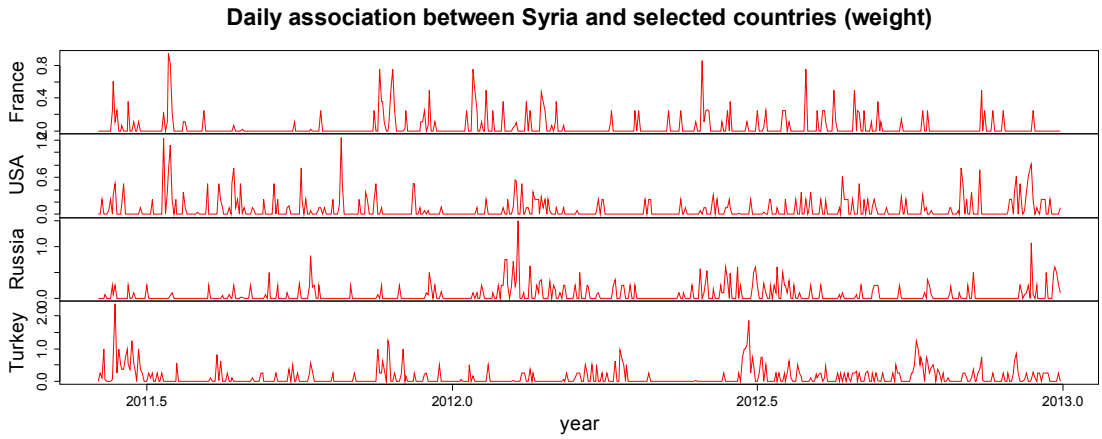


Figure 2 indicates the weighted number of items where Syria is associated to France, Turkey, Russia or USA

Figure 2



2) THE SITUATION OF SYRIA IN THE INTERNATIONAL AGENDA OF MEDIA BY TIME AGGREGATION

This section proposes different definitions of an international agenda based on the number of

country quotations in RSS flows. As our main concern is about Syria, we focus on the following questions "*What is the position of Syria in international news? Can we define periods of increasing or decreasing attention for this particular country? Is it possible to give a* global *representation of the newspapers agenda out of* daily *results?*" The identification of international events – and more generally of international *agendas* – is based on a procedure of time partition that optimizes measures from information theory to build consistent observation periods (*Lamarche-Perrin R., Vincent J.-M. and Demazeau Y., 2013*). This procedure is applied to the frequency of "Syria" quotations in the RSS flows of the four newspapers.

Figure 3 presents the frequencies of Syria quotations at the week level. The horizontal lines give the global frequency, *i.e.* the frequency of Syria quotations over the whole period of observation. It measures the global rank of Syria in the corresponding RSS flow agenda. The aggregation procedure consists in dividing the whole observation period in several sub-periods that fit with the two following requirements.

1. **The quotation frequencies should be homogeneous within the aggregated sub-periods, thus delimitating stable states of the newspaper agenda.**
2. **Breaks between successive sub-periods should mark important discontinuities in the time series, thus revealing crucial transitions in the agenda.**

The aggregation procedure proposed in (*Lamarche-Perrin R., Vincent J.-M. and Demazeau Y., 2013*) is based on the joint optimization of two dual information-theoretic measures. (1) The *Kullback Leibler divergence* quantifies the information regarding the detailed series (Figure 3) that is lost during the aggregation procedure. Minimizing the divergence consists in keeping details regarding heterogeneous time periods, in such a way that significant peaks are not suppressed. (2) The *size of the aggregated series* quantifies the information needed to encode the result of the aggregation. Minimizing the size of the aggregated series consists in gathering homogeneous periods to suppress redundant information and build stable periods of time. Aggregation thus consists in a trade-off between these two measures and can be performed at different level in order to adapt the granularity of the generated time series.

Figure 4 reports the results of the aggregation procedure. They correspond to the time partitions that preserve at least 70% of the information presented at the week level (Figure 3), while minimizing the partition size. Each time series gives an overview of the newspapers interests regarding the Syrian crisis and thus corresponds to significant media events with respect to the corresponding country (France, India, UK and USA). However, we can spot some global distinctive features:

The quotations of Syria in *Le Monde* and *The Times of India* seem to be very chaotic. Peaks and valleys irregularly alternate, indicating the time periods and the events which the newspapers are most interested in. These periods have different time scales (from one week to one month).

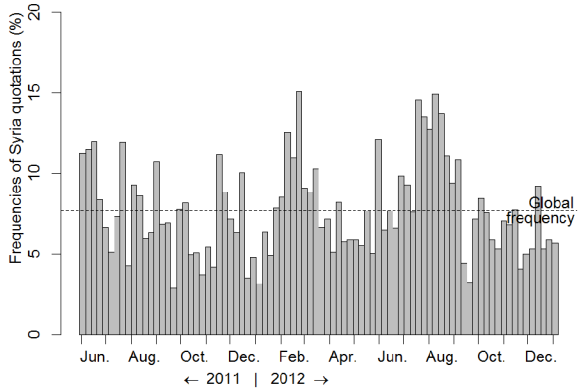
The agenda of *The Washington Post* shows two distinct periods: one from May 2011 to January 2012, where Syria is not a major topic in news, followed by a strong increase of Syria quotations, during one year, containing only two very significant peaks (in August and in December 2012).

The agenda of *The Financial Times* is more regular than the others. No very significant peak is detected.

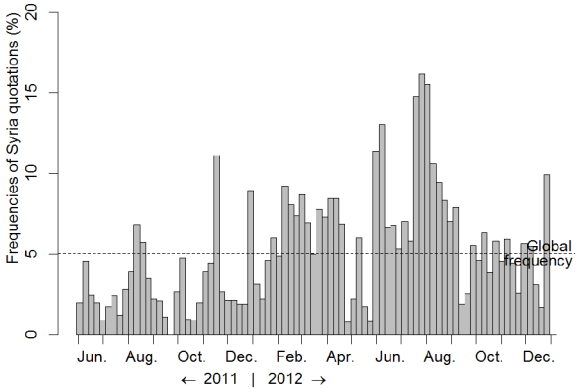
This comparison allows us to define global patterns (rythm of peaks, time scales, and overall variation on the observed period) and thus gives an abstract classification of newspaper agendas.

Figure 3: Weekly frequencies of Syria quotations in four RSS flows

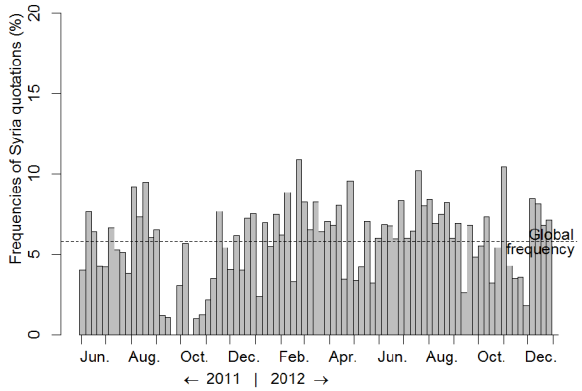
Le Monde



The Times of India



The Financial Times



The Washington Post

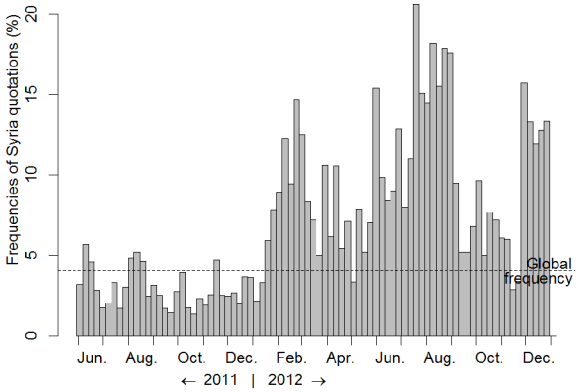
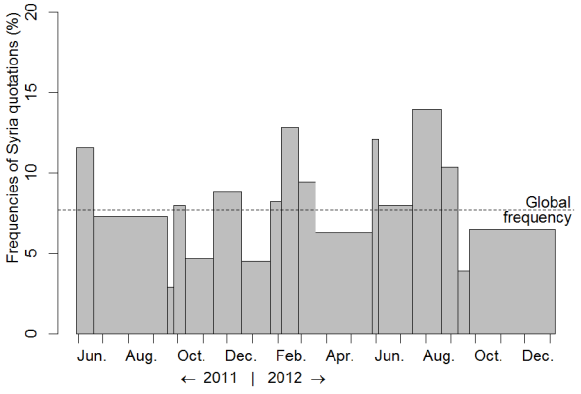
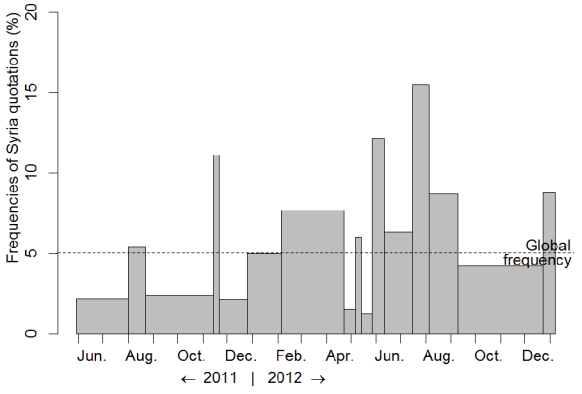


Figure 4: Optimal time aggregation of the frequencies of Syria quotations in four RSS flows

Le Monde

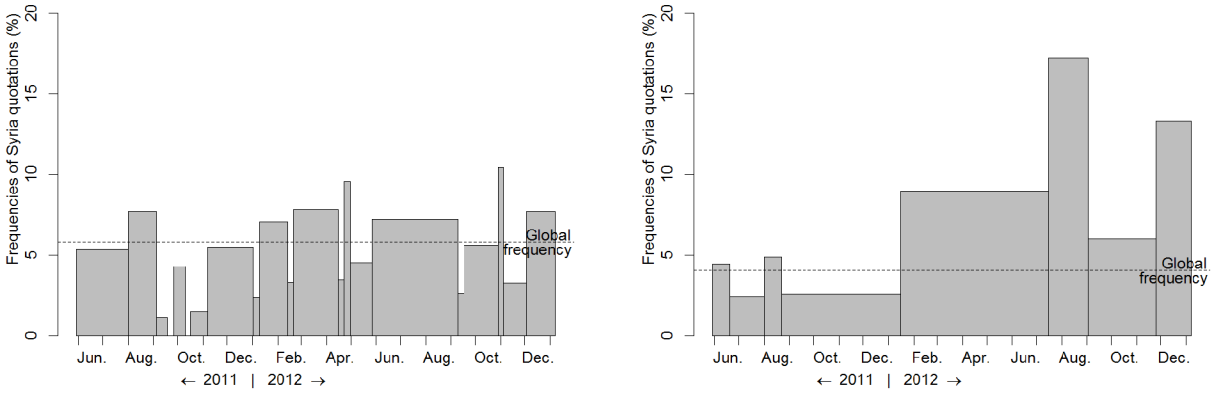


The Times of India



The Financial Times

The Washington Post



3) EVALUATION OF THE AUTONOMY OF SYRIAN CRISIS AS REGARD TO THE LINKAGES BETWEEN COUNTRIES

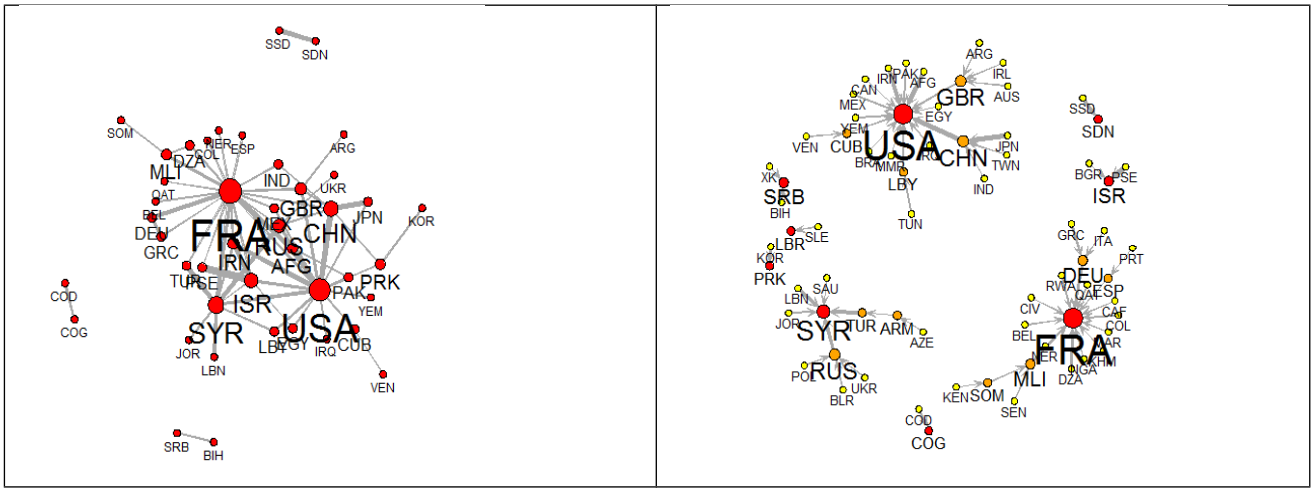
In the previous section we examined the variation of the **agenda of the Syrian Crisis** according to the four newspapers of our sample. The time aggregation procedure demonstrated that the focus on Syria does not correspond to the same time periods for the reader of an American, a British, a French or an Indian newspaper, even if some similarities are observed here and there.
In this section, we examine the variation of the **maps of international linkage of the Syrian Crisis** as revealed by the association of Syria with different countries in the same item. We assume that mentioning Syria together with one or several other countries is the sign of a relation established (voluntary or not) by the author and perceived (explicitly or not) by the reader. The network of co-quotations therefore defines a network of associated countries that can be interpreted as a geopolitical mental map of the conflict.

Following a methodology previously applied to the perception of euro crisis by financial media (*Grasland C., Giraud T., Severo M., 2012),* we analyze the countries that are associated with Syria and we establish a weighted network of co-quotations between countries which is equivalent to a matrix of flows. Then, this matrix is transformed into a hierarchical network using the algorithm of dominant flows (Nyusten J.,Dacey M.,1968) for the analysis of urban hierarchy. This algorithm is based on the application of the two following rules:

A spatial unit i is dominated by a spatial unit j if and only if:
(1) the most important flow from i is directed toward j;
(2) the sum of flows received by j is greater than the sum of flow received by i.

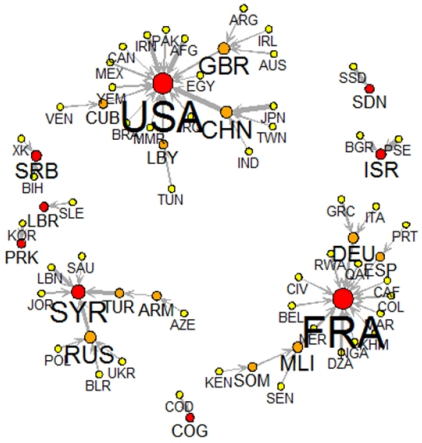
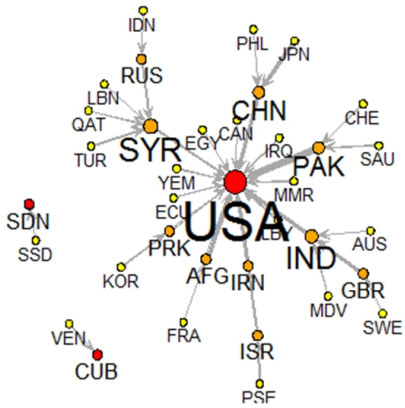
Figure 5: Result of the dominant flows method applied to items sent by the RSS flow of Le Monde in 2012

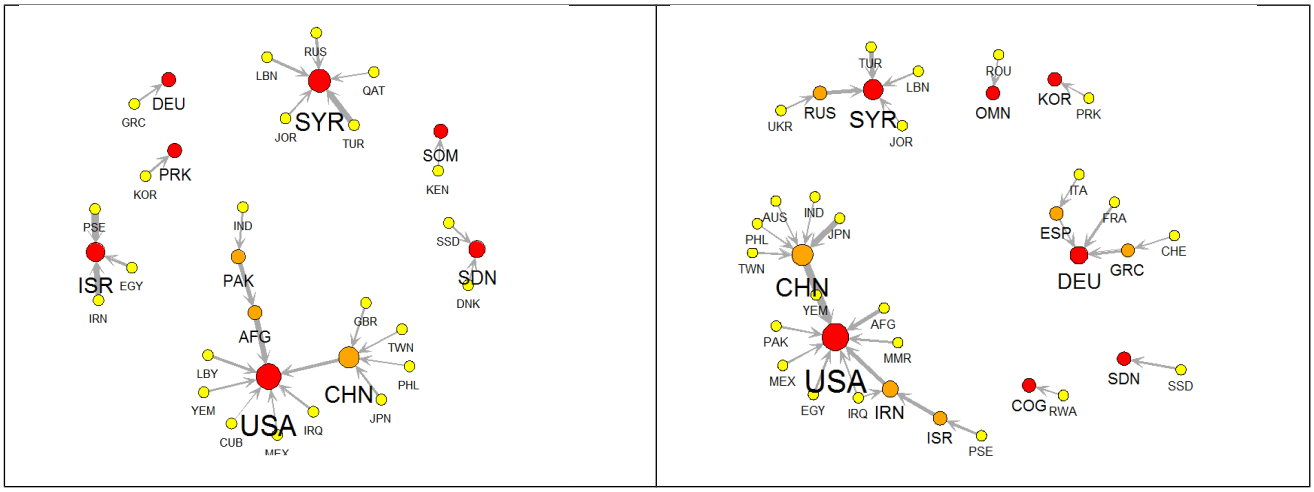
(a) Initial matrix of flows	(b) Graph of dominant flows



As presented in Figure 5, the dominant flows method creates a simplified version of the graph organized around dominant nodes (*in red*) which are characterized by the fact that their most important flow of co-quotation is directed toward a country that is less important in terms of media association. We also identify intermediate nodes (*in orange*) that are both dominated and dominant, and finally, purely dominated nodes (*in yellow*). The method therefore realizes a partition of countries in clusters of different sizes in which we identify major and minor components. The dominant countries of the most important components can be considered as the most powerful in terms of media linkages. In terms of "storytelling", they are associated with a lot of other countries but not with any country of greater importance. In our example, it is interesting to evaluate what are the different maps of international linkage of the four newspapers during the year 2012(see Figure 6).

Figure 6: International dominant networks of four international RSS flows of newspapers in 2012

(a) Le Monde – International	(b) *Times of India - World*
	
(c) *The Washington Post – World*	(d) *The Financial Times - World*



The structure of the major geopolitical components is clearly different in each newspaper. For the French journal *Le Monde*, we observe a division between two dominant states: France and USA. For *The Times of India*, the single major dominant state is clearly the USA. For *The Washington Post* and *The Financial Times* we have also a major component dominated by USA, but with a strong intermediate node representing China. Concerning the Syrian Crisis, we observe that in 3 out of 4 newspapers, Syria is a dominant state associated to its neighbors (Jordan, Turkey, Lebanon, …), but also to Russia and some Gulf countries (Saudi Arabia, Qatar). These networks are consistent with the fact that the Syrian Crisis appears as a conflict in which Western countries (France, USA, UK, …) are very reluctant to engage in, mostly due to the importance of the Syrian army, the risk of diffusion of the crisis to the whole Middle-East region, and the firm opposition of Russia (but also China) to any military supports for the rebels. It is certainly not a coincidence that *The Times of India* (which is not published in the "West") is the only newspaper of our sample where Syria is more systematically associated to USA.

4) CONCLUSION: TOWARD MULTIDIMENSIONAL AGGREGATION PROCEDURES

In previous sections, we examined the problems of aggregation according to time and space. However, this simplification is based on the over-optimistic assumption that these dimensions are independent. In future work, we will cross them and explore the changing links of country co-quotations according to the level of media focus on Syria, as introduced by the following simple example. For this example we used a time partition for *Le Monde* created using the previously described methodology. Figure 8 details the evolution of the co-quotation pattern over the whole period of observation according to the time periods defined by the aggregation procedure (periods *in red* in Figure 7).

Figure 7. Optimal time aggregation of the frequencies of Syria in* Le Monde *RSS flow

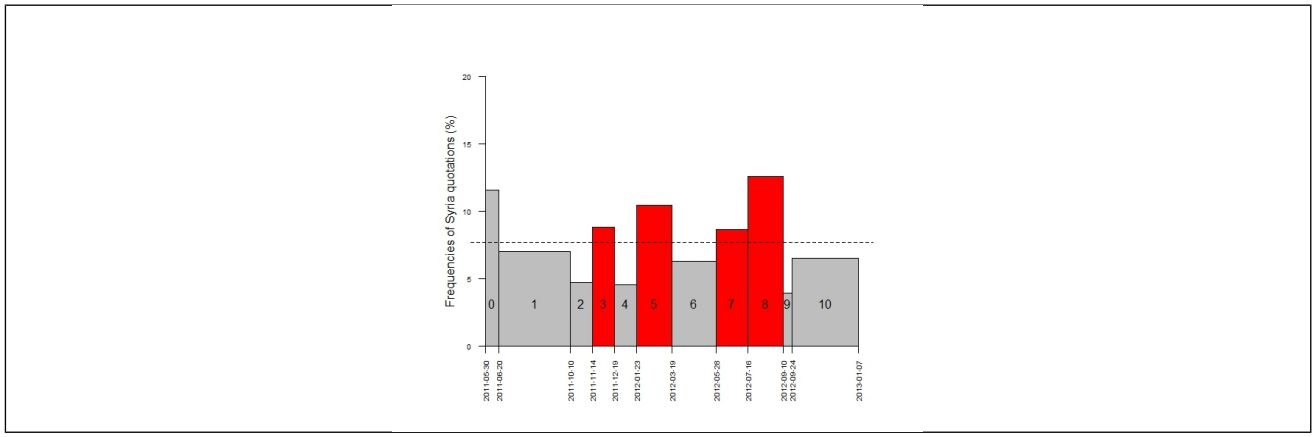
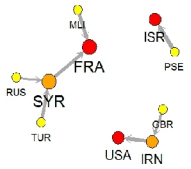
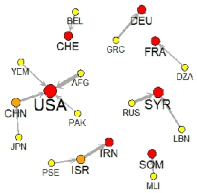
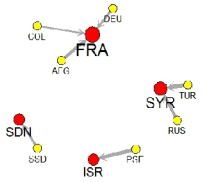
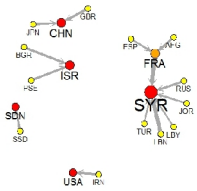


Figure 8: Graphs of dominants flows from* Le Monde *at the four red time periods

	
Period 3	Period 5
	
Period 7	Period 8

In each period, Syria is a dominant country and Russia is systematically associated to it. Besides that, there is no constant pattern. The situation of Syria varies from dominant-dominated (period 3), to dominant-isolated (period 5 and 7), and eventually to highly dominant (period 8). This shows the evolution of the media agenda related to the Syrian Crisis and the dominant-dominated relation with other countries according to *Le Monde*.

It is probably too early to turn these results toward empirical interpretation of the perception of the Syrian Crisis by the media. However, despite the limited sample of newspapers used in this experiment, we emphasized patterns revealing significant structures in time and space that could be of high interest for specialists of media and political studies. This encourages us to deepen the methods and develop large-scale harvesting tools leading to global analysis of international relations through media.

REFERENCES

Battistella D. , 2003, *Théories des relations internationales*, Presses de Science Po.

Galtung, J., & Ruge, M. H., 1965, « The Structure of Foreign News The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers ». *Journal of peace research, 2(1)*, 64-90

Gomez-Rodriguez M., Leskovec J. and Schölkopf B., 2013, "Structure and dynamics of information pathways in online media." *Proceedings of the sixth ACM international conference on Web search and data mining. ACM, 2013*.
Grasland C., Giraud T., Severo M., 2012, « Un capteur géomédiatique d'événements internationaux » in Beckouche P., Grasland C., Guerin-Pace F., Moisseron J.Y., in *Fonder les Sciences du Territoires*, Karthala, Paris.
Lamarche-Perrin R., Demazeau Y. and Vincent J.-M., 2013, "The Best-partitions Problem: How to Build Meaningful Aggregations?" *inProc of the 12th Conference on Intelligent Agent Technology, (IAT'13)* Atlanta.
Lamarche-Perrin R., Demazeau Y. and Vincent J.-M., 2013, "How to Build the Best Macroscopic Description of your Multi-Agent System? In Proc. *of the 11th International Conference on Practical Application of Agents and Multi-Agent Systems (PAAMS'13), LNAI 7879* Springer-Verlag, 2013.
Nyusten J. et M. Dacey, 1968, A graph theorie interpretaties of nodal regions, in *Spatial Analysis, a reader in statistical geography*, Berry et Marble (eds.),Prentice Hall, Englewood Cliffs, New Jersey