ANR CORPUS GEOMEDIA – Delivrable L.3.1 (M24)

Multidimensional Aggregation of RSS Flows

Robin Lamarche-Perrin
Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany
Robin.Lamarche-Perrin@mis.mpg.de

Yves Demazeau
CNRS, Laboratoire d'Informatique de Grenoble, France
Yves.Demazeau@imag.fr

Jean-Marc Vincent
Univ. Grenoble Alpes, Laboratoire d'Informatique de Grenoble, France
Jean-Marc.Vincent@imag.fr

Timothée Giraud
CNRS, UMS RIATE, Paris, France
Timothee.Giraud@ums-riate.fr

Claude Grasland
Univ. Paris Diderot, UMR Géographie-cités, Paris, France
Claude.Grasland@parisgeo.cnrs.fr

The database currently developed by the GEOMEDIA project consists in a huge collection of newspaper articles, along with automatised annotation tools conceived to extract geographical and temporal information out of the raw data. This database hence allows to represent and to analyse the media coverage of newspapers in space (at a national level) and in time (at a daily level). However, the analysis of international relations through media coverage requires to approach the data at many other representational levels, both in space (to see international events emerge from the aggregation of national data) and in time (because the dynamics of international relations can easily mix short term and long term events).

This research report presents an analysis method based on data aggregation and information theory to provide multilevel representations of geographical data. This approach has been developed especially for the data and problems addressed in the GEOMEDIA project, but also has many other applications in Artificial Intelligence (AI) for multilevel analysis of complex systems in general. This report hence contains two papers that have been published in 2014 in AI journals, summarising the research work on data aggregation that has been done within the GEOMEDIA project by computer scientists (Robin Lamarche-Perrin, Yves Demazeau and Jean-Marc Vin-

cent), and a third contribution in collaboration with geographers (Timothe Giraud and Claude Grasland) that has been presented in an international colloquium on quantitative geography. Lastly, a short video (4 minutes) has been made to briefly present our objectives and achievements to a broad audience of researchers. It hence constitutes another interesting introduction to our approach.

Paper 1, pp. 4–30.

- R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent. "Building Optimal Macroscopic Representations of Complex Multi-agent Systems. Application to the Spatial and Temporal Analysis of International Relations through News Aggregation". In: *Transactions on Computational Collective Intelligence*. Ed. by N.T. Nguyen, R. Kowalczyk, J.M. Corchado, and J. Bajo. Vol. XV. LNCS 8670. Springer-Verlag Berlin, Heidelberg, 2014, pp. 1–27.

In this paper, we present a general framework for data aggregation and its direct application to geographical data. Our main contribution is to use Shannon entropy and Kullback-Leibler divergence – two measures inherited from information theory – to respectively quantify the complexity reduction and the information loss induced by a given representation level. Data aggregation thus consists in achieving a trade-off between information and complexity, hence resulting in a multiresolution representation of the data, efficiently compressing homogeneous parts while preserving details about the irregularities (media events) appearing at different scales of space and time. The solutions to the corresponding optimisation problem hence provide different levels of understanding of the observed media events in order for the geographers to choose a representation level that is adapted to particular analysis objectives.

Paper 2, pp. 31–49.

- R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent. "A Generic Algorithmic Framework to Solve Special Versions of the Set Partitioning Problem". In: *Proceedings of the 2014 IEEE 26th International Conference on Tools with Artificial Intelligence (ICTAI'14)*. Limassol, Cyprus, 10$^{\text{th}}$-12$^{\text{th}}$ November 2014. Ed. by A. Andreou and G.A. Papadopoulos. IEEE Computer Society, Los Alamitos, CA, USA, 2014, pp. 891–897.

Increasing the representation level while preserving the information content is not sufficient to provide usable representations in geography. Indeed, domain experts usually rely on classical models of the territorial space to explain sociopolitical phenomena. In this second paper, we propose to integrate this *a priori* knowledge of the data structure within the aggregation process in order to generate representations that are consistent with the geographers' expectations. More generally, we present a class of combinatorial optimisation problems corresponding to special versions of the *Set Partitioning Problem* applied to different constraint structures on the solution set. We hence show that, beside multilevel aggregation of geographical and temporal data, our approach also benefits to a broad range of applications in Artificial Intelligence.

Paper 3, pp. 50–59.

- T. Giraud, C. Grasland, R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent. "Identification of International Media Events by Spatial and Temporal Aggregation of RSS Flows of Newspapers. Application to the Case of the Syrian Civil War between May 2011 and December 2012". In: *Proceedings of the 18th European Colloquium on Theoretical and Quantitative Geography (ECTQG'13)*. Dourdan, France, 5th-9th September 2013.

In this paper, our aggregation method is applied to the case of the Syrian Civil War in order to macroscopically describe the international agenda of four different newspapers. Time-series aggregation (using our approach) and spatial analysis of country co-citations (using dominant flow models) hence allow us to give a first description of the several steps constituting the media events related to Syria, and more generally the world vision that emerges from the data for the different newspapers.

Video

- R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent. "Multi-resolution Representations of Media Information". In: *Video Competition of the 23rd International Joint Conference on Artificial Intelligence (IJCAI'13)*. Beijing, China, 3rd-9th August 2013.

Building Optimal Macroscopic Representations of Complex Multi-agent Systems

Application to the Spatial and Temporal Analysis of International Relations through News Aggregation

Robin Lamarche-Perrin
Univ. Grenoble Alpes, Laboratoire d'Informatique de Grenoble, France
Robin.Lamarche-Perrin@imag.fr

Yves Demazeau
CNRS, Laboratoire d'Informatique de Grenoble, France
Yves.Demazeau@imag.fr

Jean-Marc Vincent
Univ. Grenoble Alpes, Laboratoire d'Informatique de Grenoble, France
Jean-Marc.Vincent@imag.fr

Abstract

The design and the debugging of large-scale MAS require abstraction tools in order to work at a macroscopic level of description. Agent aggregation provides such abstractions by reducing the complexity of the system's microscopic representation. Since it leads to an information loss, such a key process may be extremely harmful for the analysis if poorly executed. This paper presents measures inherited from information theory to evaluate abstractions and to provide the experts with feedback regarding the quality of generated representations. Several evaluation techniques are applied to the spatial and temporal aggregation of an agent-based model of international relations. The information from on-line newspapers constitutes a complex microscopic representation of the agent states. Our approach is able to evaluate geographical abstractions used by the domain experts in order to provide efficient and meaningful macroscopic representations of the world global state.

Keywords: Large-scale MAS, agent aggregation, macroscopic representation, information theory, geographical and news analysis.

1 Introduction

Because of their increasing size, complexity and concurrency, current multi-agent systems (MAS) can no longer be understood from a microscopic point of view. Design, debugging and optimization of such large-scale distributed applications need tools that proceed at a higher representational level by providing insightful abstractions regarding the system's global dynamics. Among abstraction techniques (dimension reduction, subsetting, segmentation, clustering, and so on [4]), this paper focuses on *data aggregation*. It consists in losing some information regarding the agent level to build simpler – yet meaningful – macroscopic representations. Such a process is not trivial for the interpretation of the data by the observer. In particular, unsound aggregations

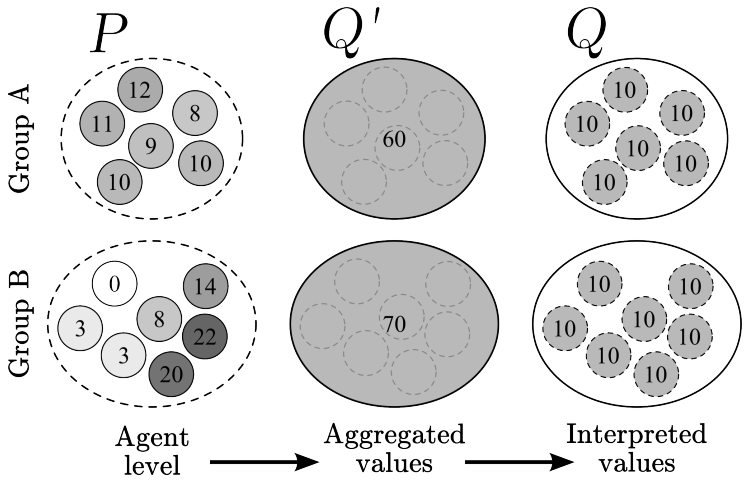


Figure 1: Averaging the behavior of groups of agents may reduce the redundant information (group A) or may lead to an unwanted information loss (group B)

may lead to critical misrepresentations of the MAS behavior. Hence, we need to determine what are the *good* abstractions and how to properly use them. At each stage of MAS development, aggregation processes should be carefully monitored and feedback should be provided regarding the quality of the generated macroscopic representations.

A simple example can demonstrate how critical the aggregation process can be. Fig. 1 shows two groups of agents that are simplified by two abstract entities with an average behavior. Intuitively, group A constitutes a *good* abstraction since the induced global behavior is relatively similar to the microscopic one, unlike group B. Hence, in order to scale-up, aggregation of redundant information should be encouraged to reduce the representation complexity (group A), but details regarding heterogeneous behaviors should be preserved in order to control the information loss and proceed to a sound analysis (group B).

Very little work has been done in the MAS community to quantify such aggregation properties. The main contribution of this paper consists in introducing measures from information theory (Kullback-Leibler (KL) divergence [11] and Shannon entropy [17]) to clarify the notion of *good* aggregation. From these measures, we provide generic feedback techniques and an algorithm that builds multi-resolution representations out of hierarchically organized MAS. These techniques and algorithms are applied to the agent-based modeling of international relations: agents represent countries, and their behavior is extracted from on-line newspapers. Geographers exploit multi-level aggregates to build statistics regarding world areas. We show how these geographical abstractions should be used to better understand the system states and its evolution through time.

Section 2 presents the work related to the main concern of this article. Section 3 presents the agent-based model of the GEOMEDIA application. Sections 4 and 5 introduce KL divergence and the size of representations to respectively estimate *information*

loss and *complexity reduction*. Section 6 shows how these measures can be combined to identify *best* aggregations and to build multi-resolution representations of hierarchically organized MAS. Section 7 applies these aggregation techniques to the time dimension in order to provide macroscopic representations of the system's dynamics.

2 Related Work

Aggregation can take place in every stage of the MAS development: from its design to its use. Even if abstraction techniques may differ, each stage should carefully take into consideration the quality of the provided aggregations. First, from a software perspective, this section shows that very few research efforts have been done to tackle this issue. (1) Most classical simulation platforms and monitoring systems do not even provide the user with abstraction tools; (2) some do handle the issue, but are still at an early stage of thought. Secondly, on a theoretical aspect, this section explains why classical techniques (*e.g.* data clustering, graph analysis) are not entirely satisfying to build consistent abstractions.

In a comprehensive survey of agent-based simulation platforms [16], Railsback *et al.* evaluate some tools by testing classical features of MAS modeling and analysis. Unfortunately, the abstraction problem is not tackled by this survey, thus indicating that such considerations are seldom if ever taken into account. Most platforms (Java Swarm, Repast, MASON, NetLogo and Objective-C Swarm) are limited to the microscopic simulation of agents. Railsback warns about the lack of "a complete tool for statistical output" in these platforms [16]. The provision of global views on the MAS macroscopic behavior thus constitutes an on-going research topic. Some tools for large-scale MAS monitoring however address this issue. For example, in some debugging systems, abstractions are used to reduce the information complexity of execution traces; however, they are either limited to the simplification of agents internal behavior, and do not tackled multi-agent organizational patterns [21], or they are provided without any feedback regarding their quality for the analysis [1, 19].

Some techniques from graph analysis and data clustering build groups of agents out of their *microscopic properties* (see for example [18, 8, 15]). Such considerations may meet ours from a theoretical point of view, but the approach presented in this report supports a very different philosophy: *abstractions should be built regarding some macroscopic semantics.* We claim that, to be meaningful, the aggregation process needs to rely on exogenous high-level abstractions defined by the experts. Hence, our approach should rather be related to researches on multi-level agent-based models [6]. These works openly tackle the abstraction problem by designing MAS on several levels of organization according to expert definitions. Such approaches aim at reducing the computational cost of simulations depending on the expected level of detail. The algorithm and measures presented in this report may provide a formal and quantitative framework to such researches.

To conclude, aggregation techniques should be more systematically implemented on MAS platforms in order to handle large-scale systems. They should combine consistent macroscopic semantics from the experts and feedback regarding the abstractions quality. For example, in this paper, abstractions used by geographers are evaluated

according to their information content.

3 Agent-based Modeling of International Relations

This section presents the GEOMEDIA agent-based model. It consists in the microscopic representation of countries with agents and the macroscopic representation of world geographical areas with groups and organizations.

3.1 Microscopic Data: the agent level

Let A be a set of agents constituting the MAS microscopic level. Visualization tools aim at displaying and explaining the properties of these agents: their behavior and internal states, the events they are associated with, the messages they exchange, and so on. Given a variable v that expresses such properties, the set of values $\{v(a)\}_{a \in A}$ constitutes the *microscopic representation* of the system (illustrated by distribution P in Fig. 1).

The GEOMEDIA project[1] is interested in the analysis of world international relations through a media point of view. This project is conducted in collaboration with geographers and media experts from the CIST (*Collège International des Sciences du Territoire*, Paris). In that context, we make the assumption that citations or co-citations of countries, within news, are good indicators to represent and understand their political, economical and cultural relations. For example, we may assume that an often-cited country is likely to politically interact with the newspaper country. Our agent-based model has two dimensions:

- The agents of the model represent the $|A| = 193$ United Nation member states, selected by geographers depending on their significance for the analysis of international relations.
- The temporal dimension contains $|T| = 90$ weeks, from the 3rd of May 2011 to the 20th of January 2013. This preliminary aggregation to the week level aims at reducing the chaotic variations of the day level and to focus on the more significant variations related to media events.

The experiments presented in this paper focus on a very basic variable: the number of articles that cite a country during a given time period. We use the 59,234 articles published by the "world" RSS flow of *The Guardian*[2] during the analyzed period and stored in the GEOMEDIA database. For each article, we look for the occurrences of the country names, the country adjectives, and the inhabitants names (*e.g.*, "Spain", "Spanish", and "Spaniard(s)" for the `Spain` agent). Thus, for each agent a and time period t, we count the number of articles $v(a, t)$ that "cite a during t". A total of 138,811 citations have been found within the dataset, distributed within 77% of the

[1] Project from the *Agence Nationale de la Recherche* (ANR-GUI-AAP-04). See the dedicated website: `http://geomediatic.net/`

[2] `http://www.theguardian.com/world`

articles (3 citations/article in average if we set aside the 23% that contain no citation at all).

In order to spot critical aspects of the international systems, geographers are interested in detecting significant events in the news. Such events correspond to unexpected values of the variable according to the following hypothesis: **the citation numbers of countries are homogeneous through time.** In that sense, the marginal values of the dataset give the expected citation number. For an agent a and a time period t, we thus expect the observed value $v(a,t)$ to be close to an expected one $v^*(a,t)$, defined as follows:

$$v^*(a,t) = \frac{v(a,.)\ v(.,t)}{v(.,.)}$$

where:

- $v(a,.)$ is the citation number of agent a on the whole observation period T;
- $v(.,t)$ is the total citation number – regarding all agents in A – during the time period t;
- $v(.,.)$ is the total citation number within the dataset.

A media event thus correspond to a high observed value $v(a,t)$ compared to the expected value $v^*(a,t)$. Fig. 2 displays the *observed-to-expected ratio of citation number* $v(a,t)/v^*(a,t)$ for each country $a \in A$ during $t =$ "the month of July 2011". A detailed survey of this map allows to identify geographical areas that have been unexpectedly over-cited during this time period: at the national level (*e.g.*, Norway, Djibouti, Guinea-Bissau) and at higher levels, *i.e.* for groups of countries (*e.g.*, Europe, Horn of Africa).

The quantity of information displayed in such a microscopic representation makes it quite hard to read. In particular, the visual clutter in dense areas prevents the proper interpretation of data. To overcome this difficulty, Fig. 3 and 4 propose to focus on areas of particular interest (resp. Europe and Africa). In the following sections, we will detail two particular events that occurred in these geographical areas:

1. The observed citation number of the `Norway` agent is 3.7 times higher than expected (Fig. 3). This is explained by the terrorist attacks that occurred in Norway the 22th of July 2011[3]. This event belongs to the national level and thus constitutes a *microscopic event* within the system's spatial dimension.

2. Countries of the `Horn of Africa` area also present unexpected citation numbers (from 1.9 times to 3.4 times the expected value for `Rwanda`, `Sudan`, `Somalia`, `Ethiopia` and `Djibouti`). This is explained by the food crisis that has been reported in this world area starting from the beginning of July 2011[4]. Unlike the previous one, this event is not located at the national level, but regards a group of agents located in a spatially spread out area.

[3] `http://en.wikipedia.org/wiki/2011_Norway_attacks`
[4] `http://en.wikipedia.org/wiki/2011_East_Africa_drought`

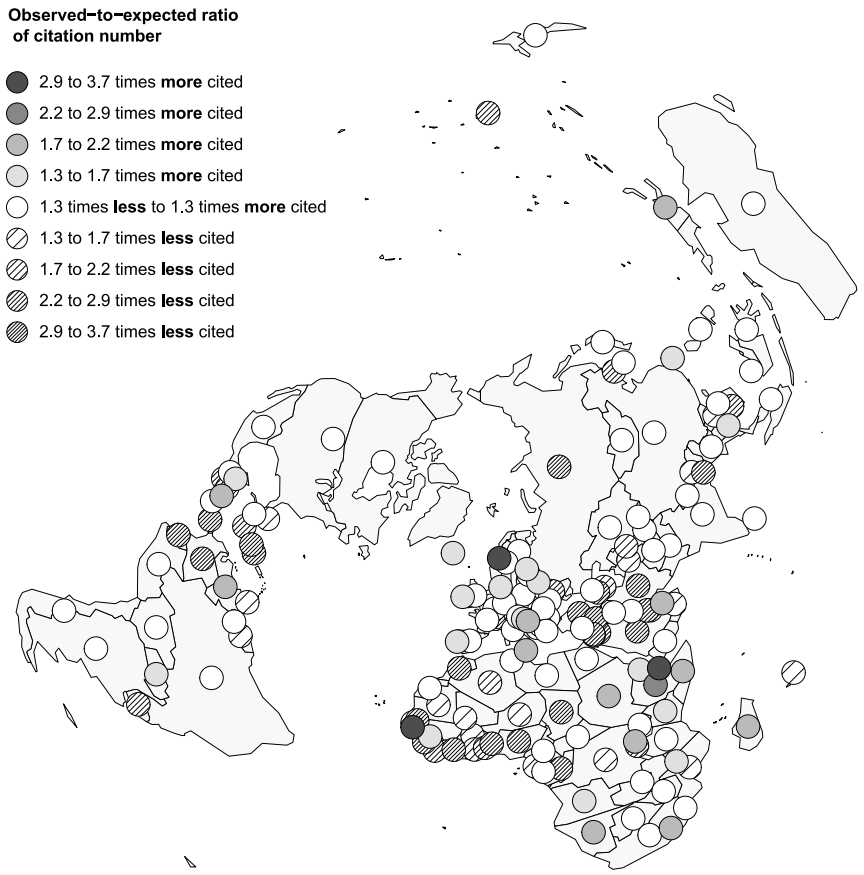


Figure 2: Observed-to-expected ratio of countries citation number in the articles published by *The Guardian* during July 2011

3.2 Macroscopic Data: groups and organizations

Even if the maps in Fig. 3 and 4 allow to easily spot the two events we are interested in, they do not manage to give the global overview of the world-wide system that is necessary for an informed analysis. *Data aggregation* aims at resuming the microscopic information to provide such an overview.

A *group* $G \subset A$ is subset of agents that are members of a consistent organizational pattern. It can be interpreted as an *abstract entity* that sums up the behavior of its underlying agents. Hence, groups satisfy a recursive definition: a group is either an agent or a set of groups. Quantitative variables – expressing *agents* properties – may be extended on *groups* according to an aggregation operator: sum, mean, median, extrema, and so on [4]. In our case, since we work with *extensive* variables (*i.e.* variables that are proportional to the aggregate size), $v(G, t)$ is defined as the *sum* of the values of

Figure 3: Observed-to-expected ratio of citation number (zoom on European countries)

the group G underlying agents (see distribution Q' in Fig. 1):

$$v(G,t) = \sum_{a \in G} v(a,t)$$

We define an *organization* O as a set of groups that constitutes a *partition* of the agent set A. Thus, in the scope of this paper, each agent is always a member of one and only one group. The set of group values $\{v(G,t)\}_{G \in O}$ composes a *macroscopic representation* of the system with respect to a given organization. It simplifies the variable distribution, from the detailed microscopic representation (distribution P in Fig. 1) to an aggregated one (distribution Q').

In order to be consistent with the observer's background knowledge, groups and organizations should be derived from the structural and semantical properties of the agent space. In our context, the world's *social*, *political*, and *economic* organization is used by geographers to represent and explain the data. Moreover, in this paper, we also focus on the world's *topological* organization in order to be consistent with classic geographical representations. Groups thus aggregate nearby territories that share a cultural and historical background. In the following experiments, we consider two hierarchical organizations of countries that meet these needs, namely `WUTS` [7] and `UNEP` [20]. Such organizations define multilevel nested groups commonly used by geographers to build global statistics regarding world areas, from the microscopic level

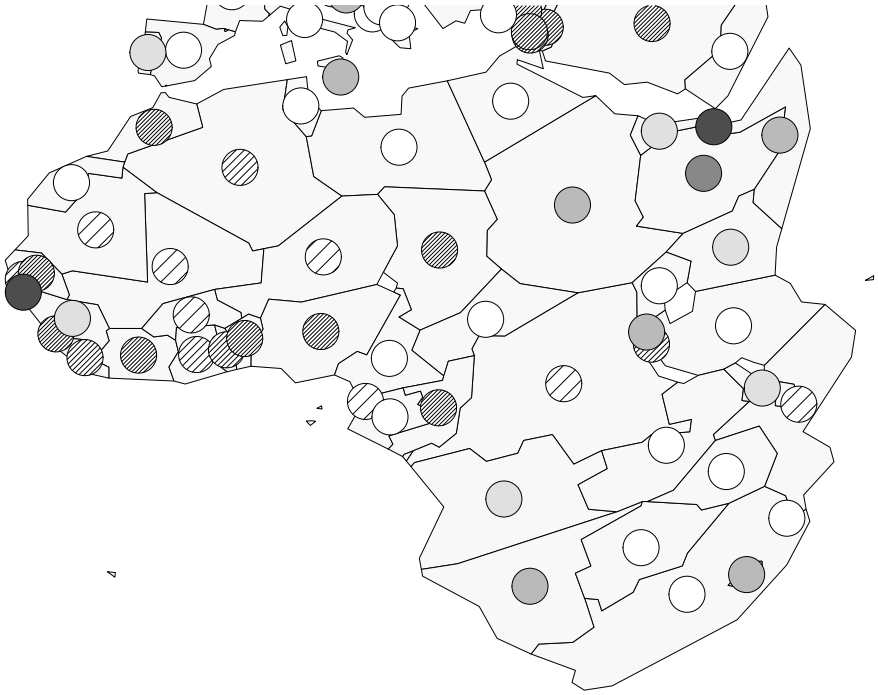


Figure 4: Observed-to-expected ratio of citation number (zoom on African countries)

of agents to the full aggregation (see [7] for a detailed presentation of these multiscale organizations).

As an example, Fig. 5 gives the observed-to-expected ratio of citation numbers aggregated according to the 3rd level of the `WUTS` hierarchy (or `WUTS_3`). Because of the data reduction, this map is much easier to analyze than the microscopic one (see Fig. 2). In particular, the food crisis that occurred in the `Horn of Africa` group is resumed by one observed macroscopic value that is globally 1.8 times higher than the expected citation number. In that case, the aggregation macroscopically represents the corresponding event. However, most of the microscopic variations have been suppressed by the aggregation process: for example, the events that occurred in the `Norway` agent are no longer represented. We thus need to control the aggregation process in order to visualize events at different levels depending on their spatial granularity. The following sections present an aggregation technique to automatically build such multiresolution representations of MAS.

4 KL Divergence as a Measure of Organization Quality

When an observer tries to interpret the data that is contained in a macroscopic representation, he or she necessarily makes an assumption regarding the distribution of the aggregated values over the underlying agents. For example, in Fig. 1, the observer may

Figure 5: Observed-to-expected ratio of citation number for the groups of agents defined by the 3rd level of the `WUTS` hierarchical organization

consider that each agent has the same weight in the group. It is thus underlined that aggregated values are *uniformly distributed* over the agents (from Q' to Q). Consequently, some groups are more suitable than others to summarize the microscopic information: using group A seems relevant since P is close to Q, unlike group B. Hence, organizations should be carefully chosen in order to provide accurate abstractions. In particular, they should only aggregate homogeneous and redundant distributions of the displayed variable.

Among classical similarity measures to compare a source distribution P with a model distribution Q, Kullback-Leibler (KL) divergence is of high interest because of its interpretation in terms of information content. This section shows how it can be exploited to provide feedback regarding the quality of groups and organizations and to ensure their proper interpretation by the observer.

4.1 Formalization and Semantics of Kullback-Leibler Divergence

Formally, KL divergence measures the number of bits of information that one loses by using the model distribution Q to find the optimal binary coding of countries associated to articles, instead of using the source distribution P [11]. In other words, KL divergence estimates the information that is lost by the aggregation process. But more generally, it is a measure of dissimilarity between two probability distributions. Hence, it can be interpreted as a *fitness function* between a source P and a model Q.

In our case, the "uniform hypothesis" is not suitable to interpret an aggregated representation. Indeed, for a given group, countries *do not have* the same weight regarding citation number. For example, the observer may assume that, within the `Northern America` group, the `USA` agent accumulates much more citations than the `Canada` and the `Mexico` agents. The aggregated value should thus be interpreted depending on that fact. The marginal values can be used to interpret an aggregated representation Q' and to give the corresponding model Q: the citations associated to a group of agents during a time period are distributed according to the total citation numbers of the underlying agents over the whole dataset. Given an agent a in a group G and a time period t, the interpreted citation number is thus given by the following formula:

$$Q(a,t) = v(G,t)\ \frac{v(a,.)}{v(G,.)}$$

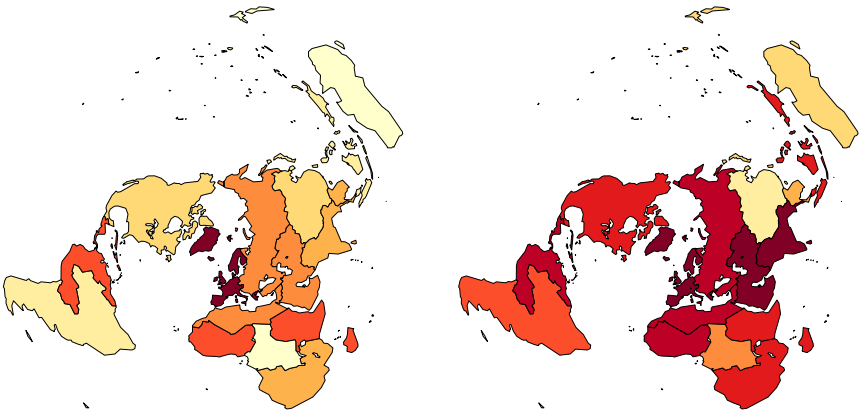
This interpreted value is then compared to the observed microscopic value: $P(a,t) = v(a,t)$. From the KL formula in [11], we define the *divergence* of a group G (or *information loss*, in bits) as follows:

$$\begin{aligned}\text{loss}(G,t) &= \sum_{a\in G} P(a,t)\log_2\left(\frac{P(a,t)}{Q(a,t)}\right)\\ &= \sum_{a\in G} v(a,t)\log_2\left(\frac{v(a,t)\ v(G,.)}{v(G,t)\ v(a,.)}\right)\end{aligned}$$

As we assume that aggregated values are thus distributed among underlying agents, a group whose internal distribution is very close to the observed distribution (as group A in Fig. 1) will have a low divergence, and conversely (as group B). KL divergence verifies the *sum property* [2], meaning that the divergence of disjoint groups is the sum of their divergences. Therefore, for an organization O, we have: $\text{loss}(O,t) = \sum_{G\in O}\text{loss}(G,t)$

4.2 Divergence is Correlated with the Source of Information

This first experiment aims at showing an essential feature of the aggregation process: its quality depends on the context of the analysis. Fig. 6 presents the KL divergence of groups defined by the `WUTS_3` macroscopic organization for two different newspapers (*The Guardian* and *The New York Times*) that have been observed during the month of July 2011. The darker a group is, the higher its KL divergence is, the more heterogeneous its internal distribution is. Such groups should not be used for the aggregation



(a) *The Guardian* (UK) (b) *The New York Times* (USA)

Figure 6: Spatial variations of the KL divergence for groups of the `WUTS_3` organization (the darker, the higher)

since they induce a misleading interpretation of the data. In this case, the real microscopic representation significantly diverges from the macroscopic model, making these groups unsuitable for the analysis. To the contrary, bright groups of countries constitutes *good* abstractions in terms of information content. The aggregated representation they provide regarding the corresponding geographical area fits with the microscopic data and can thus be properly interpreted by the observer.

In the case of *The Guardian* (*cf.* Fig. 6(a)), the groups with a high divergence are the location of microscopic events that can be spotted in Fig. 2. In these cases, the suppression of the corresponding microscopic variations induces a significant information loss. Divergence thus indicates heterogeneous behaviors in lower levels that should be detailed in order to produce properly interpretable representations. In the case of the *The New York Times* (*cf.* Fig. 6(b)), the `WUTS_3` groups have not the same divergence than in the previous case. First, divergence is globally higher, thus indicating a more heterogeneous microscopic behavior. This newspaper should then be analyzed at a lower level of representation than *The Guardian*. Moreover, events are not reported in the same way, or with the same intensity, depending on the newspaper editorial policies. We do not aim at making explicit the various positive and negative factors explaining the citation number (*e.g.*, geographical, cultural, historical factors [5, 10]), but at showing that groups should be chosen with respect to the dataset. In our case, this is partly correlated with the source of the information. As a consequence, if an analyst uses distributed probes to observe a MAS, she does not want to use only one abstraction pattern to summarize the information. This is consistent with the *subjectivist* account of emergence, according to which emergent phenomena strongly rely on the observation process [3].

4.3 Divergence of Groups Varies over Time

Fig. 7 presents the time variation of the KL divergence (thick plain line) of the `Northern America` group (compounded of the `USA`, the `Canada` and the `Mexico` agents). Each value have been computed at the week level, by comparing the interpreted and the observed citation numbers for the countries of this group. The graph shows that a group with a globally low divergence through time can nonetheless be the source of significant information losses during specific time periods. Henceforth, the choice of the representation level should also fit with the analyzed time period. Moreover, the graph shows that the divergence variations are not strictly correlated with the variation of the analyzed variable itself (dashed line): an increasing of the observed-to-expected ratio of citation number does not implies an increasing of the divergence, and conversely. Hence, the citation number is not a sufficient criterion to evaluate the information content of organizations, by contrast with divergence.

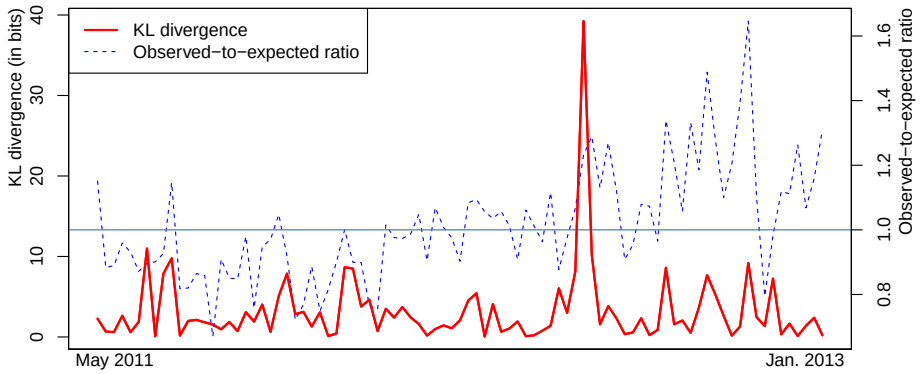


Figure 7: Time variation of the KL Divergence and the observed-to-expected ratio of citation number of the `Northern America` group for *The Guardian*

4.4 Divergence is Correlated with the Shape of the Groups

The purpose of this third experiment is to compare two mesoscopic agent organizations: `WUTS_2` and `UNEP_reg` (see Fig. 8). First, a global comparison indicates which organization minimizes the KL divergence. In order to compare the results from different newspapers, the information loss induced by organizations is normalized by the total citation number of the corresponding newspaper (in the following array, "b/c" stands for "bits/citation"):

	The Vancouver Sun	*The Daily Mail*	*The Ph. Daily Inquirer*
`WUTS_2`	1.80 b/c	1.46 b/c	2.07 b/c
`UNEP_reg`	1.57 b/c	1.51 b/c	2.26 b/c

It appears that, both for *The Daily Mail* and *The Philippine Daily Inquirer*, divergence is slightly lower for the `WUTS_2` organization than for the `UNEP_reg` organi-

zation. Hence, if one should choose between these two, WUTS_2 should be preferred. However, for *The Vancouver Sun*, UNEP_reg is better. Once again, abstractions should then be chosen with respect to the source of information.

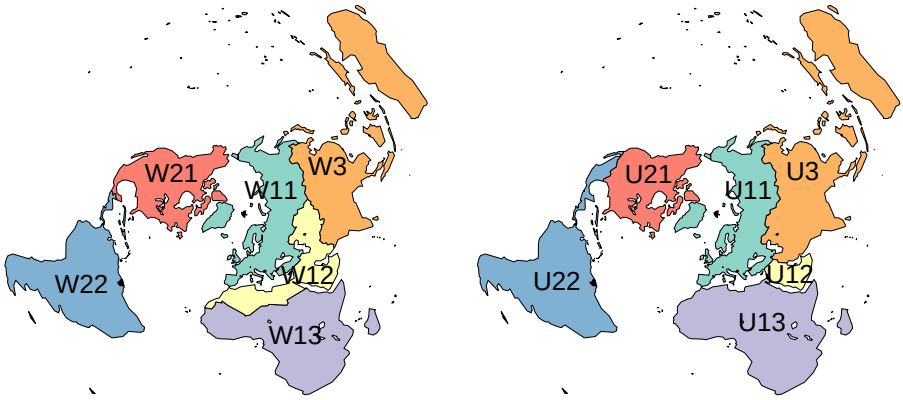
We can perform a more subtle analysis in order to determine the groups *best* shapes. For example, we notice in Fig. 8 that U22 = W22 ∪ Mexico and W21 = U21 ∪ Mexico. Hence, one may ask "what is the best location of the Mexico agent?" Should it be aggregated with the Northern America group (W21/U21) or with the Latin America group (W22/U22)? For *The Daily Mail*, we have:

$$\text{loss}(\texttt{W21}) + \text{loss}(\texttt{W22}) \;=\; 0.048 \text{ b/c} \quad < \quad 0.055 \text{ b/c} \;=\; \text{loss}(\texttt{U21}) + \text{loss}(\texttt{U22})$$

Therefore, the observed-to-expected ratio of citation number of the Mexico agent is closer to those of the Northern America group. Mexico should be aggregated accordingly. This technique allows to evaluate the geographical abstractions used by the experts in terms of information content and to choose their best shape for the macroscopic analysis of a given dataset.

5 The Complexity Reduction Induced by the Aggregation

The information content is never increased by the aggregation process: for every pair of disjoint groups, we have: $\text{loss}(G_1 \cup G_2) \geq \text{loss}(G_1) + \text{loss}(G_2)$. Hence, if we only rely on KL divergence, the more detailed is always the better. That is why we need a measure that also expresses what one *gains* by aggregating the microscopic data. To do so, this section presents two measures of *complexity reduction*. They estimate the



(a) The WUTS_2 organization (b) The UNEP_reg organization

Figure 8: Two organizations of the agents space in six similar (but not equivalent) groups of countries: locations of the Northern Africa group, the Western Asia group and the Mexico agent differ

information quantity that one saves by encoding a group G rather than its underlying agents:

$$\text{gain}(G) = \left(\sum_{a \in G} Q(a)\right) - Q(G)$$

where Q estimates the quantity of information needed to represent the agent a or the group G.

5.1 Number of Encoded Values

One way of measuring information quantities consists in estimating the number of bits needed to encode the values of a given representation. We may assume that it is constant for each agent or group: $Q(a) = Q(G) = q$, where q depends on the data type of the encoded values. Hence, for a group, we have:

$$\text{gain}(G) = (|G| - 1) \times q$$

This function gives a basic complexity measure that fits well with classic visualization techniques (as for the maps of this paper) since the number of displayed values defines the granularity of the visualization. For example, according to the map expected complexity, the user can determine the number of groups that should be displayed. Fig. 9 gives on the left the organization sizes (numbers of groups) and the associated gain for each level of the `WUTS` hierarchy. However, all groups do not contain the same number of agents. The same figure gives on the right, for each level, the sizes (numbers of agents) of the three high-level groups: `Euro-Africa`, `Americas` and `Asia-Pacific`. The user may want to adapt the level of these three groups depending on the amount of detail he or she expects for the corresponding geographical areas. The following section presents a criterion that automatically combines KL divergence and complexity reduction to adapt the size of groups depending on their quality, thus leading to multiresolution organizations.

5.2 Shannon Entropy

The number of encoded values only depends on the groups partitioning proposed by a given organization. In contrast, Shannon entropy also depends on the variable distribution. It is a classical complexity measure that is consistent with KL divergence (it can be defined as *the divergence from the uniform distribution* [11]). Briefly, entropy evaluates the quantity of information needed to encode the countries associated to *each citation* (and not to only encode the citation number for *each agent*). Based on Shannon's formula [17], we define the *entropy reduction* (or *gain*, in bits) of a group G as follows:

$$\text{gain}(G,t) = v(G,t)\ \log_2 v(G,t)\ -\ \sum_{a \in G} v(a,t)\ \log_2 v(a,t)$$

The choice of either one of these two complexity measures depends on the performed analysis. *Shannon entropy* should rather be used for the visualization of individuated citations, whereas *the number of encoded values* is more consistent with the

Organization	$\|O\|$	gain/q
WUTS_0	1	192
WUTS_1	3	190
WUTS_2	7	186
WUTS_3	17	176
WUTS_4	36	157
WUTS_5	193	0

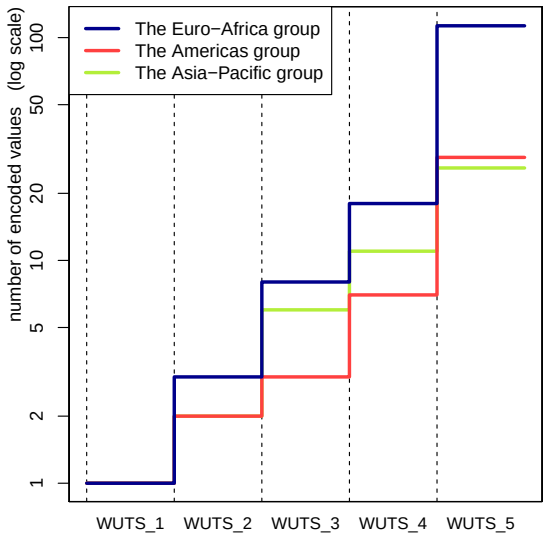


Figure 9: Complexity reduction (number of encoded values) of the six levels of the WUTS hierarchy (on the left) and of the three high-level groups of WUTS_1 (on the right)

visualization of aggregated values. In any case, techniques presented in this paper are meant to be generic. They can be used with any complexity measure as long as it fits with some algebraic properties (see [13] for more details).

6 Multiresolution Organizations of Systems

As a conclusion to the previous sections, finding a *good* organization relies on two aspects: (1) the complexity reduction (or *gain*) quantifies the granularity of the macroscopic representation and (2) the KL divergence (or *loss*) quantifies the amount of information that have been lost during the aggregation process. Choosing an organization thus consists in finding a compromise between these two aspects.

6.1 Parametrized Information Criterion

A *parametrized Information Criterion* can express the trade-off between complexity reduction and information loss for a given group G:

$$\mathrm{pIC}(G,t) \;=\; p \times \frac{\mathrm{gain}(G,t)}{\mathrm{gain}(A,t)} \;-\; (1-p) \times \frac{\mathrm{loss}(G,t)}{\mathrm{loss}(A,t)}$$

where $p \in [0, 1]$ is a parameter used to balance the trade-off. For $p = 0$, maximizing the pIC is equivalent to minimizing the loss: the observer wants to be as precise as possible (microscopic level). For $p = 1$, she wants to be as simple as possible (full aggregation). When p varies from 0 to 1, a whole class of nested organizations arises. The observer has to choose the ones that fulfill her requirements, between the expected amount of details and the computational resources available for the analysis.

Fig. 10 represents the groups of the `WUTS` hierarchy (squares) depending on the two criteria that have been previously defined: the ratio of *KL divergence* ($\mathrm{loss}(G)/\,\mathrm{loss}(A)$) and the ratio of *encoded values reduction* ($\mathrm{gain}(G)/\,\mathrm{gain}(A)$). In this plot, quality groups are easily spotted:

- The closer to the *bottom right* corner (red squares), the higher the information loss relatively to the complexity reduction. This is for example the case of the `Northern Europe` group (`W1111`): the unexpected citation number of the `Norway` agent (see subsection 3.1) makes the group very heterogeneous. Higher-level European groups (`W111` and `W11`) also induces a significant information loss (their are close to the right side). Thus, avoiding such groups during the aggregation ensures that we preserve the details regarding the significant microscopic variations.

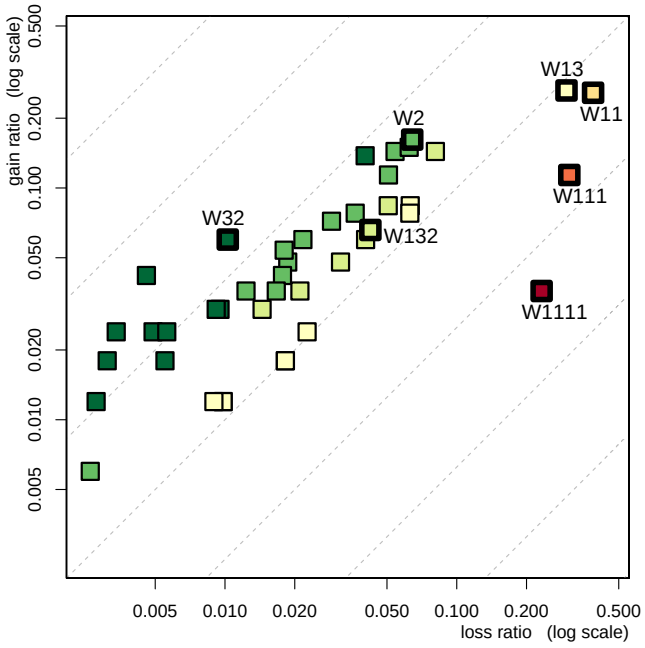


Figure 10: Comparison of the ratio of information loss and the ratio of complexity reduction (logarithmic scales) for the groups defined by the `WUTS` hierarchy applied to the map in Fig. 2

- On the contrary, the closer to the *top left* corner (green squares), the more the information loss is compensated by the complexity reduction. This is the case of the `Americas` and the `South Pacifica` groups (resp. `W2` and `W32`). This indicates that these representation levels are particularly interesting to give a synthetic view of the system. Indeed, these two groups correspond to relatively homogeneous geographical areas where no significant event have occurred during the observed time period (see map in Fig. 2).

- The `Horn of Africa` group (`W132`) has a better gain/loss ratio than the higher-level `Sub-Saharan Africa` group (`W13`). This indicates that, if some details are necessary to analyze the events occurring in Africa, the `Horn of Africa` group can however be described as a whole, without giving more details regarding this particular area. Hence, by choosing groups depending on their gain/loss ratio, the observer can represent the system with several spatial granularities in order to perfectly fit with the microscopic data.

This method allows to spot interesting groups of agent to build a synthetic but consistent macroscopic representation of the system. The rest of this section proposes an algorithm to automatize this evaluation process and to find the combinations of groups (*i.e.*, the organizations) that optimize the two criteria.

6.2 Organizations within a Hierarchy

Given a value of the trade-off parameter p, *best* organizations are those that maximize the parametrized information criterion. Clustering techniques using *gain* and *loss* measures as distances could find such optimal partitions. However, results may have very little meaning for the MAS analysis since agents would be aggregated regardless of their location within the system. In contrast, we assume that, in most spatial MAS, there is a correlation between topology and behavior. Hence, we propose that organizations should fit with the topological constraints defined by the domain experts. In other agent-based applications, such constraints may also be derived from *semantic* properties of the system (and not necessarily *topological* properties).

In this subsection, we consider hierarchically organized MAS. A *hierarchy* H is a set of nested groups, defined from the microscopic level (each agent is a group) to the whole MAS (only one group). The number of possible multi-resolution organizations within such a hierarchy *exponentially* depends on the number of groups and the number of levels. For `UNEP` (31 groups in 3 levels) and `WUTS` (64 groups in 5 levels), we respectively have 1.3×10^6 and 3.8×10^{12} possible organizations. Finding the best one can thus be computationally expensive in case of large-scale systems.

Algorithm 1 below finds topologically-consistent organizations that maximize our parametrized information criterion. Its complexity *linearly* depends on the number of groups in the hierarchy (respectively 196 and 231 groups) by doing a classical linear search within the branches of the hierarchy. Indeed, according to the *sum property* [2] of the defined information-theoretic measures, each branch can be independently evaluated (see [12] for more details).

Algorithm 1 linearly finds best organizations within a hierarchy

Require: A hierarchy H and a trade-off parameter p in $[0, 1]$.
Ensure: An organization made of groups in H that maximizes the pIC.

```
function FINDBESTORGANIZATION(H, p)
    if H contains only one group G then return {G}
    G ← biggest group of H
    bestMicroOrganization ← ∅
    for each direct subhierarchy S of H do
        aux ← FINDBESTORGANIZATION(S, p)
        bestMicroOrganization ← UNION(bestMicroOrganization, aux)
    if pIC of {G} > pIC of bestMicroOrganization then return {G}
    else return bestMicroOrganization
end function
```

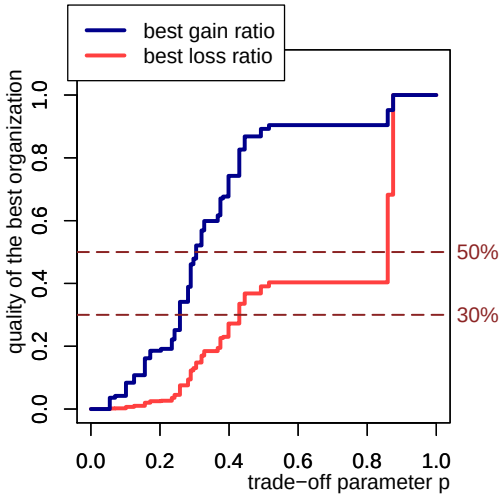


Figure 11: Ratio of complexity reduction (*gain*) and information loss (*loss*) of the best organizations of the map in Fig. 2 according to the trade-off parameter p

6.3 Hierarchical Aggregation to Build Spatial Macro-representations

The above algorithm has been ran on the `WUTS` hierarchy for the articles published by *The Guardian* during July 2011 (see Fig. 2). The plot in Fig. 11 gives the complexity reduction and the information loss associated to the organizations provided by the algorithm depending on the trade-off parameter p specified by the observer. For $p = 0$, the best organization corresponds to the microscopic representation (no information loss and no complexity reduction). As p increases, groups of countries are chosen within the `WUTS` hierarchy in order to focus on significant events. The observer can adjust the granularity of the generated representation depending on the expected level of detail. Fig. 12 and 13 present two organizations respectively preserving at least 50% and 70% of the microscopic information: $\text{loss}(O, t)/\text{loss}(A, t) < 0.5$ and

Figure 12: Best geographical organization preserving at least 50% of the microscopic information ($p \in< 0.86$)

$\text{loss}(O,t)/\text{loss}(A,t) < 0.3$ respectively for $p < 0.86$ and $p < 0.43$ (see Fig. 11).

The map in Fig. 12 represents the observed-to-expected ratio according to 17 groups of countries, two of which are high-level groups of countries (the `Americas` group and the `Asia-Pacific` group) where the observed citation number is quite close to the expected value: no significant event have been spotted at this level of detail. By contrast, two events are immediately highlighted in Europe and in Africa. They correspond to the most unexpected citation numbers for which an aggregation would induce a misleading information loss (at most 40% of information loss when $p < 0,85$ and at least 68% when $p > 0,85$, see Fig. 11).

1. The map in Fig. 12 displays the national details regarding agents of the `Northern Europe` group. Among them, the observed citation number of the `Norway`

agent is 3.7 times higher than expected. The observer is thus informed that a significant event took place at the national level during July 2011 (the two terrorist attacks of the 22th, *cf.* subsection 3.1).

2. This map also displays some details regarding the `Sub-Saharan Africa` group, but at a mesoscopic level constituted of 4 intermediary groups. Among them, the observer notices that the citation number of the `Horn of Africa` group is 1.8 times higher than expected. As the national details are not represented for this particular group, the observer may consider that this aggregated value is – at least at first glance – a good approximation of the underlying values. She concludes that *the observed-to-expected ratio of citation number is uniformly high in the Horn of Africa.* This group thus highlights an event that occurred in an extended geographical area (the food crisis that have been declared at the beginning of July 2011, *cf.* subsection 3.1)

The aggregation algorithm thus allows to represent significant events that occur at different granularities of the system organization.

The map in Fig. 13 is a little bit more detailed than the previous one, in particular for countries of the `Asia-Pacific` group and those of the `Western Africa` group. Some other – less significant – microscopic events are thus represented:

3. The severe floods that occurred in Thailand at the end of July 2011[5].

4. The development cooperation project that began the 16th of July between the European Union and the Republic of Guinea-Bissau.

However, in this second map, the `Horn of Africa` group of agents is still aggregated. This is only when the observer asks for at least 83% of the microscopic information ($p < 0.39$) that the algorithm provides the national details regarding this geographical area. In that way, the algorithm can adapt the generated representations to the observer's expectations.

7 Generalization to Time Aggregation

The time series in Fig. 14 gives the week-level variations of the observed-to-expected ratio of citation number of the `Greece` agent by *The Guardian*. Peaks of unexpected citation number reveal significant events in Greece recent history. In the following, we will focus on three particular events:

1. The peak of citation that appears at the beginning of November 2011 is explained by the announcement the 31st of October of a referendum regarding the setting up of an austerity plan to reduce the Greek public debt. This announcement

[5] `http://en.wikipedia.org/wiki/2011_Thailand_floods`

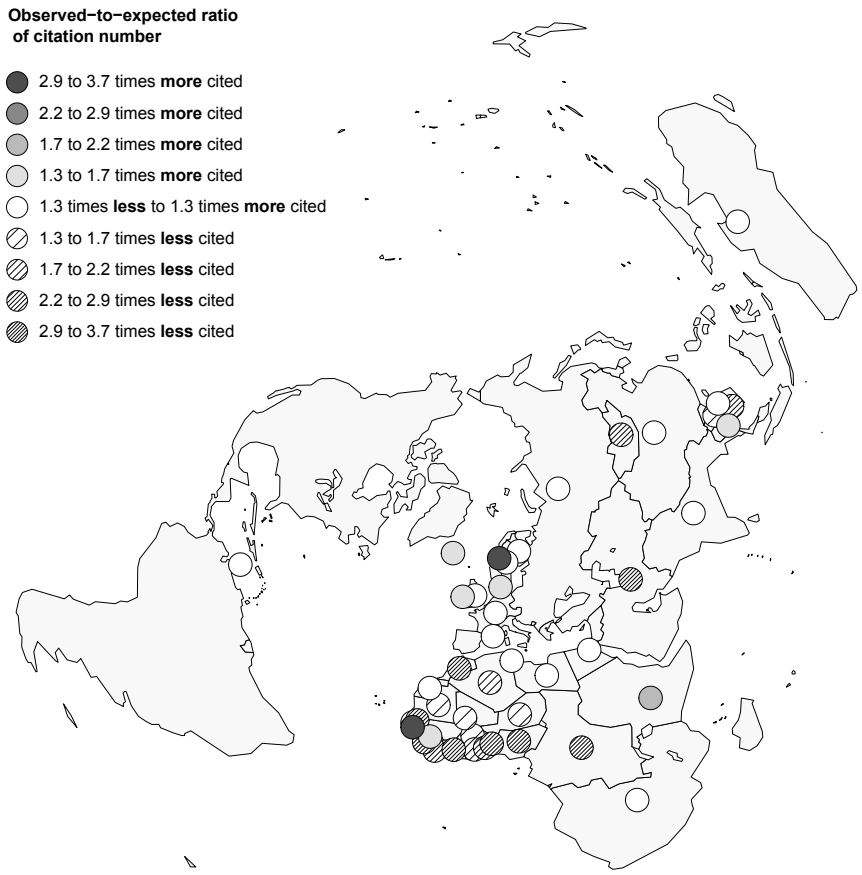


Figure 13: Best geographical organization preserving at least 70% of the microscopic information ($p < 0.43$)

causes everyone's surprise and is widely reported by the media. On the 4th of November, the Minister of Finance announces the referendum abandonment and the Prime Minister George Papandreou arranges, on the same evening, a vote of confidence in the Parliament that could lead to his resignation.

2. The peak that appears in the middle of May 2012 is explained by the failure of the legislative elections that are held the 6th of May and concluded the 16th by the establishment of an interim government until the organization of new elections.

3. The peak that appears at the end of June 2012 is explained by the holding, on the

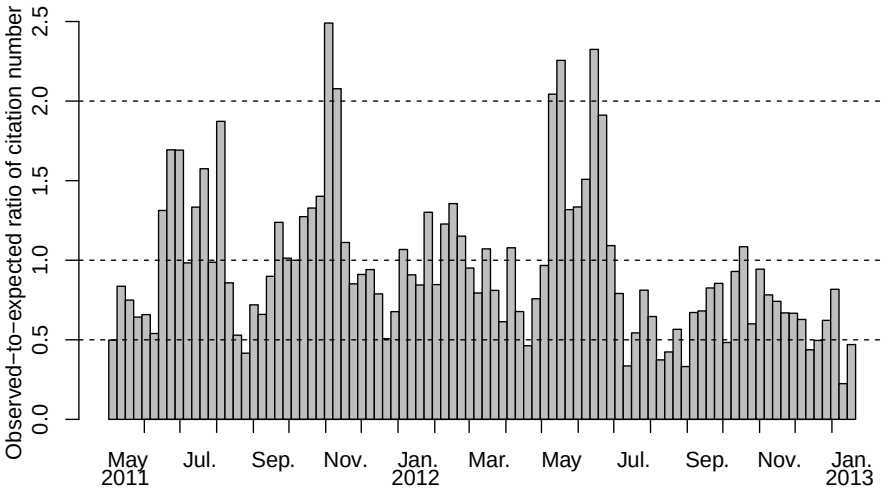


Figure 14: Microscopic time series (week level) of the observed-to-expected ratio of citation number of the `Greece` agent by *The Guardian*

17th of June, of these second legislative elections[6].

The same aggregation technique is applied to the system's temporal dimension. In this case, macroscopic events refer to time periods during which the citation number of a given country (or group of countries) has been much higher than expected. Such periods – that breaks with the system's stable state – may also be defined at different time scales (days, weeks, month, years, and so on). When interpreting an aggregated time period, the observer can use – as for the spatial aggregation – the marginal values to distribute the aggregated citation number over the underlying microscopic time periods (*i.e.*, the week level). For a microscopic time period t in an aggregated time period $T' \subset T$, the interpreted citation number is:

$$Q(a,t) = v(a,T') \, \frac{v(.,t)}{v(.,T')}$$

This interpreted value is then compared to the observed value: $P(a,t) = v(a,t)$, according to KL divergence:

$$\begin{aligned} \text{loss}(a,T') &= \sum_{t \in T'} P(a,t) \log_2 \left(\frac{P(a,t)}{Q(a,t)} \right) \\ &= \sum_{t \in T'} v(a,t) \log_2 \left(\frac{v(a,t) \; v(.,T')}{v(a,T') \; v(.,t)} \right) \end{aligned}$$

[6]See the Wikipedia page dedicated to the Greek government-debt crisis to get more details regarding the chronology of these political events: `http://en.wikipedia.org/wiki/Greek_government-debt_crisis`

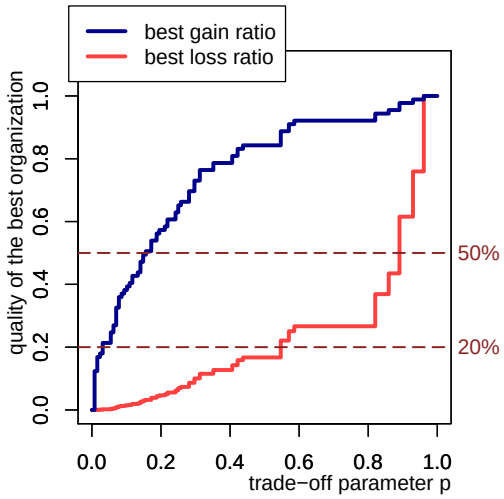


Figure 15: Ratio of complexity reduction (*gain*) and information loss (*loss*) of the best partitions of the time series in Fig. 14 according to the trade-off parameter p

The optimization of the corresponding *parametrized Information Criterion* (see subsection 6.1) can be achieved by the algorithm of Jackson *et al.* [9]. It consists in slicing the time series in intervals that maximize a given fitness function. We have shown in [12] that this time-aggregation algorithm is part of a larger class of algorithms – including the space-aggregation algorithm proposed in this paper – that consist in computing the optimal partitions of a dataset under some constraints (*e.g.*, hierarchical organization for spatial aggregation, ordered dataset for temporal aggregation). The time-aggregation algorithm of Jackson *et al.* is thus exploited as previously to build multiresolution representations of the agent dynamics.

The plot in Fig. 15 gives the complexity reduction and the information loss of the time partitions provided by the algorithm depending on the trade-off parameter p. Series in Fig. 16 and 17 present two such optimal partitions respectively preserving at least 80% and 50% of the week-level information (resp. $p < 0.55$ and $p < 0.89$). The time series in Fig. 16 thus summarizes the microscopic data by aggregating groups of weeks for which the observed-to-expected ratio of citation number is quite homogeneous. Even if the result contains less details, it still provides significant information for the analysis of Greek news. In particular, the three significant aforementioned peaks are highlighted and easily spotted by the observer. Moreover, the variations between the peaks are synthetically represented. This allows to describe the system dynamics according to macroscopic time periods.

The time series in Fig. 17 gives an even more aggregated representation of time variations. Only the most significant events are represented:

1. The first peak, corresponding to the announcement the 31st of October of a Greek referendum, is strongly highlighted by the aggregation process.

2. The two peaks of May and June 2012, respectively corresponding to the legislative elections of the 6th of May and the 17th of June, are aggregated to-

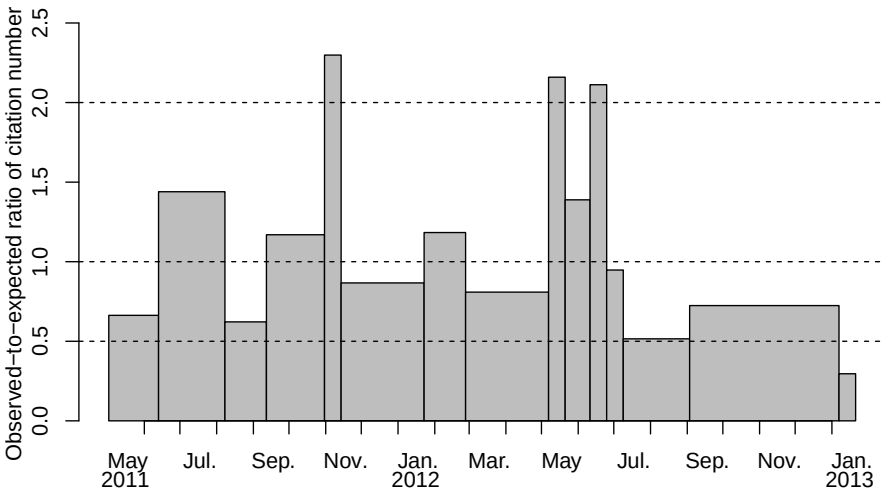


Figure 16: Best time partition preserving 50% of the microscopic information ($p < 0.55$)

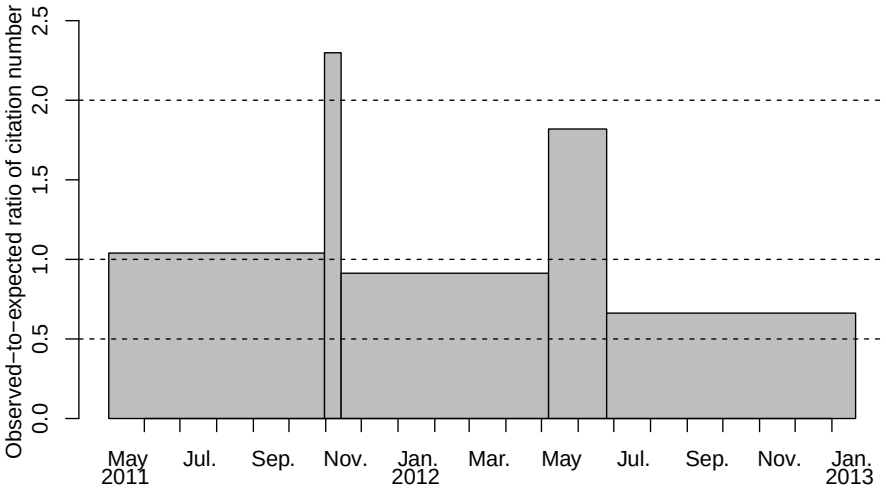


Figure 17: Best time partition preserving 80% of the microscopic information ($p < 0.89$)

gether, now constituting a unique homogeneous time period of 7 weeks. The corresponding event is interpreted as "the Greek legislative elections of 2012", without giving the week-level details of this month-scale event. The aggregation process thus provides consistent temporal abstractions to describe and analyze Greek news at a higher level of representation.

3. This third time series also gives an interesting macroscopic information: the citation number of the `Greece` agent has been globally decreasing during the observation period. This is explained by the declining media interest regarding the Greek crisis after the arrival in news of other economical crises concerning European countries such as Spain and Italy. The aggregation process thus allows to represent the system's temporal dynamics at several time-scales. It highlights the macroscopic variations: if we had the sufficient depth, we would be able to identify a year-long time period in Greek history corresponding to "the government-dept crisis" described as a whole.

8 Conclusion and Perspectives

The design and debugging of complex MAS need abstraction tools to work at a higher level of representation. However, such tools have to be developed and exploited with the greatest precaution in order to preserve useful information regarding the system behavior and to guarantee that generated representations are not misleading for the observer. To that extent, this paper focuses on aggregation techniques for large-scale MAS and gives clues to estimate their quality in term of complexity and information content. They are applied to the geographical and temporal aggregation of international relations through the point of view of on-line newspapers. We show that, by combining information theoretic measures, one can give interesting feedback regarding abstractions and build multiresolution representations of the dataset that adapt to the effective information content.

We believe that these measures and algorithms can be generalized to a large class of MAS, provided that :

- one can observe and describe the agents microscopic behavior according to several discretized microscopic dimensions (here: space and time);
- one can define measures to express the descriptions *quality* (here: complexity and information content);
- these measures have the *sum property* [2];
- the semantic and topological properties of the aggregated dimensions can be used to provide meaningful abstractions for the domain experts (here: hierarchical organizations and order of time).

Future work will apply these techniques to other dimensions of the analysis: *e.g.* for aggregation of newspapers, thematic aggregation, multi-dimensional aggregation

[12]. Besides this work, we are currently exploiting these techniques for performance visualization of large-scale distributed systems [14]. This kind of application shows that our techniques can be scaled up to one million agents.

Acknowledgement

This work was partially funded by the ANR CORPUS GEOMEDIA project (ANR-GUI-AAP-04). We would like to thank Claude Grasland, Timothée Giraud and Marta Severo for their work on this project; and Lucas M. Schnorr for his close participation to previous work.

References

[1] Luis Búrdalo et al. "A Tracing System Architecture for Self-adaptive Multiagent Systems". In: *PAAMS'10*. 2010, pp. 205–210.

[2] I. Csiszár. "Axiomatic Characterizations of Information Measures". In: *Entropy* 10.3 (2008), pp. 261–273.

[3] J. Deguet, Y. Demazeau, and L. Magnin. "Element about the Emergence Issue: A Survey of Emergence Definitions". In: *ComPlexUs* 3 (2006), pp. 24–31.

[4] N. Elmqvist and J.D. Fekete. "Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines". In: *IEEE Transactions on Visualization and Computer Graphics* 16.3 (2010), pp. 439–454.

[5] Johan Galtung and Mari Holmboe Ruge. "The Structure of Foreign News: The Presentation of the Congo, Cuba and Cyprus Crises in Four Norwegian Newspapers". In: *Journal of Peace Research* 2.1 (1965), pp. 64–91.

[6] Javier Gil-Quijano, Thomas Louail, and Guillaume Hutzler. "From Biological to Urban Cells: Lessons from Three Multilevel Agent-Based Models". In: *Principles and Practice of Multi-Agent Systems*. Vol. 7057. LNCS. 2012, pp. 620–635.

[7] C. Grasland and C. Didelon. *Europe in the World – Final Report. Volume 1 of ESPON Project 3.4.1*. Vol. 1. 2007.

[8] P. Iravani. "Multi-level network analysis of multi-agent systems". In: *RoboCup 2008: Robot Soccer World Cup XII*. Vol. 5399. LNCS. 2009, pp. 495–506.

[9] Brad Jackson et al. "An Algorithm for Optimal Partitioning of Data on an Interval". In: *IEEE Signal Processing Letters* 12.2 (2005), pp. 105–108.

[10] Ruud Koopmans and Rens Vliegenthart. "Media Attention as the Outcome of a Diffusion Process–A Theoretical Framework and Cross-National Evidence on Earthquake Coverage". In: *European Sociological Review* 27.5 (2011), pp. 636–653.

[11] S. Kullback and R.A. Leibler. "On Information and Sufficiency". In: *Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86.

[12] Robin Lamarche-Perrin, Yves Demazeau, and Jean-Marc Vincent. “The Best-partitions Problem: How to Build Meaningful Aggregations”. In: *Proceeding of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'13)*. IEEE Computer Society, 2013.

[13] Robin Lamarche-Perrin, Jean-Marc Vincent, and Yves Demazeau. *Informational Measures of Aggregation for Complex Systems Analysis*. Tech. rep. RR-LIG-026. France: Laboratoire d'Informatique de Grenoble, 2012.

[14] Robin Lamarche-Perrin et al. *Evaluating Trace Aggregation Through Entropy Measures for Optimal Performance Visualization of Large Distributed Systems*. Technical Report. France: Laboratoire d'Informatique de Grenoble, 2012, forthcoming.

[15] Wilbur Peng et al. “Graph-Based Methods for the Analysis of Large-Scale Multiagent Systems”. In: *AAMAS'09*. 2009, pp. 545–552.

[16] Steven F. Railsback, Steven L. Lytinen, and Stephen K. Jackson. “Agent-based Simulation Platforms: Review and Development Recommendations”. In: *Simulation* 82 (9 2006), pp. 609–623.

[17] C.E. Shannon. “A mathematical theory of communication”. In: *Bell System Technical Journal* 27 (1948), pp. 379–423, 623–656.

[18] Alexei Sharpanskykh and Jan Treur. “Group Abstraction for Large-Scale Agent-Based Social Diffusion Models with Unaffected Agents”. In: *Agents in Principle, Agents in Practice*. Vol. 7047. LNCS. 2011, pp. 129–142.

[19] Jakob Tonn and Silvan Kaiser. “ASGARD – A Graphical Monitoring Tool for Distributed Agent Infrastructures”. In: *PAAMS'10*. 2010, pp. 163–175.

[20] United Nations Environment Programme. *Global Environmental Outlook: environment for development*. Vol. 4. Nairobi, 2007.

[21] Marc H. Van Liedekerke and Nicholas M. Avouris. “Debugging multi-agent systems”. In: *Information and Software Technology*. Vol. 37. 1995, pp. 103–112.

A Generic Algorithmic Framework to Solve Special Versions of the Set Partitioning Problem

Robin Lamarche-Perrin
Max-Planck-Institute for Mathematics in the Sciences, Leipzig, Germany
Robin.Lamarche-Perrin@mis.mpg.de

Yves Demazeau
CNRS, Laboratoire d'Informatique de Grenoble, France
Yves.Demazeau@imag.fr

Jean-Marc Vincent
Univ. Grenoble Alpes, Laboratoire d'Informatique de Grenoble, France
Jean-Marc.Vincent@imag.fr

Abstract

Given a set of individuals, a collection of subsets, and a cost associated to each subset, the *Set Partitioning Problem* (SPP) consists in selecting some of these subsets to build a partition of the individuals that minimizes the total cost. This combinatorial optimization problem has been used to model dozens of problems arising in specific domains of Artificial Intelligence and Operational Research, such as coalition structures generation, community detection, multilevel data analysis, workload balancing, image processing, and database optimization. All these applications are actually interested in *special versions* of the SPP where assumptions regarding the admissible subsets constraint the search space and allow tractable optimization algorithms. However, there is a major lack of unity regarding the identification, the formalization, and the resolution of these strongly-related problems. This paper hence proposes a generic framework to design dynamic programming algorithms that fit with the particular algebraic structure of special versions of the SPP. We show how this framework can be applied to two well-known versions, thus opening a unified approach to solve new ones that might arise in the future.

Keywords: Problem Solving, Constraint satisfaction, Combinatorial algorithms, Dynamic programming.

1 Introduction

Let a *population* $\Omega = \{x_1, \dots, x_n\}$ be a finite set of individuals, a set of *admissible parts* $\mathcal{P} = \{X_1, \dots, X_m\}$ be a set of nonempty subsets of Ω, and a *cost function* c be an application from $\mathcal{P}$ to $\mathbb{R}$. The *Set Partition Problem* (SPP) *with an additive objective* consists in finding a subset $\mathcal{X}^* \subset \mathcal{P}$ that partitions Ω and minimizes c:

$$\mathcal{X}^* \in \operatorname*{arg\,min}_{\mathcal{X} \in \mathfrak{P}} \left(\sum_{X \in \mathcal{X}} c(X) \right)$$

where $\mathfrak{P}$ is the set of *admissible partitions*:
$\{\mathcal{X} \subset \mathcal{P} \setminus \bigcup_{X \in \mathcal{X}} X = \Omega,\ \forall (X_1 \neq X_2) \in \mathcal{X}^2,\ X_1 \cap X_2 = \emptyset\}$.

This combinatorial optimization problem has been used to model a colossal amount of problems in the neighboring fields of Operational Research and Artificial Intelligence (see Subsection 2.1). In fact, the SPP naturally arises as soon as one wants to organize a set of objets into covering and pairwise disjoint subsets such that an additive objective is optimized. This is for example the case in *algorithmic game theory*, where a population of self-motivated agents must be partitioned into coalitions that optimally exploite the interagent synergies to achieve a given task [26, 24, 3], or in *multilevel data aggregation*, where a set of data points needs to be partitioned into homogeneous classes preserving the system's structure and optimizing the compression rate [15, 20, 23, 21, 17, 12, 16].

It is well-known that the SPP is NP-complete in the *general* case [9]. This still holds when all costs are equal and when the admissible parts do not contain more than three individuals [25]. Hence, even with these very restrictive additional constraints, one cannot hope for a general-purpose algorithm that can efficiently solve every instance of the SPP. Among the strategies that have been proposed to tackle this computational challenge (see Subsection 2.2), this paper is interested in those that propose to consider stronger additional constraints regarding the set of admissible parts $\mathcal{P}$ in order to define easier versions of the SPP. Such constraints may arise from the semantical, structural, functional, or topological properties of the modeled system. For example, in data aggregation, the partitioning should be consistent with the dataset's topological properties so that the generated macroscopic data is meaningful for the domain expert [21, 17, 12, 16]. In particular, in image processing, the two-dimensional structure of the dataset implies very strong constraints for compression [20].

Although many special versions of the SPP have been addressed in the past (see Subsection 2.3), there is a major lack of unity regarding the identification, the formalization, and the resolution of these however strongly-related problems. Many of the papers herein referenced deal with special versions of the SPP without explicitly referring to it. They cannot therefore benefit from the tools that have been extensively developed by the combinatorial optimization community. Consequently, some results have been proved several times by independent work, such as the *Ordered Set Partitioning Problem* (see Subsection 2.3.3) that has been solved at least five times in 30 years [9, 2, 27, 25, 15]. Rothkopf *et al.* have addressed several versions in [25], leading to deep results regarding their tractability in a unified *applicative* context, but without proposing any unified *algorithmic* framework to solve them. This paper aims at providing such a framework to design optimization algorithms for any versions of the SPP.

Section 2 presents in further details the broad range of applications of the SPP and identify several special versions that have been addressed in the past. Section 3 presents the generic algorithmic framework. It relies on a proper understanding of the search space's algebraic properties and uses dynamic programming to efficiently exploit them. Section 4 applies this framework to two special versions, namely the *Hierarchical Set Partitioning Problem* (HSPP) and the *Ordered Set Partitioning Problem* (OSPP). We show that the computational complexity of the resulting optimization algorithms meets the one of past algorithms dedicated to the same problems [9, 2, 27, 25, 23, 16], thus opening a unified approach to solve new versions of the SPP that might arise in the future.

2 Related Work

The prolific applicability of the SPP is partly explained by its closeness to the extensively-studied *set packing* and *set covering problems*, respectively corresponding to the relaxation of the "covering" and the "pairwise disjoint" constraints [4]. The SPP also generalizes numerous combinatorial optimization problems, in particular in *computational graph theory* where individuals are vertices of a graph and admissible parts are defined according to specific subgraph structures [26, 6, 10], thus leading to a tremendous amount of other possible applications. The next subsection focuses on applications in Artificial Intelligence and Operational Research.

2.1 Applying the SPP

2.1.1 The *airline crew scheduling problem* and other operational problems

Historically, the best-known application of the SPP is the *airline crew scheduling problem* [4]: given a set Ω of flight legs (takeoff and landing) and a set $\mathcal{P}$ of feasible sequences of legs by airline crews, each of these $X \in \mathcal{P}$ having a cost $c(X)$, the airline company would like to find a collection of feasible sequences minimizing the total cost such that each flight leg is covered by exactly one crew. This setting is easily generalizable to a broader class of *transportation, delivery, routing, and location problems* [4, 13, 7, 14], and to other well-known operational problems such as the *circuit partitioning problem* [1] or the *political districting problem* [13]. Many other examples can be found in [4] and [2].

2.1.2 The *winner determination problem* in combinatorial auctions

Given a set of assets Ω and a set of bids associating a price $c(X)$ to groups of asset $X \in \mathcal{P}$, the auctioneer wants to find an allocation of the assets to the bidders that maximizes his revenue. Rothkopf *et al.* [25] have shown that, when considering the classical *OR bidding language* to express the bidder objectives [18], the *winner determination problem* is equivalent to the *Set Packing Problem* [25]. Moreover, since the bids are always positive ($c \geq 0$), and by interpreting the absence of bid as a null price, there is always an optimal packing that is also an optimal partition. Hence, in this context, the *winner determination problem* is equivalent to the SPP.

2.1.3 The *coalition structure generation problem* in algorithmic game theory

A population of self-motivated agents Ω must collaborate with one another to perform a given task. Coalition structure generation consists in partitioning the agents into feasible teams – expressed by $\mathcal{P}$ – so as to achieve better result by maximizing the social welfare – expressed by c [26, 24, 3]. This field itself leads to many applications in *e-commerce* (buyers form coalitions to purchase a product in bulk), *e-business* (groups form to satisfy particular market niches), *distributed sensor network* (sensors work together to track targets), and *information gathering* (servers form coalitions to answer queries) [24].

2.1.4 The *clustering problem* for multilevel data analysis

The SPP can be seen as a general formulation of classical *clustering problems*: data points have to be partitioned into classes such that the intra-class similarity and/or the inter-class dissimilarity are maximized [8, 22, 6]. In this context, it also relates to *data aggregation problems*, where data points are partitioned into homogeneous classes preserving the system's structure [23, 21, 17, 12, 16] leading to applications in *time series analysis* [15, 21, 16], *database optimization*, and *image processing* [19, 20].

2.2 Solving the SPP

The SPP is usually tackled the following strategies:

2.2.1 Exploiting the Algebraic Structure of the SPP

The set of parts and the set of partitions have several useful algebraic properties when one tries to directly tackle the general problem. Strategies formalizing and exploiting such properties to cleverly run through the search space are usually based on *integer programming* [4, 13, 14], *dynamic programming* [28, 2, 25, 15], *implicit enumeration* [4, 14], and/or *automatic reformulation* [14]. However, since the SPP is NP-complete, one should not expect any worst-case polynomial algorithm to emerge from such strategies, unless P = NP.

2.2.2 Approaching Optimality with Tractable Algorithms

Heuristics limiting the search space in some way have been proposed to find suboptimal solutions in reasonable time, including *genetic algorithms* [11], *dual ascent* [7], *simulated annealing*, and *neural networks* [13]. However, such approaches do not provide any worst-case guarantee regarding the closeness to optimality [24]. To the contrary, *approximation algorithms* provide provable solution quality and run-time bounds [20, 18, 6, 10, 3], but are still limited by severe inapproximability results [18].

2.2.3 Exploiting Properties of the Cost Function

Much work has focused on special cases of the SPP by making additional assumptions regarding the cost function. For example, the SPP has been proved to be polynomially solvable when costs are defined by an aggregative measure applied to some attributes of the individuals [8, 22, 2, 19, 20, 3]. In these cases, each considered cost function requires a dedicated treatment that can hardly be generalized to a broader context. Other work has hence focused on more general properties of the cost function, such as *concavity* [9, 2], *submodularity* [2, 18, 10], *superadditivity* [25, 26, 20], and *subadditivity* [26]. These approaches all assume that the costs are somehow monotonously defined with respect to the set inclusion. Although considerable results have been achieved for such settings, Sandholm *et al.* [26] argue that, because of communication costs or anti-trust penalties, many applications of the *coalition generation problem* are neither superadditive nor subadditive. This is also the case in multilevel data analysis,

where the information-theoretic measures are usually non-monotonous regarding the set inclusion [23, 21, 17, 12, 16].

2.2.4 Exploiting Structures of the Admissible Parts

As announced in the introduction, this paper focus on strategies exploiting constraints on the set of admissible parts $\mathcal{P}$ to define easier versions of the SPP. In this case, one should guarantee that the constraints do not exclude solutions that would be optimal otherwise. For example, in the *winner determination problem*, if the assets are known to be more valuable in given combinations, the auctioneer may anticipate the bids of greatest economic significance and only allow such valuable combinations [25, 18]. However, constraints might also arise from semantics considerations when some subsets are not meaningful for the partition purposes. In data aggregation, for example, the partitioning should be consistent with the dataset's structural and topological properties so that the compressed data is usable by the domain expert [21, 17, 12, 16]. In particularly, in image processing, the two-dimensional structure implies very strong constraints for compression [20]. The following subsection present some structures that have been addressed in previous work.

2.3 Special Versions of the SPP

In the following, we indicate the worst-case time complexity of algorithms according to the size n of the population.

2.3.1 The Complete Set Partitioning Problem (CSPP)

Definition The CSPP arises when all subsets are admissible: $\mathcal{P} = 2^{\Omega}$.

Applications The CSPP has been extensively used for *coalition structure generation*, assuming that every possible group of agents is an adequate candidate to constitute a coalition [26, 24]. It has also been applied to *corporate tax structuring* to find an optimal aggregation of corporate subsidiaries to pay state unemployment compensation tax [28].

Results The CSPP is NP-complete [26]. However, exponential algorithms have been provided to solve the problem on small instances: a $O(3^n)$ dynamic programming algorithm [28, 18] and a $O(n^n)$ anytime algorithm that quickly generates and slowly improves a suboptimal solution [26, 24].

2.3.2 The Hierarchical Set Partitioning Problem (HSPP)

Definition The HSPP arises when the set of admissible parts $\mathcal{P}$ forms a *hierarchy*, that is when every two admissible parts are either disjoint or one is included in the other: $\forall (X_1, X_2) \in \mathcal{P}^2,\ X_1 \cap X_2 = \emptyset$ or $X_1 \subset X_2$ or $X_1 \supset X_2$. If one also assumes that the population and the singletons are admissible: $\Omega \in \mathcal{P}$ and $\forall x \in \Omega,\ \{x\} \in \mathcal{P}$,

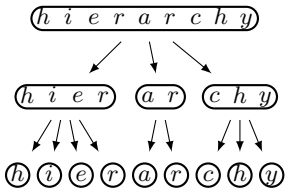


Figure 1: A 3-levels hierarchy defined on a population of size 9

then the hierarchy can be described as a *rooted tree* where the tree-order is the subset relation (see Fig. 1).

Applications The HSPP has been mainly applied in data aggregation to model systems with multilevel nested structures. This include *community representation of networks* (individuals are nodes of a graph representing a social structure and admissible parts are highly-connected groups of nodes) [23], *aggregation of geographical data* (individuals are territorial units and admissible parts are defined according to nested geographical partitions) [16], and *analysis of distributed systems* (individuals are computational resources and admissible parts are defined according to the system's hierarchical structure) [17, 12].

Results All algorithms solving the HSPP consist in a $O(n)$ depth-first search of the hierarchy [23, 17, 16].

2.3.3 The Ordered Set Partitioning Problem (OSPP)

Definition The OSPP arises when a total order $<$ is defined on Ω and the admissible parts are the intervals induced by this order: $\mathcal{P} = \{\{x_i < \ldots < x_j\} \subset \Omega \backslash i \leq j\}$. This set can be represented as a "pyramid of intervals" (see Fig. 2).

Applications The OSPP very naturally apply to partition any population that has a temporal component (*e.g.*, sets of dates, events, or time periods). For example, the OSPP has been addressed for the aggregation of time series [15, 21, 16]. This setting might also receive a unidimensional-space interpretation, such as the North-South geographical ordering of the cities on the East Coast [25]. It has also been applied to *inventory control* and *production planing* [9, 2, 27].

Results Chakravarty *et al.* [9] have shown that, when the optimal partition is a sequence of intervals, solving the SPP is equivalent to solving the *shortest path problem*, resulting in a $O(n^2)$ optimization algorithm. A $O(n^2)$ dynamic programming algorithm has also been proposed in many independent works [2, 27, 25, 15].

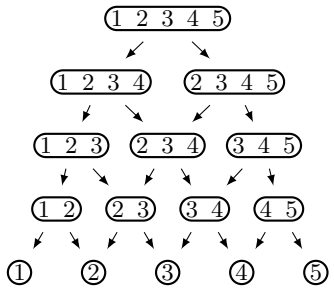


Figure 2: The "pyramid of intervals" of an ordered population of size 5

Extensions Some other order-related structures are sometime considered, such as *monotone partitions* (when the size of intervals increases with the order) [2], *extremal partitions* (the last individual of each interval can be associated to the next interval without losing optimality) [2], and *cyclic orders* defining intervals on a circle [25].

2.3.4 The Array Partitioning Problem (APP)

Definition The APP consists in partitioning a two-dimensional array into rectangular tiles. It naturally arises when one considers the Cartesian product $\Omega_1 \times \Omega_2$ of two ordered populations Ω_1 and Ω_2. The set of admissible parts is $\mathcal{P} = \{X \subset \Omega_1 \times \Omega_2 \backslash \exists X_1 \in \mathcal{P}_1, \exists X_2 \in \mathcal{P}_2, X = X_1 \times X_2\}$, where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the sets of intervals of Ω_1 and Ω_2.

Applications As for the OSPP, the APP may be used to model spatial structures, such as the rectangular partitioning of geographic locations on a two-dimensional grid [25, 5], the clustering of points with fairly uniform color in image processing, computer graphics, and video compression [19, 20], and the building of histograms to approximate multi-dimensional data distributions in database systems [20]. Moreover, the APP has been used to model *load balancing problems* in parallel computation when the computational space corresponds to a matrix [20]. In this case, the tiles are rectangular in order to respect the computation space topology [20] and to reduce the communication costs between subproblems [19].

Results Rothkopf *et al.* [25] have shown that the APP is NP-complet, even if one only considers 2×2 rectangles and singletons as admissible parts. However, the APP is manageable if one only considers rows, columns, and singletons as admissible parts, for example to represent assets that have two different properties of interest for a collector [25].

Extensions Other multidimensional versions of the SPP has been addressed, such as the partition of d-dimensional hypercubes [25], and the Cartesian product of a hierarchy

and a total order leading to a $O(n_1(n_2)^3)$ dynamic algorithm, where n_1 and n_2 are respectively the sizes of the hierarchical and the ordered populations.

2.3.5 Bounds on the Size of Admissible Parts

Definition Some versions of the SPP constraint the size of admissible parts: $\mathcal{P} = \{X \subset \Omega \backslash |X| \leq k\}$ or $\mathcal{P} = \{X \subset \Omega \backslash |X| > k\}$ for a given $k \in \mathbb{N}$.

Applications Such assumptions are very generic and might apply to any problem bringing in only small groups (or only large groups) of individuals to partition the population.

Results Rothkopf *et al.* [25] have shown that, when admissible parts are limited to size $k = 3$, the SPP is still NP-complete [25], when limited to size $k = 2$, that solving the SPP is equivalent to solving the *maximum-weight matching problem*, leading to a $O(n^3)$ optimization algorithm, and when only large parts and singletons are admissible ($\mathcal{P} = \{X \subset \Omega \backslash |X| = 1$ or $|X| < n/k\}$ for $k \in \mathbb{N}$), the SPP can be solved in $O(n|\mathcal{P}|^k)$ time.

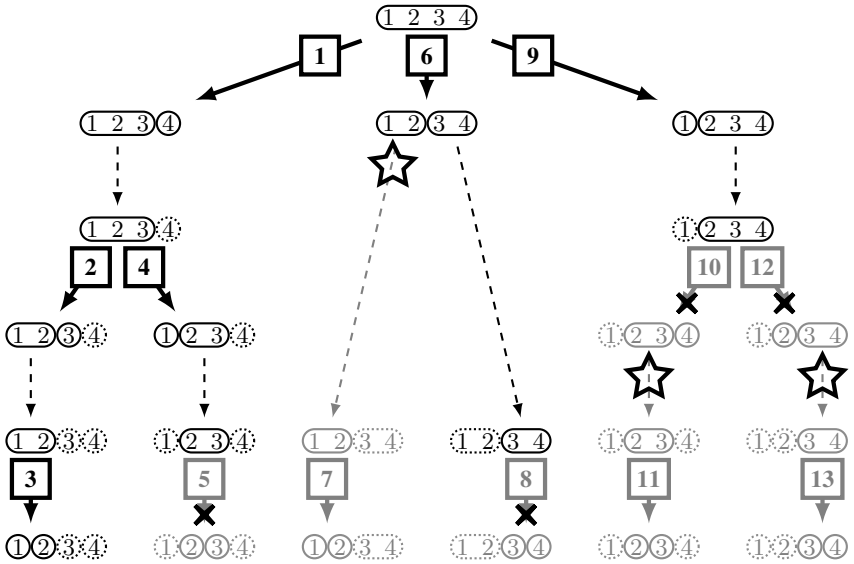


Figure 3: Execution trace of the generic algorithm for an ordered population of size 4. Plain numbered arrows represent branchings (step 2). Dashed arrows represent recursive calls (step 2.a). Crosses and stars give the cuttings of branches that may improve the algorithm.

3 A Generic Framework to Solve Special Versions of the SPP

This section proposes to rely on *multi-branching recursion* and *dynamic programming* to efficiently solve the SPP. The search space is first broke down into smaller covering subspaces. Then, thanks to a *principle of optimality* that fits with the algebraic structure of the partition set, these subproblems are recursively solved. Finally, locally-optimal solutions are compared to globally solve the initial problem.

This algorithm is *generic* in the sense that it can be applied to any set of admissible parts. However, it should be considered as an abstract tool to build optimization algorithms for special version of the SPP. Section 4 shows how to derive such *specialized implementations* for the HSPP and the OSPP.

3.1 Algebraic Structure and Principle of Optimality

Restricting the size the number of admissible part $|\mathcal{P}|$ is not sufficient to restrict the size of the search space $|\mathfrak{P}|$. For example, in the case of the OSPP, we have $|\mathcal{P}| = O(n^2)$ and $|\mathfrak{P}| = O(2^n)$ (see Subsection 4.2 hereinafter). Since the number of admissible partitions grows exponentially with the population size in most of the special versions reported in Subsection 2.3, the introduction of strong constraints is often not sufficient to make these problems tractable. Rothkopf *et al.* have argued that the computational complexity does not actually depend on the size of the search space, but rather on its structure [25]. In this subsection, we present such a structure and we propose a *principle of optimality* that exploit it to evaluate partitions in a computationally-efficient fashion.

3.1.1 The Algebraic Structure of the Search Space

The set of admissible partitions $\mathfrak{P}$ is structured by an essential algebraic relation, usually referred to as the *refinement relation* $\subset$. A partition $\mathcal{X}$ *refines* a partition $\mathcal{Y}$ if and only if each part in $\mathcal{X}$ is a subset of a part in $\mathcal{Y}$:

$$\mathcal{X} \subset \mathcal{Y} \quad \Leftrightarrow \quad \forall X \in \mathcal{X}, \quad \exists Y \in \mathcal{Y}, \quad X \subset Y$$

As this binary relation is reflexive, antisymmetric, and transitive, it defines a partial order on the partition set $\mathfrak{P}$ that consequently forms a poset.

The *covering relation* $\sqsubset$ is the transitive reduction of the refinement relation, that is the binary relation which holds between immediate "neighbors" regarding $\subset$. Hence, a partition $\mathcal{X}$ is *covered* by a partition $\mathcal{Y}$ if and only if $\mathcal{X} \neq \mathcal{Y}$, $\mathcal{X}$ refines $\mathcal{Y}$, and there is no other partition "between" them:

$$\mathcal{X} \sqsubset \mathcal{Y} \quad \Leftrightarrow \quad \mathcal{X} \subsetneq \mathcal{Y} \quad \text{and} \quad \nexists \mathcal{Z} \in \mathfrak{P}, \ \mathcal{X} \subsetneq \mathcal{Z} \subsetneq \mathcal{Y}$$

For a given admissible partition $\mathcal{X} \in \mathfrak{P}$, we define $\mathfrak{R}(\mathcal{X})$ as the *set of admissible partitions refining* $\mathcal{X}$, $\mathfrak{C}(\mathcal{X})$ as the *set of admissible partitions covered by* $\mathcal{X}$ and, respectively, $\mathfrak{R}^*(\mathcal{X})$ and $\mathfrak{C}^*(\mathcal{X})$ as the sets of optimal partitions among $\mathfrak{R}(\mathcal{X})$ and $\mathfrak{C}(\mathcal{X})$.

Note that if the *minimal partition* $\{\{x\}\}_{x\in\Omega}$ is admissible, it refines all admissible partitions: $\mathcal{X} \in \mathfrak{P},\ \{\{x\}\}_{x\in\Omega} \in \mathfrak{R}(\mathcal{X})$, and if the *maximal partition* $\{\Omega\}$ is admissible, it is refined by all admissible partitions: $\mathfrak{R}(\{\Omega\}) = \mathfrak{P}$.

3.1.2 A Principle of Optimality for the SPP

In dynamic programming, finding a principle of optimality consists in showing that the search space has an optimal substructure: the solution to the optimization problem can be obtained by recursively combining optimal solutions to several subproblems. Intuitively, in the case of the SPP, one can rely on the fact that *the union of optimal partitions on subsets of the population is an interesting candidate to form an optimal partition of the whole population*. Hence, by appropriately decomposing the population, one might provide a computationally efficient procedure to build such an optimal solution.

Theorem. *Let Ω be a population and c be an additive cost function defining partition optimality. For any partition $\mathcal{Y}$ of Ω, the union of optimal partitions of the parts of $\mathcal{Y}$ is optimal among the refinements of $\mathcal{Y}$:*

$$\forall Y \in \mathcal{Y},\ \mathcal{Y}^*_Y \in \mathfrak{P}^*(Y) \quad \Rightarrow \quad \left(\bigcup_{Y\in\mathcal{Y}} \mathcal{Y}^*_Y\right) \in \mathfrak{R}^*(\mathcal{Y}) \tag{1}$$

Proof. *Let $\mathcal{Y}$ be a partition of Ω and, $\forall Y \in \mathcal{Y}$, let $\mathcal{Y}^*_Y$ be an optimal partition of Y. We mark $\mathcal{Y}^* = \left(\bigcup_{Y\in\mathcal{Y}} \mathcal{Y}^*_Y\right)$. Any partition $\mathcal{Y}' \in \mathfrak{R}(\mathcal{Y})$ can be decomposed the following way: $\mathcal{Y}' = \left(\bigcup_{Y\in\mathcal{Y}} \mathcal{Y}'_Y\right)$ such that, $\forall Y \in \mathcal{Y}$, $\mathcal{Y}'_Y$ is a partition of Y. Since $c(\mathcal{Y}^*) = \bigcup_{Y\in\mathcal{Y}} c(\mathcal{Y}^*_Y) \leq \bigcup_{Y\in\mathcal{Y}} c(\mathcal{Y}'_Y) = c(\mathcal{Y}')$, we have $\mathcal{Y}^* \in \mathfrak{R}^*(\mathcal{Y})$.* □

3.2 Branching the Search Space

Given a admissible part $X \in \mathcal{P}$ for which one wants to compute an optimal admissible partition $\mathcal{X}^* \in \mathfrak{P}^*(X)$, a *branching* consists in building subspaces $\mathfrak{P}_1, \ldots, \mathfrak{P}_k$ that cover the search space: $\mathfrak{P}_1 \cup \ldots \cup \mathfrak{P}_k = \mathfrak{P}(X)$. Then, if one finds locally-optimal partitions $\mathcal{X}^*_1 \in \mathfrak{P}^*_1, \ldots, \mathcal{X}^*_k \in \mathfrak{P}^*_k$ for each of these subspaces, one can easily solve the optimization problem the following way:

$$\underset{\mathcal{X}\in\{\mathcal{X}^*_1,\ldots,\mathcal{X}^*_k\}}{\arg\min} c(\mathcal{X}) \quad \subset \quad \mathfrak{P}^*(X) \tag{2}$$

The covering relation indicates "atomic disaggregations" of a given set. For example, in the case of the OSPP, it consists in dividing the population into *two* intervals (see for example arrows **1**, **6**, and **9** in Fig. 3). The covering relation can thus be used to branch the search space. First, we know that all admissible partitions of X refine the maximal partition $\{X\}$: $\mathfrak{P}(X) = \mathfrak{R}(\{X\})$. Second, for any partition $\mathcal{X} \in \mathfrak{P}(X)$, a refining partition of $\mathcal{X}$ is either *the partition $\mathcal{X}$ itself*, or *a partition that refines a partition covered by $\mathcal{X}$*. Hence, the search space can be branched the following way:

$$\mathfrak{P}(X) = \{\{X\}\} \cup \left(\bigcup_{\mathcal{Y}\in\mathfrak{C}(\{X\})} \mathfrak{R}(\mathcal{Y})\right) \tag{3}$$

3.3 A Recursive Algorithm

The computation of an optimal partition thus consists in computing *locally-optimal partitions refining the partitions covered by the maximal partition.* Thanks to the principle of optimality, such a computation can be recursively performed (see Eq. 1) Hence, the three branching and recursion equations 1, 2, and 3 allow to define a divide and conquer algorithm that computes a locally-optimal partition $\mathcal{X}^* \in \mathfrak{P}^*(X)$ for any $X \in \mathcal{P}$ according to the following recursive formula:

$$\mathcal{X} \in \{\{X\}\} \cup \left(\bigcup_{\mathcal{Y} \in \mathfrak{C}(\{X\})} \left\{ \bigcup_{Y \in \mathcal{Y}} \mathcal{Y}^*_Y \right\} \right) \quad \underset{\arg\min}{} \quad c(\mathcal{X}) \ \subset \ \mathfrak{P}^*(X) \tag{4}$$

where $\mathcal{Y}^*_Y$ designates a partition in $\mathfrak{P}^*(Y)$. Here are the steps of the resulting algorithm:

(step 1) Compute the set $\mathfrak{C}(\{X\})$ of partitions covered by the maximal partition $\{X\}$.

(step 2) For each partition $\mathcal{Y} \in \mathfrak{C}(\{X\})$, do the following:

(step 2.a) for each part $Y \in \mathcal{Y}$, recursively compute an admissible locally-optimal partition $\mathcal{Y}^*_Y \in \mathfrak{P}^*(Y)$;

(step 2.b) compute the union $\mathcal{Y}^* = \bigcup_{Y \in \mathcal{Y}} \mathcal{Y}^*_Y$ (the principle of optimality ensures that $\mathcal{Y}^* \in \mathfrak{R}^*(\mathcal{Y})$);

(step 3) Return a partition that minimizes c among $\{X\}$ and the $\mathcal{Y}^* \in \mathfrak{R}^*(\mathcal{Y})$ computed for each $\mathcal{Y} \in \mathfrak{C}(\{X\})$.

Fig. 3 gives an example of execution of this algorithm in the case of an ordered population of size 4. The starting point is the maximal partition at the top (1 2 3 4). The plain numbered arrows represent the sequence of branchings executed by the algorithm (step 2): for example branching [1] evaluates the covering partition (1 2 3)(4). The dashed arrows represent the recursive calls on an admissible part (step 2.a): for example (1 2 3):4: corresponds to the execution of the algorithm on the first part of the partition (1 2 3)(4).

3.4 Dynamic Programming Improvements

This first algorithm is not computationally-optimal. In the rest of this subsection, we thus propose two improvements to reduce its time complexity and we give a more detailed description of the final algorithm.

3.4.1 Recording Intermediary Results

According to the dynamic programming paradigm, recursive algorithms can be easily improved by recording the results of time-consuming recursive calls. For each part on which the algorithm is once applied, by keeping trace of the resulting locally-optimal

partition, one can immediately return this result when posterior calls occur on the same part. This way, the algorithm is applied only once to each admissible part $X \in \mathcal{P}$. For example, in Fig. 3, the algorithm is initially applied *twice* on parts ①②, ②③, and ③④. Thanks to this *memoization* procedure, one can avoid the second calls (see stars on dashed lines).

3.4.2 Avoiding Redundant Evaluations

The branching of the search space proposed in Eq. 3 is redundant, *i.e.* subspaces are not disjoints. For example, in Fig. 3, branches **2** and **3** allow the evaluation of partitions ①②③④ and ①②③④, and branches **4** and **5** the evaluation of partitions ①②③④ and ①②③④. Hence, ①②③④ is evaluated *twice* and **5** is useless. In order to avoid such redundant branches, one can keep trace of the covered partitions $\mathcal{X}_1, \ldots, \mathcal{X}_k$ that have already been evaluated during step 2. When the algorithm is recursively applied to a part $X \in \mathcal{X}_{k+1}$ (step 2.a), one also retains the *complementary partition* $\overline{\mathcal{X}} = \mathcal{X}_{k+1} \backslash \{X\}$. Hence, within the "lower" calls, when a covered partition $\mathcal{Y} \in \mathfrak{C}(\{X\})$ is considered for branching, one first checks if $\overline{\mathcal{X}} \cup \mathcal{Y}$ does not refine any of the previously-evaluated partitions. If it does ($\exists i \leq k,\ \overline{\mathcal{X}} \cup \mathcal{Y} \in \mathfrak{R}(\mathcal{X}_i)$), one deduces that the branch has already been evaluated, and steps 2.a and 2.b may be avoided. Fig. 3 indicates the result of this improvement by crosses cutting the plain arrows.

However, this improvement should be implemented with the greatest care in order to be computationally efficient. In Section 4, where this generic algorithmic framework is applied to special versions of the SPP, the specific algebraic structures resulting from the covering relation are fully known. We then show how one can directly generate, during step 2, the covered partitions that have not been evaluated yet, without actually recording them in memory. This way, the avoidance of redundant evaluations does not require additional ressources.

The branching method and these improvements directly lead to the following algorithm:

A Generic Algorithm to Solve the SPP

Global Inputs:

c	a cost function;
$\mathcal{P}$	a set of admissible parts defining admissible partitions;
$\mathfrak{L}$	a set of locally-optimal admissible partitions of parts on which the algorithm has already been applied.

Local Inputs:

X	an admissible part;
$\overline{\mathcal{X}}$	the complementary partition of X inherited from the "higher" call ($\overline{\mathcal{X}}$ is a partition of $\Omega \backslash X$);
$\mathfrak{D}$	the set of admissible partitions which refinements have already been evaluated during "higher" calls.

Output:

$\mathcal{X}^*$	a locally-optimal admissible partition of X.

- If the algorithm has already been applied to part X, return the locally-optimal partition recorded in $\mathfrak{L}$.
- Initialization: $\mathcal{X}^* \leftarrow \{\{X\}\}$ and $\mathfrak{D}' \leftarrow \mathfrak{D}$.
- For each $\mathcal{Y} \in \mathfrak{C}(\{X\})$ such that $\overline{\mathcal{X}} \cup \mathcal{Y}$ does not refine any partition in $\mathfrak{D}$, do the following:
  - For each part $Y \in \mathcal{Y}$, call the algorithm with local inputs $X \leftarrow Y$, $\overline{\mathcal{X}} \leftarrow \overline{\mathcal{X}} \cup \mathcal{Y} \backslash \{Y\}$, and $\mathfrak{D} \leftarrow \mathfrak{D}'$ to compute a locally-optimal partition $\mathcal{Y}^*_Y \in \mathfrak{P}^*(Y)$.
  - $\mathcal{Y}^* \leftarrow \bigcup_{Y \in \mathcal{Y}} \mathcal{Y}^*_Y$.
  - If $c(\mathcal{Y}^*) > c(\mathcal{X}^*)$, then $\mathcal{X}^* \leftarrow \mathcal{Y}^*$.
  - $\mathfrak{D}' \leftarrow \mathfrak{D}' \cup \{\mathcal{Y}\}$.
- Return $\mathcal{X}^*$ and record this result in $\mathfrak{L}$.

4 From the Generic Framework to Specialized Implementations

The above algorithm may *in theory* be applied to any set of admissible parts $\mathcal{P} \subset 2^\Omega$. In that sense, it provides a dynamic solving of the *general* SPP. However, a generic implementation would not be computationally-optimal for *special versions* of the SPP. Indeed, assuming that one explicitly knows the partition set that is meant to be searched, one can adapt the algorithm and the data structures to the problems' specific algebraic structures. That is what we call *deriving a specialized implementation of the algorithm*. For example, in the general case, there are 2^n possible admissible parts (where n is the size of Ω), so one needs at least n bits to identify one of them (*e.g.*, a binary string of size n indicating which individuals are contained in the identified part). In the case of the OSPP, the admissible parts are the $\frac{1}{2}\, n\, (n-1)$ intervals of Ω, so one roughly needs $2 \log n$ bits to identify one of them (*e.g.*, two integers indicating the indexes of the interval extrema). Hence, the generic algorithm should be used as a starting point to build specialized ones. This section makes the specialization process explicite for the HSPP and the OSPP.

4.1 Solving the HSPP

As it has been previously defined, the HSPP arises when the set of admissible parts $\mathcal{P}$ forms a *hierarchy* (see 2.3.2).

Execution of the Generic Algorithm For each admissible part $X \in \mathcal{P}$, the corresponding maximal partition $\{X\}$ covers only one admissible partition, that is the set $\mathfrak{C}(X) = \{Y \in \mathcal{P} \backslash Y \subsetneq X, \nexists Z \in \mathcal{P}, Y \subsetneq Z \subsetneq X\}$ of children of the node X in the tree representing the hierarchy. Hence, the branching of the search space proposed in Eq. 3 is not redundant: $\mathfrak{P}(X) = \{\{X\}\} \cup \mathfrak{R}(\mathfrak{C}(X))$. The resulting algorithm is then a simple recursive procedure consisting in a depth-first search of the tree: a recursive execution on each $Y \in \mathfrak{C}(X)$ (step 2.a), the building of the resulting locally-optimal partition $\bigcup_{Y \in \mathfrak{C}(X)} \mathcal{Y}^*_Y$ (step 2.b), and the comparison of its cost with

the one of the maximal partition $\{X\}$ (step 3). The algorithm is naturally called only once per admissible part (no need for memoization) and the branching of the partition set is never redundant.

Data Structures and Implementation The hierarchy $\mathcal{P}$ is implemented by a tree data structure. Each node represents an admissible part $X \in \mathcal{P}$ and has three labels instantiated and exploited by the algorithm:

- $cost$ stores the cost $c(X)$ of the corresponding part;
- $optimalCost$ stores the cost $c(\mathcal{X}^*)$ of a locally-optimal partition of X;
- $optimalCut$ is a Boolean value that is *true* if and only if the maximal partition $\{X\}$ is optimal among the admissible partitions of X.

The depth-first search computes the $optimalCost$ and the $optimalCut$ of each node according to its $cost$ and to the sum of the $optimalCost$ of its children. After the execution, the optimal partition of Ω is the union of the higher nodes in the tree such that $optimalCut = true$.

Algorithm 2 for the HSPP

Require: A tree with a label $cost$ on each node representing the cost of the corresponding admissible part.

Ensure: Each node of the tree has a Boolean label $optimalCut$ representing an optimal partition (see above).

procedure SOLVEHSPP($node$)
  if $node$ has no child **then**
    $node.optimalCost \leftarrow node.cost$
    $node.optimalCut \leftarrow true$
  else
    $MCost \leftarrow node.cost$
    $\mu Cost \leftarrow 0$
    for each $child$ of $node$ **do**
      SOLVEHSPP($child$)
      $\mu Cost \leftarrow \mu Cost + child.optimalCost$
    $node.optimalCost \leftarrow \max(\mu Cost, MCost)$
    $node.optimalCut \leftarrow (\mu Cost < MCost)$

Linear Complexity The space complexity of this algorithm is bounded by the size of the data tree representing the hierarchy. As it contains between $n + 1$ (only the root and the leaves) and $2n - 1$ nodes (in the case of a complete binary tree), the space complexity is *linear*. The time complexity is the one of a depth-first search and is also *linear*, meeting the results of past algorithms dedicated to the HSPP (see 2.3.2).

4.2 Solving the OSPP

As it has been previously defined, the OSPP arises when the admissible parts are the intervals of Ω induced by a total order $<$ (see Subsection 2.3.3).

Execution of the Generic Algorithm Given a population Ω of n ordered individuals $x_1 < \ldots < x_n$, for each interval $[\![x_i, x_j]\!] = \{x_i, \ldots, x_j\}$ with $1 \leq i \leq j \leq n$, the admissible partitions covered by the maximal partition $\{[\![x_i, x_j]\!]\}$ are the couples of subintervals $\{[\![x_i, x_k]\!], [\![x_{k+1}, x_j]\!]\}$ with $i \leq k < j$. In the following, for the sake of conciseness, we simply mark such an interval $[i, j]$ and its covered partitions $[i, k][k+1, j]$. We thus have the following branching: $\mathfrak{C}([i, j]) = \{[i][i+1, j], \ldots, [i, j-1][j]\}$.

The generic algorithm is applied to $\Omega = [1, n]$. Let us assume that the covered partitions are evaluated (step 2) in the following order: $[1, n-1][n]$, $[1, n-2][n-1, n]$, $\ldots$, $[1][2, n]$ (see for example arrows **1**, **6** and **9** in Fig. 3). The covered partition $[1, n-1][n]$ is evaluated first (**1**). The algorithm is thus recursively applied (step 2.a) on part $[1, n-1]$ (**2**), then on part $[1, n-2]$ (**3**), and so on, until locally-optimal partitions of parts $[1], [1, 2], \ldots, [1, n-1]$ have been computed and recorded. All that remains is the computation of an optimal partition of part $[1, n]$. For the k^{th} evaluation, with $1 < k < n$, the covered partition $[1, n-k][n-k+1, n]$ has to be evaluated (for example **6**) knowing that the covered partitions $\{[1, n-i][n-i+1, n]\}_{1 \leq i < k}$ have already been evaluated along with their refined partitions. The algorithm is recursively applied to parts $[1, n-k]$ and $[n-k+1, n]$ (step 2.a):

- Since the algorithm has already been applied to part $[1, n-k]$ during the first evaluation, the optimal partition is simply read from memory (see cross below (1 2)).
- Regarding part $[n-k+1, n]$, all covered partitions are now considered for evaluation. But, since $[n-k+1, n-i][n-i+1, n]$ refines $[1, n-i][n-i+1, n]$ for all $1 \leq i < k$, each covered partition has already been evaluated during the previous evaluations. Hence, steps 2.a and 2.b may be avoided (see star on arrow **8** in Fig. 3) and the algorithm uses the maximal partition $[n-k+1, n]$.

To sum up, in order to compute an optimal partition $\mathcal{X}^*_{[1,n]}$, the generic algorithm recursively computes locally-optimal partitions $\mathcal{X}^*_{[1]}, \mathcal{X}^*_{[1,2]}, \ldots, \mathcal{X}^*_{[1,n-1]}$. Then, it exploits the results to compare partitions $\mathcal{X}^*_{[1]} \cup \{[2, n]\}, \ldots, \mathcal{X}^*_{[1,n-1]} \cup \{[n]\}$ and it returns one that has the highest cost.

Data Structures and Implementation The following specialized algorithm is a *bottom-up* implementation of the *top-down* generic algorithm. Each admissible part $[i, j] \in \mathcal{P}$ is represented as a couple of integer (i, j). The costs of admissible parts are recorded in a $n \times n$ upper triangular matrix $cost$. Each cell $cost[i, j]$ gives the cost $c([i, j])$ of the corresponding part. Optimal partitions are encoded in a vector $optimalCut$ containing n integers such that, for all $1 \leq j \leq n$, $optimalCut[j]$ is the indice of the first individual of the last part of an optimal partition of $[1, j]$. Hence, $optimalCut[n] = k$ indicates that part $[k, n]$ is in the optimal partition of $[1, n]$ and, if $k > 1$, then $optimalCut[k-1]$ again indicates the first individual of the last part of an optimal

partition of $[1, k-1]$, and so on. The optimal partition of $[1, n]$ thus consists in a sequence of indices $k_1, \ldots, k_m$ recorded in $optimalCut$ and indicating the m individuals where the population is divided: $[1, k_1 - 1][k_1, k_2 - 1] \ldots [k_m, n]$. The costs of these optimal partitions are recorded in a vector $optimalCost$ of size n. Each cell $optimalCost[j]$, with $1 \leq j \leq n$, gives the cost of the optimal partitions of part $[1, j]$. The algorithm iteratively runs through the triangular matrix to build the two vectors and thus computes an optimal admissible partition of Ω.

Algorithm 3 for the OSPP

Require: A matrix $cost$ recording the costs of intervals.
Ensure: The vector $optimalCut$ represents an optimal partition (see text above).

```
for j ∈ ⟦1, n⟧ do
    optimalCost[j] ← cost[1, j]
    optimalCut[j] ← 1
    for cut ∈ ⟦2, j⟧ do
        μCost ← optimalCost[cut − 1] + cost[cut, j]
        if μCost > optimalCost[j] then
            optimalCost[j] ← μCost
            optimalCut[j] ← cut
```

Quadratic Complexity For a population of size n, the upper triangular matrix contains $n(n-1)/2$ values and the two vectors each contains n integers. Hence, the space complexity is *quadratic*. In the proposed implementation, for each part $[1, j]$ with $1 \leq j \leq n$, the algorithm performs $j - 1$ comparisons to identify the optimal partitions among the covering ones. Hence, overall, $(n-1)(n-2)/2$ comparisons are performed and the time complexity is also *quadratic*. This result meets the ones of the previous algorithms that have been developed for the OSPP (see Subsection 2.3.3).

5 Conclusion

By making strong assumptions regarding the structure of the search space, dozens of problems have been modeled as tractable versions of the SPP. The algorithmic framework we propose in this paper provides a unified dynamic programming approach to design such computationally-efficient algorithms by exploiting the algebraic properties of such structures. We have shown how this framework applies on two well-known versions of the SPP, for structures induced by a *hierarchy* (HSPP) and by a *total order* (OSPP). By following the same specialization steps (formalization of admissible parts and admissible partitions, analysis of the generic algorithm execution, and design of data structures that fits with the induced algebraic structure), this programming method can be applied to numerous other versions expressing interesting spatial or temporal properties: *e.g.*, partitioning graphs in connected components, partitioning partially ordered sets representing causal relations, partitioning the state space of a dynamical process, partitioning multidimensional populations mixing spatial and temporal constraints.

For any new version of the SPP, the computational complexity of the corresponding specialized implementation will be bounded from below by the number of admissible parts. For both versions addressed in this paper, the proposed algorithms achieve such a lower bound ($O(n)$ for the HSPP and $O(n^2)$ for the OSPP). However, this is not always the case (*e.g.*, for a circular order, the number of admissible part is $O(n^2)$, but dynamic programming only provides a $O(n^3)$ algorithm [25]). This leads to an interesting research question: *for which versions of the SPP does the framework provide an optimization algorithm which complexity is bounded by the number of admissible parts?*

References

[1] C.J. Alpert and A.B. Kahng. "Recent developments in netlist partitioning: a survey". In: *Integration: the VLSI Journal* 19 (1-2 1995), pp. 1–81.

[2] S. Anily and A. Federgruen. "Structured Partitioning Problems". In: *Operations Research* 39.1 (Jan. 1991), pp. 130–149.

[3] Y. Bachrach et al. "Optimal Coalition Structure Generation in Cooperative Graph Games". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. Ed. by Marie desJardins and Michael L. Littman. Bellevue, Washington, USA: AAAI Press, July 14-18, 2013, pp. 81–87.

[4] E. Balas and M.W. Padberg. "Set Partitioning: A Survey". In: *SIAM Review* 18.4 (1976), pp. 710–760.

[5] R. Becker et al. "Max-Min Partitioning of Grid Graphs into Connected Components". In: *Networks* 32.2 (1998), pp. 115–125.

[6] A. Björklund, T. Husfeldt, and M. Koivisto. "Set Partitioning via Inclusion-Exclusion". In: *SIAM Journal on Computing* 39.2 (2009), pp. 546–563.

[7] M.A. Boschetti, A. Mingozzi, and S. Ricciardelli. "A dual ascent procedure for the set partitioning problem". In: *Discrete Optimization* 5.4 (2008), pp. 735–747.

[8] P. Brucker. "On the Complexity of Clustering Problems". In: *Optimization and Operations Research*. Lecture Notes in Economics and Mathematical Systems 157 (1978), pp. 45–54.

[9] A. K. Chakravarty, J. B. Orlin, and U. G. Rothblum. "A Partitioning Problem with Additive Objective with an Application to Optimal Inventory Groupings for Joint Replenishment". In: *Operations Research* 30.5 (1982), pp. 1018–1022.

[10] C. Chekuri and A. Ene. "Approximation Algorithms for Submodular Multiway Partition". In: *Proceedings of the 2011 IEEE Fifty-second Annual Symposium on Foundations of Computer Science (FOCS'11)*. Oct. 2011, pp. 807–816.

[11] P.C. Chu and J.E. Beasley. "Constraint Handling in Genetic Algorithms: The Set Partitioning Problem". In: *Journal of Heuristics* 4.4 (1998), pp. 323–357.

[12] D. Dosimont et al. “A Spatiotemporal Data Aggregation Technique for Performance Analysis of Large-scale Execution Traces”. In: *Proceedings of the 2014 IEEE International Conference on Cluster Computing (CLUSTER'14)*. Madrid, Spain: IEEE, Sept. 22-26, 2014.

[13] K. Hoffman and M. Padberg. “Set Covering, Packing and Partitioning Problems”. In: *Encyclopedia of Optimization*. Ed. by Christodoulos A. Floudas and Panos M. Pardalos. Springer US, 2001, pp. 2348–2352.

[14] K.L. Hoffman and T.K. Ralphs. “Integer and Combinatorial Optimization”. In: *Encyclopedia of Operations Research and Management Science*. Ed. by Saul I. Gass and Michael C. Fu. Springer US, 2013, pp. 771–783.

[15] Brad Jackson et al. “An Algorithm for Optimal Partitioning of Data on an Interval”. In: *IEEE Signal Processing Letters* 12.2 (2005), pp. 105–108.

[16] R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent. “Building the Best Macroscopic Representations of Complex Multi-Agent Systems”. In: *Transactions on Computational Collective Intelligence*. LNCS. Forthcoming. Springer-Verlag Berlin, Heidelberg, 2014.

[17] R. Lamarche-Perrin et al. “Evaluating Trace Aggregation for Performance Visualization of Large Distributed Systems”. In: *Proceedings of the 2014 IEEE International Symposium on Performance Analysis of Systems and Software*. 2014.

[18] D. Lehmann, R. Müller, and T. Sandholm. “The Winner Determination Problem”. In: *Combinatorial Auctions*. Ed. by Peter Cramton, Yoav Shoham, and Richard Steinberg. MIT Press, 2006, pp. 297–317.

[19] A. Mingozzi and S. Morigi. “Partitioning a matrix with non-guillotine cuts to minimize the maximum cost”. In: *Discrete Applied Mathematics* 116.3 (2002), pp. 243–260.

[20] S. Muthukrishnan and T. Suel. “Approximation algorithms for array partitioning problems”. In: *Journal of Algorithms* 54.1 (2005), pp. 85–104.

[21] G. Pagano et al. “Trace Management and Analysis for Embedded Systems”. In: *Proceedings of the 7th International Symposium on Embedded Multicore SoCs (MCSoC'13)*. IEEE Computer Society Press, 2013.

[22] J. Pintér and G. Pesti. “Set partition by globally optimized cluster seed points”. In: *European J. of Operational Research* 51 (1991), pp. 127–135.

[23] P. Pons and M. Latapy. “Post-processing hierarchical community structures: Quality improvements and multi-scale view”. In: *Theoretical Computer Science* 412.8-10 (2011), pp. 892–900.

[24] T. Rahwan et al. “An Anytime Algorithm for Optimal Coalition Structure Generation”. In: *Journal of Artificial Intelligence Research* 34.1 (Jan. 2009), pp. 521–567.

[25] M.H. Rothkopf, A. Pekeč, and R.M. Harstad. “Computationally Manageable Combinational Auctions”. In: *Management Science* 44.8 (Aug. 1998), pp. 1131–1147.

[26] T. Sandholm et al. “Coalition structure generation with worst case guarantees”. In: *Artificial Intelligence* 111.1-2 (1999), pp. 209–238.

[27] R.V.V. Vidal. “Optimal Partition of an Interval – The Discrete Version”. In: *Applied Simulated Annealing*. Vol. 396. LNEMS. Springer Berlin, Heidelberg, 1993, pp. 291–312.

[28] D. Yun Yeh. “A Dynamic Programming Approach to the Complete Set Partitioning Problem”. In: *BIT Numerical Mathematics* 26.4 (1986), pp. 467–474.

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

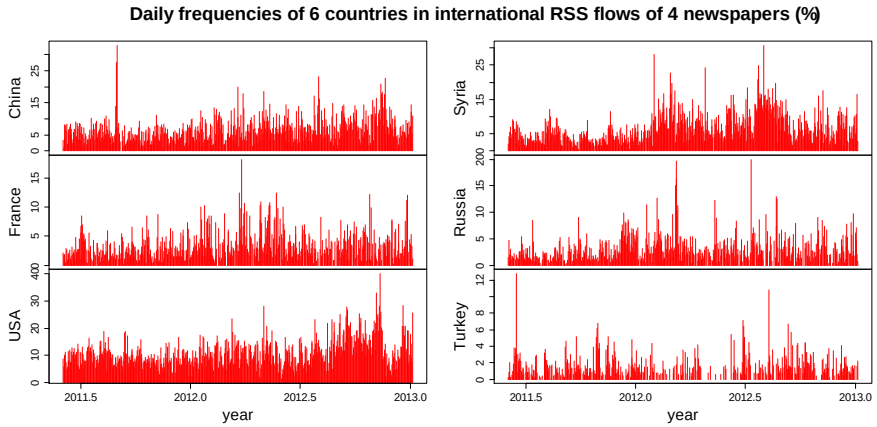
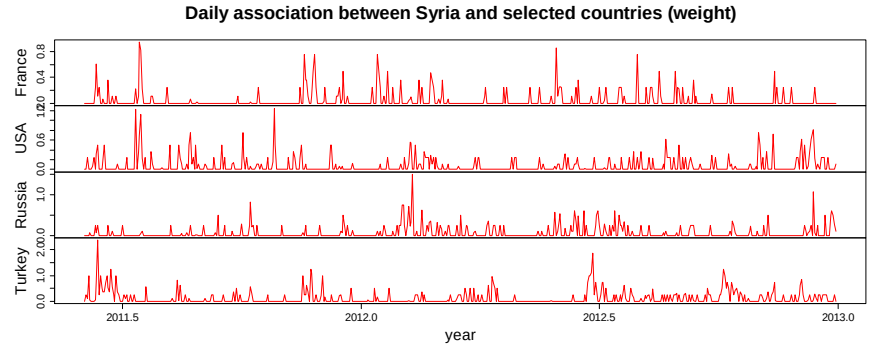


Figure 2 indicates the weighted number of items where Syria is associated to France, Turkey, Russia or USA

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

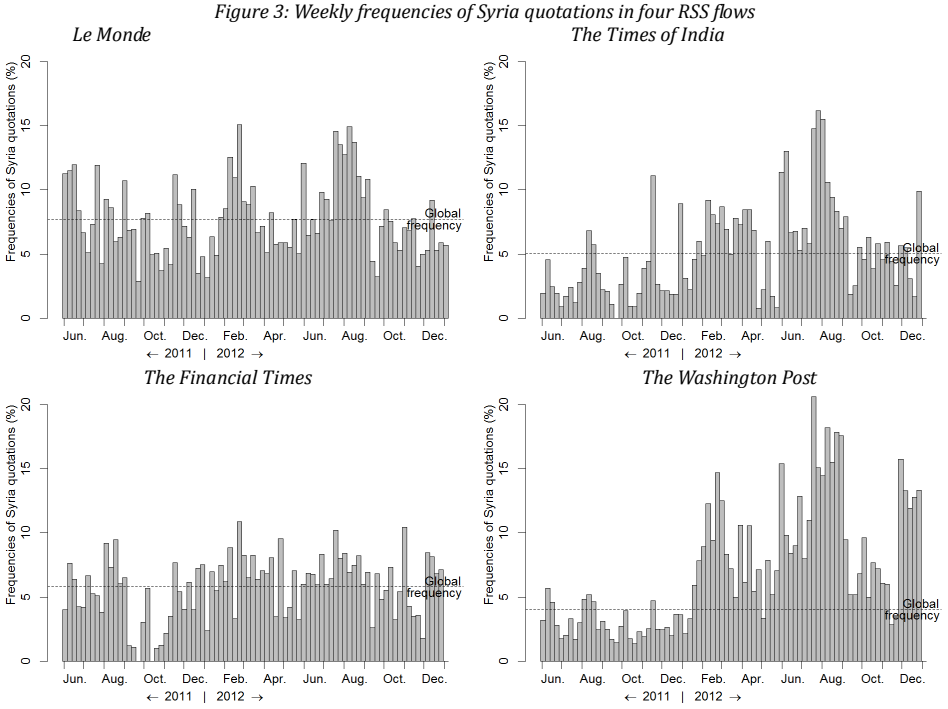
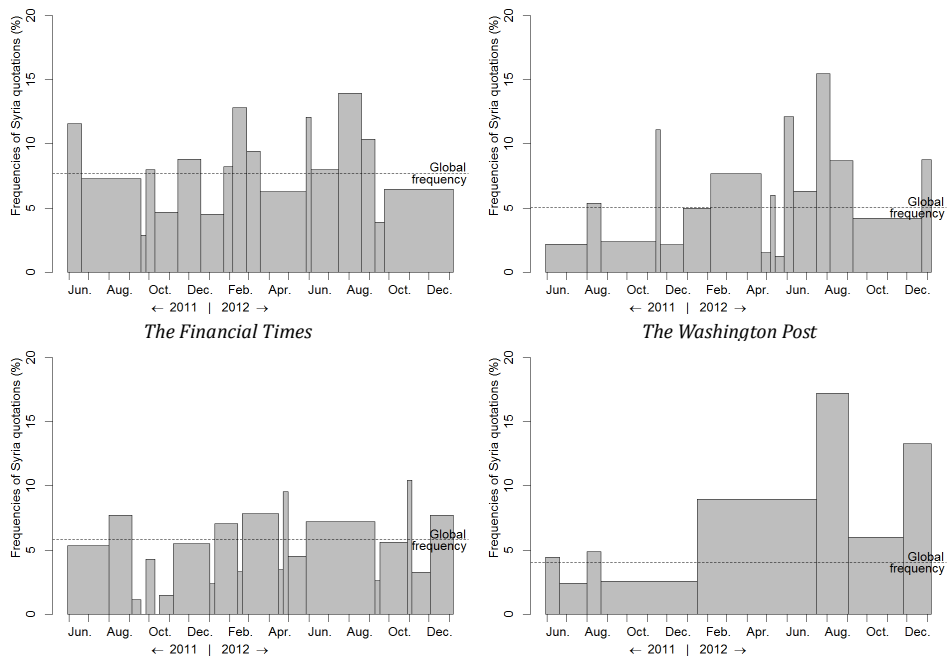


Figure 4: Optimal time aggregation of the frequencies of Syria quotations in four RSS flows



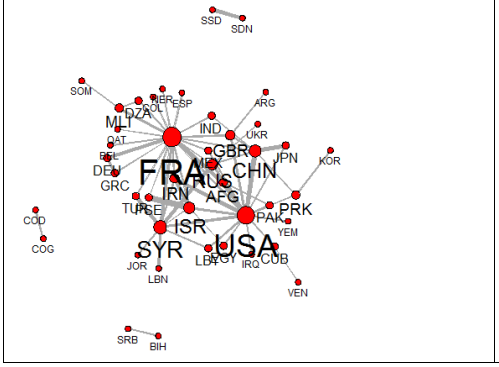
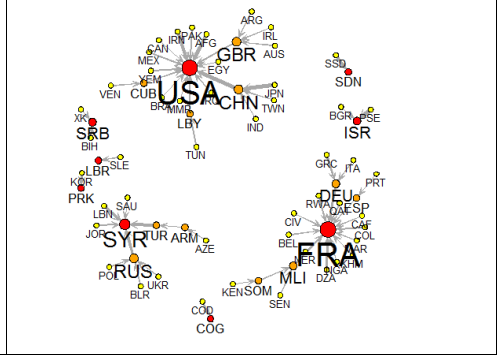
3) EVALUATION OF THE AUTONOMY OF SYRIAN CRISIS AS REGARD TO THE LINKAGES BETWEEN COUNTRIES

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.**

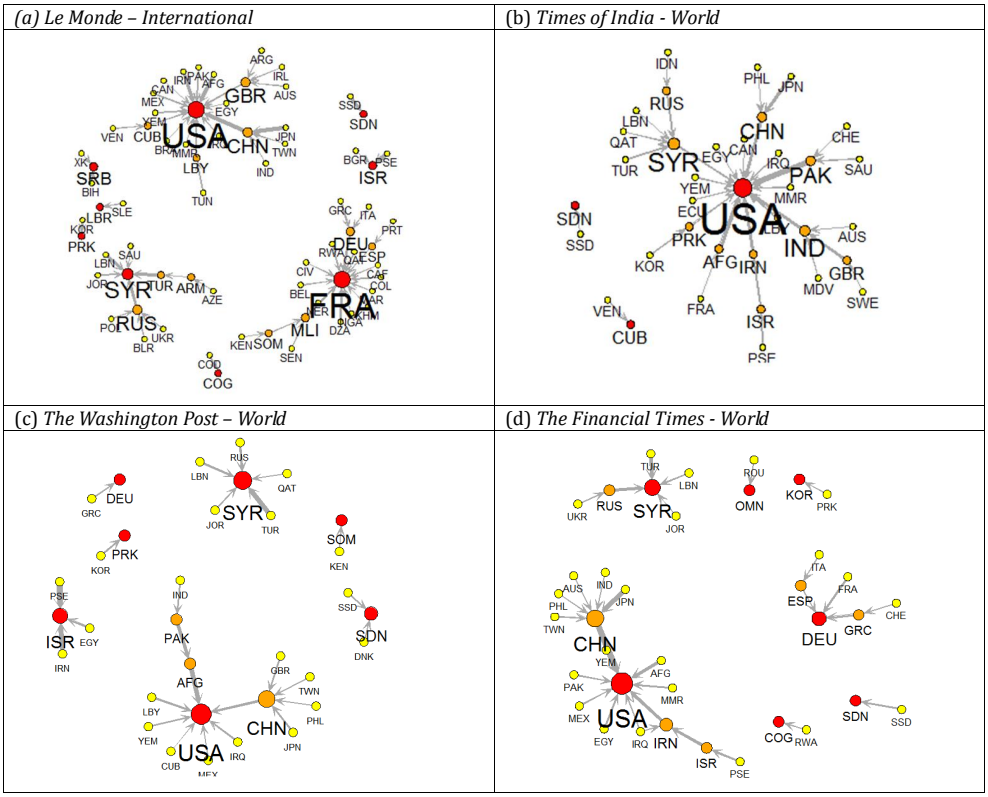
Figure 5: Result of the dominant flows method applied to items sent by the RSS flow of Le Monde in 2012

(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



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

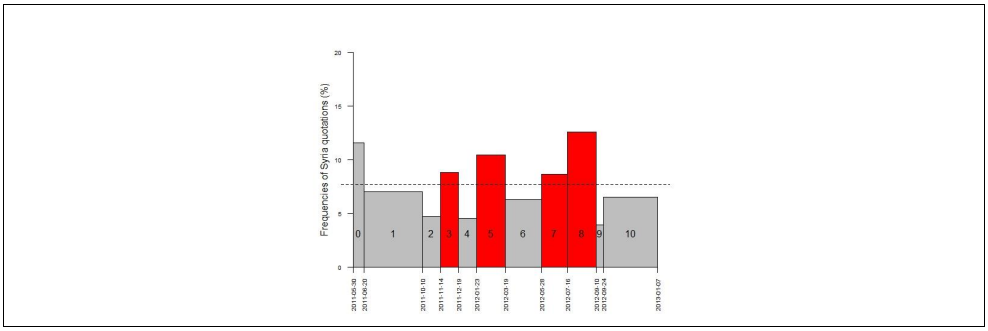
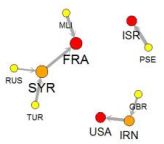
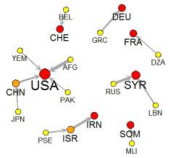


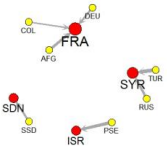
Figure 8: Graphs of dominants flows from* Le Monde *at the four red time periods



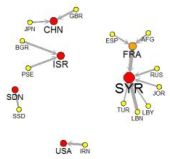
Period 3



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." In *Proc. of the 6th 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?" *in Proc. 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