HAL
open science

The Best-partitions Problem: How to Build Meaningful Aggregations ?

Robin Lamarche-Perrin, Yves Demazeau, Jean-Marc Vincent

To cite this version:

Robin Lamarche-Perrin, Yves Demazeau, Jean-Marc Vincent. The Best-partitions Problem: How to Build Meaningful Aggregations ?. [Research Report] RR-LIG-044, 2013, pp.18. hal-00947934

HAL Id: hal-00947934

https://inria.hal.science/hal-00947934

Submitted on 18 Feb 2014

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

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

The Best-partitions Problem: How to Build Meaningful Aggregations?

Laboratoire d'Informatique de Grenoble
February 11th, 2013

Robin Lamarche-Perrin
Université de Grenoble
Robin.Lamarche-Perrin@imag.fr

Yves Demazeau
CNRS
Yves.Demazeau@imag.fr

Jean-Marc Vincent
Université Joseph Fourier
Jean-Marc.Vincent@imag.fr

Abstract

The design and the debugging of large distributed AI systems require abstraction tools to build tractable macroscopic descriptions. Data aggregation can provide such abstractions by partitioning the systems dimensions into aggregated pieces of information. This process leads to information losses, so the partitions should be chosen with the greatest caution, but in an acceptable computational time. While the number of possible partitions grows exponentially with the size of the system, we propose an algorithm that exploits exogenous constraints regarding the system semantics to find best partitions in a linear or polynomial time. We detail two constrained sets of partitions that are respectively applied to temporal and spatial aggregation of an agent-based model of international relations. The algorithm succeeds in providing meaningful high-level abstractions for the system analysis.

Keywords: Data aggregation, multi-agent systems, algorithmic complexity, spatial and temporal analysis, news analysis.

1 Introduction

The design, debugging, and optimization of large-scale distributed AI systems need tools that proceed at high level of description, with insightful abstractions regarding the system global dynamics. Among abstraction techniques (dimension reduction in multivariate analysis, generalization in machine learning, subsetting, segmentation, and so on [3]), this report focuses on *data aggregation*. It consists in loosing some information about the agent level to build simpler – yet meaningful – macroscopic descriptions. Such a process is not harmless: an unfortunate aggregation may lead to a critical misunderstanding of the system behavior. Hence, we have to determine what are the *good* abstractions and how to properly choose them.

Given a dimension of the system, each partition of its elements determines a potential aggregation. Yet, the number of partitions of a set grows exponentially with its size. Finding the partitions that maximize a given measure of quality is a computationally-expensive problem. In this report, we propose an algorithm that finds the "best partitions" in a linear or polynomial time, assuming that: (1) parts of the dimension can be evaluated independently one from each other and (2) all partitions are not necessarily eligible for aggregation. We show that, by introducing exogenous constraints into the best-partitions problem, the computational complexity of the algorithm can be highly reduced. In a multi-agent perspective, we present and apply two constrained sets of partitions, inherited from the spatial and temporal semantics of multi-agent systems (MAS).

The state of the art in section 2 shows that the MAS community lacks of tools to build high-level abstractions. Section 3 formalizes the best-partitions problem and presents an algorithm that efficiently solves it. Section 4 presents two constrained sets of admissible partitions, their semantics in MAS analysis, and their associated complexity. Section 5 introduces information-theoretic quality measures to define what are the "good" partitions depending on their *complexity reduction* and their *information loss*. Section 6 applies the algorithm and measures for the analysis of international relations through press media. Temporal and spatial dimensions are aggregated to build meaningful multi-resolution representations of the dataset. Section 7 concludes this research report and gives some perspectives.

2 Macroscopic Analysis of Large-scale MAS

This section shows that very few research efforts have been done in the MAS domain to evaluate and choose the right aggregations. In a comprehensive survey of agent-based simulation platforms [14], Railsback *et al.* evaluate simulation tools by implementing classical features of MAS modeling and analysis. Unfortunately,

the abstraction problem is not tackled, thus indicating that such considerations are seldom if ever taken into account. Indeed, most platforms (Java Swarm, Repast, MASON, NetLogo and Objective-C Swarm) are confined to the microscopic simulation of agents. Railsback warns against the lake of "a complete tool for statistical output" in these platforms. 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 [18], or provided without any feedback regarding their quality [1]. In the ASGARD monitoring system [17], the level of detail is grounded on the distance between the observer and the agents in a 3D space. Such a visual aggregation is not controlled by the user and, worst, it does not give feedback regarding the induced information loss.

Some techniques from graph analysis and data clustering build groups of agents out of their *microscopic properties* (see for example [16, 13, 7]). Such considerations may meet ours on 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 [5]. 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 in order to handle complex MAS. They should use consistent semantics from the experts and provide tools to choose the best macroscopic descriptions.

3 Solving the Best-partitions Problem

Finding the partition that maximizes a given measure is a complex computational problem. Indeed, the number of partition of a set grows exponentially with its size. This section formalizes the best-partitions problem and its computational complexity (3.1). Then it presents an algorithm that solves the problem without the need of evaluating each partition independently (3.2). In section 4, we show that the complexity of the best-partitions problem can be reduced by constraining the set of meaningful partitions.

3.1 The Best-partitions Problem

Definition 1. *A* population Ω *is a set of individuals* $\{x_1, \ldots, x_n\}$*. We mark* $|\Omega| = n$ *the* size *of the population,* $\mathcal{P}$ *the set of its parts and* $\mathfrak{P}$ *the set of its partitions.*

The population generically represents a dimension of the MAS that can be aggregated: agents, simulation steps, events, communications, and so on. Each part X in $\mathcal{P}$ represents an aggregate that summarizes the behavior of its underlying individuals. Each partition $\mathcal{X}$ in $\mathfrak{P}$ represents a way of using aggregates to build a macroscopic description of the population, both *complete* (each individual is *at least* in one aggregate) and *non-redundant* (resp. *at most* in one aggregate).

Definition 2. *A* measure of partitions quality m *is an application that associates a value in* $\mathbb{R}$ *to each partition in* $\mathfrak{P}$*.*

The measure m represents the quality of the aggregated descriptions induced by the partitions. It thus gives a criterion to choose the aggregates that will summarize the MAS dimensions, according to an expected quality level.

Definition 3. *The* best partitions $\overline{\mathfrak{P}}$*, according to a measure* m*, are the partitions in* $\mathfrak{P}$ *that maximizes* m*:*

$$\overline{\mathfrak{P}} = \arg\max_{\mathcal{X} \in \mathfrak{P}} m(\mathcal{X})$$

Computing $\overline{\mathfrak{P}}$ is what we call the *best-partitions problem*.

The problem complexity is related to the number of partitions. For a population of size n, it is given by the Bell formula:

$$|\mathfrak{P}| = B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k = O\left(\left(\frac{n}{\ln(n)}\right)^n\right)$$

Hence, $|\mathfrak{P}|$ grows worse than exponentially with the size of Ω. So we need a way to find best partitions without measuring every possible ones.

3.2 An Algorithm Based on the Sum Property

An interesting algebraic property of quality measures can make the best-partitions problem tractable, namely the *sum property* [2].

Definition 4. *A measure of partitions quality has the* sum property *if it is the additive extension of a measure on parts:*

$$\forall \mathcal{X} \in \mathfrak{P}, \quad m(\mathcal{X}) = \sum_{X \in \mathcal{X}} m(X)$$

In other words, a measure has the sum property if it can be defined as the sum of the measures of its parts. A trivial example of such a measure is the *size of partitions*.

$$m(\mathcal{X}) = |\mathcal{X}| = \sum_{X \in \mathcal{X}} 1$$

In section 5, in order to build meaningful aggregations of MAS, we propose more relevant quality measures.

The sum property guaranties that the quality of an aggregate does not depend on the way other individuals are themselves aggregated (in fact, it is a stronger hypothesis since it assumes that the measure m is also additive).

Definition 5. *A partition $\mathcal{X}$ is a* refinement *of a partition $\mathcal{Y}$ if every part in $\mathcal{X}$ is a subset of some part in $\mathcal{Y}$. We mark $\mathcal{X} \prec \mathcal{Y}$. Moreover, $\mathcal{X}$ is* covered *by $\mathcal{Y}$ if there is no partition $\mathcal{Z}$ such that $\mathcal{X} \prec \mathcal{Z} \prec \mathcal{Y}$. In this case, we mark $\mathcal{X} \prec\!: \mathcal{Y}$.*

The covering relation thus expresses the atomic disaggregations that can be applied to a given partition. Therefore, once we computed the best partitions $\overline{\mathfrak{P}}(X)$ of a given part X, it can be used to find the best partitions among all covering ones. Moreover, the best *refining* partitions of a partition $\mathcal{X}$ can be recursively found as *the Cartesian product of the best partitions of its parts*:

$$\overline{\mathfrak{P}}(\mathcal{X}) = \bigtimes_{X \in \mathcal{X}} \overline{\mathfrak{P}}(X)$$

Besides, each partition of a part X is either a refining partition of a covering partition of $\{X\}$, or the $\{X\}$ partition itself. Hence, best partitions can also be recursively founded in the following set of partitions:

$$\{X\} \cup \left(\bigcup_{\mathcal{Y} \prec\!: \{X\}} \overline{\mathfrak{P}}(\mathcal{Y}) \right)$$

The following algorithm uses this "refinement and covering structure" to make the least possible evaluations. It begins by the whole aggregated population and proceeds step by step by comparing it with the covered partitions, expressing the possible atomic desegregations one can perform. To evaluate these next partitions, the algorithm is recursively applied on each of their parts.

Algorithm. *Find the best partitions $\overline{\mathfrak{P}}$ wrt to a measure m:*

1. Get the partitions in $\mathfrak{P}$ that are covered by $\{\Omega\}$:

$$\mathrm{cov}(\{\Omega\}) = \{\mathcal{X} \in \mathfrak{P} \backslash \mathcal{X} \prec: \{\Omega\}\}$$

2. For each covered partition $\mathcal{X} \in \mathrm{cov}(\{\Omega\})$:

   Recursively find the best partitions $\overline{\mathfrak{P}}(X)$ of each part $X \in \mathcal{X}$;

   Merge these best partitions to get the best refining partitions of $\mathcal{X}$:

$$\overline{\mathfrak{P}}(\mathcal{X}) = \bigtimes_{X \in \mathcal{X}} \overline{\mathfrak{P}}(X)$$

3. Compare the best refining partitions, and the partition $\{\Omega\}$, with each other;

4. Return the ones that maximize m.

3.3 Algorithmic Complexity

The intermediary results of the recursion can be stored so that the algorithm is applied only once to each part X in $\mathcal{P}$. Regarding the complexity in space, hereafter marked $\mathrm{space}(\mathfrak{P})$, the memory size required by the algorithm is then proportional to the number of parts $|\mathcal{P}| = 2^n$. Hence, the space complexity is **exponential**:

$$\mathrm{space}(\mathfrak{P}) = \Theta(2^n)$$

To determine the complexity in time, hereafter marked $\mathrm{time}(\mathfrak{P})$, we are interested in the total number of comparisons (step 3 of the algorithm). There are $\binom{n}{m}$ parts of size m in $\mathcal{P}$ and each as exactly $(2^{m-1} - 1)$ covering partitions.

$$\begin{aligned}
\mathrm{time}(\mathfrak{P}) &= \sum_{k=0}^{n} \binom{n}{k} (2^{k-1} - 1) \\
&= \frac{1}{2} \sum_{k=0}^{n} \binom{n}{k} 2^k 1^{n-k} - \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} \\
&= \frac{3^n}{2} - 2^n
\end{aligned}$$

Hence:

$$\mathrm{time}(\mathfrak{P}) = \Theta(3^n)$$

Even if these complexities are much lower than the total number of partitions $|\mathfrak{P}|$, the algorithm as an **exponential complexity** in space and in time. Therefore, it cannot be exploited in practice for the whole set of partitions $\mathfrak{P}$. Classically, heuristics may be developed to partially solve the problem in a non-exponential time. However, we claim that most real-world problems bring into play a kind or another of exogenous constraints that, by introducing some semantics, can reduce the complexity of the best-partitions problem.

4 Introducing Semantics in the Best-partitions Problem

In classical clustering problems individuals are considered regardless of any exogenous topological properties. However, in MAS analysis, the system dimensions often have a precise semantics. Agents are for example organized in groups, roles, or tasks. Events, communications, and dynamics rely on the natural order of time. These semantics are essential for the analysis and, thereby, they should be included in the aggregation process.

4.1 The Constrained Best-partitions Problem

The set of partitions $\mathfrak{P}$ corresponds to an unconstrained aggregation. Let $\mathfrak{P}_a$ be a subset of *admissible partitions*. The *constrained best-partitions problem* consists in computing:

$$\overline{\mathfrak{P}_a} = \arg\max_{\mathcal{X} \in \mathfrak{P}_a} m(\mathcal{X})$$

Our algorithm can be applied to every subset $\mathfrak{P}_a$ of admissible partitions. Its complexity depends on the "covering structure" of such a set. In particular, the less a partition admits covered partitions within $\mathfrak{P}_a$, *i.e.* the more the "covering structure" is constrained, the more the algorithm is tractable.

In the following subsections, we present two sets of admissible partitions that are essential for MAS analysis: from a hierarchy and from an order. For each, we indicate:

- $|\mathfrak{P}_a|$, the number of admissible partitions;
- $\mathrm{space}(\mathfrak{P}_a)$, the number of underlying parts, to which the memory size needed by the algorithm is proportional;
- $\mathrm{time}(\mathfrak{P}_a)$, the number of comparisons computed by the algorithm (step 3), determining its time complexity.

4.2 Admissible Partitions from a Hierarchy

Agents in MAS can be hierarchically organized: *e.g.* for decision making [12], for multi-level simulation [5], from a network analysis point-of-view [7]. Admissible partitions should be consistent with such organizations by forbidding aggregates that mix disjoint elements of the hierarchy. In section 6, we apply this organizational pattern to the spatial aggregation of world countries [4]. Hierarchies used by geographers to build global statistics about world areas are exploited to compute geographically consistent aggregations.

Definition 6. *A* hierarchy $\mathcal{H}$ *on a population* Ω *is a subset of* $\mathcal{P}$ *such that every parts in* $\mathcal{H}$ *are either disjoint or included in each other. A* hierarchical set of partitions $\mathfrak{P}_{\mathcal{H}}$ *is the subset of* $\mathfrak{P}$ *containing all partitions which parts are elements of* $\mathcal{H}$*.*

Fig. 1 represents a hierarchy of 5 individuals $\{M, U, S, I, C\}$. More generally, a hierarchically organized population of n individuals cannot have more than $(2n - 1)$ parts. Hence:

$$\text{space}(\mathfrak{P}_{\mathcal{H}}) = O(n)$$

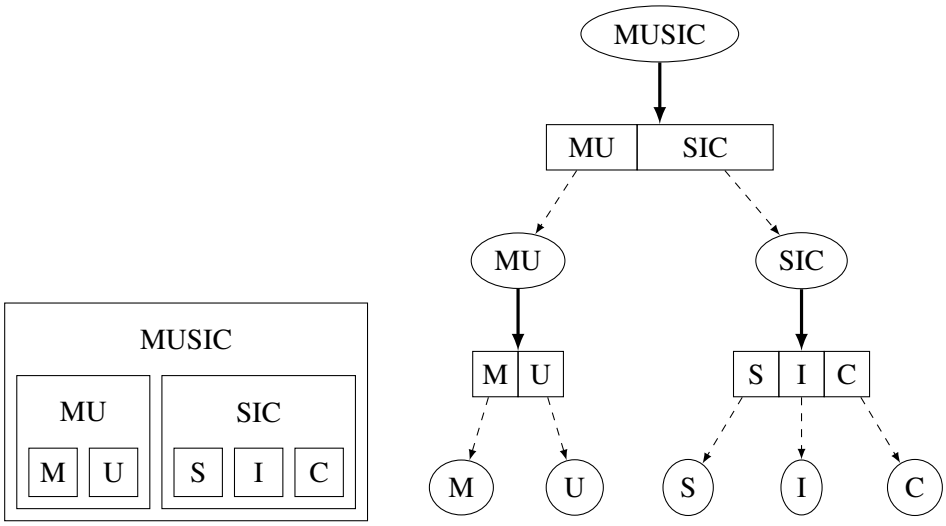

Figure 1: On the left: the nested-box representation of a 5-individuals hierarchy. [Each box corresponds to an admissible part to build partitions.] On the right: the associated "covering and refinement structure" followed by the algorithm. [Dashed arrows, from a partition to one of its parts, represent recursive calls (step 2). Thick arrows, from a part to a covered partition, represent comparisons (step 3).]

This complexity corresponds to the largest possible hierarchies (binary trees). For less complete hierarchies, it can be highly reduced. The "covering structure" of a hierarchy forms a simple tree (see thick arrows in Fig. 1) where *each part has only one covered partition*. Hence:

$$\text{time}(\mathfrak{P}_{\mathcal{H}}) = \text{space}(\mathfrak{P}_{\mathcal{H}}) = O(n)$$

The number of admissible partitions still grows exponentially. Indeed, for each fork of the tree, the number of admissible partitions within a branch is multiplied by those of the other branches. An analysis of the series $U_n = (U_{n-1})^2 + 1$ gives the following asymptotic behavior:

$$|\mathfrak{P}_{\mathcal{H}}| = O(c^n) \quad \text{where } c \approx 1,2$$

As a consequence, in this very constrained case, even if the number of induced partitions still grows exponentially, the algorithm has a **linear complexity** in space and in time.

4.3 Admissible Partitions from an Order

A MAS dimension may also be subject to an order, often when it has a temporal aspect. In this cases, the partitions should preserve this relation by aggregating only intervals of individuals. For example, in section 6, such an ordered structure is exploited to build consistent aggregations of weeks.

Definition 7. *Assuming a total order $<$ on Ω, the* set of ordered partitions $\mathfrak{P}_<$ *is the subset of $\mathfrak{P}$ containing all partitions which parts are intervals $[x, y]$ of individuals ($x < y$).*

For a population of n individuals, there are $(n + 1 - k)$ parts of size k.

$$\text{space}(\mathfrak{P}_<) = \sum_{k=1}^{n} (n + 1 - k) = \frac{1}{2}(n + 1)n$$

Hence:

$$\text{space}(\mathfrak{P}_<) = \Theta(n^2)$$

This corresponds to the number of intervals $[x, y]$ one can build with *two* degrees of freedom.

Fig. 2 represents the "covering structure" of the admissible partitions for an ordered population of size 4. It is far more intricate than for a hierarchical set. We can see that each part of size k as $k-1$ covered partitions.

$$\mathrm{time}(\mathfrak{P}_<) = \sum_{k=1}^{n} (n+1-k)(k-1) = \frac{1}{6}(n+1)n(n-1)$$

Hence:

$$\mathrm{time}(\mathfrak{P}_<) = \Theta(n^3)$$

In this case equally, the number of admissible partitions grows exponentially with the size of the population ($|\mathfrak{P}_<| = \Theta(2^n)$), but the algorithm still has a **polynomial complexity** in space and in time.

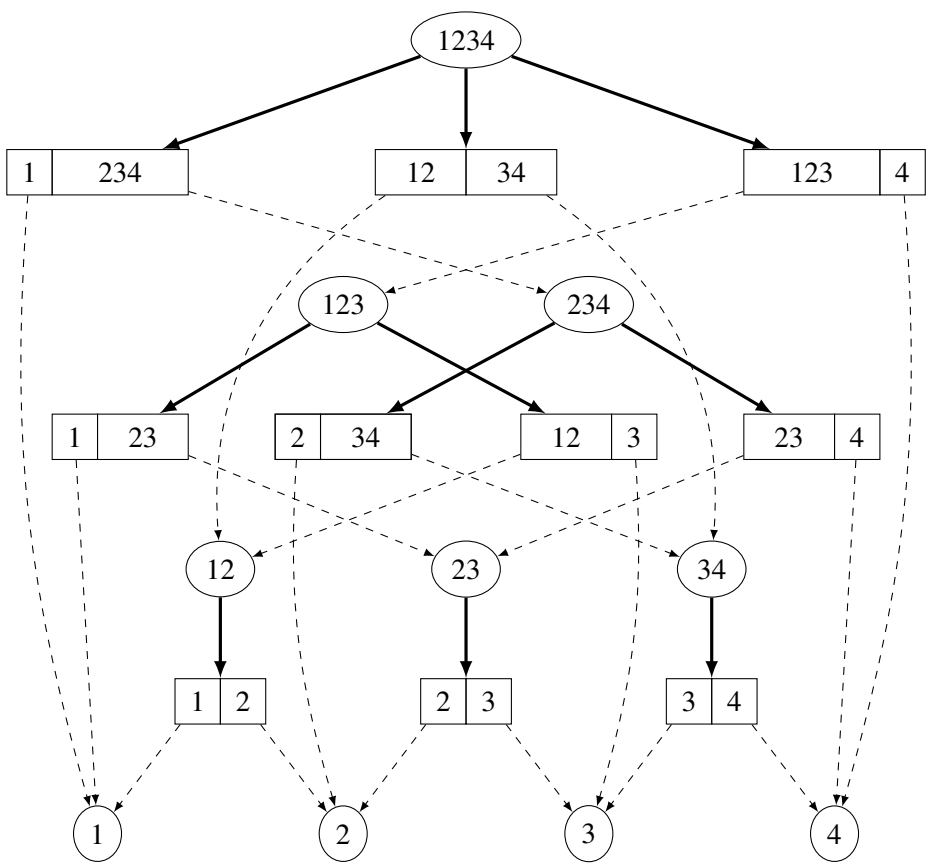

Figure 2: The "covering and refinement structure" of a 4-individuals ordered population. [As in Fig. 1, dashed arrows represent recursive calls of the algorithm and thick arrows represent executed comparisons.]

4.4 Multi-dimensional Partitions

Given two populations Ω_1 and Ω_2, a set of bidimensional individuals is given by the Cartesian product $\Omega_1 \times \Omega_2$. Bidimensional parts are rectangles. Bidimensional partitions correspond to couples of unidimensional partitions within $\mathfrak{P}_1 \times \mathfrak{P}_2$. A multi-dimensional set of partitions thus expresses the simultaneous aggregation of several correlated populations.

Definition 8. *Let $\mathfrak{P}_1$ and $\mathfrak{P}_2$ be two sets of admissible partitions on two populations Ω_1 and Ω_2. The induced* multi-dimensional set of admissible partitions *on $\Omega_1 \times \Omega_2$ is given by the Cartesian product $\mathfrak{P}_1 \times \mathfrak{P}_2$.*

The number of admissible partitions is multiplied, as well as the number of admissible parts:

$$|\mathfrak{P}_1 \times \mathfrak{P}_2| = |\mathfrak{P}_1||\mathfrak{P}_2|$$

$$\mathrm{space}(\mathfrak{P}_1 \times \mathfrak{P}_2) = \mathrm{space}(\mathfrak{P}_1)\,\mathrm{space}(\mathfrak{P}_2)$$

Each bidimensional part has the covered partitions within its own dimension, plus those of the other dimension. A simple computation gives:

$$\mathrm{time}(\mathfrak{P}_1 \times \mathfrak{P}_2) = \mathrm{time}(\mathfrak{P}_1)\,\mathrm{space}(\mathfrak{P}_2) + \mathrm{time}(\mathfrak{P}_2)\,\mathrm{space}(\mathfrak{P}_1)$$

Therefore, when we simultaneously aggregate hierarchical or ordered dimensions, the algorithm still has a **polynomial complexity** in space and in time.

5 Quality Measures from Information Theory

This section presents some measures of partitions quality that fit the sum property. Their interest is discussed in previous works [11, 9]. We do not claim that they are the only relevant measures, but that they express interesting properties and meet the MAS analysis purposes.

5.0.1 The Dual Problem of Aggregation

We consider an attribut that takes a real positive value $v(x)$ for each individual x in Ω (*e.g.* size or weight of the individual, its events or communications number). It can be additively extended on parts as follows:

$$\forall X \in \mathcal{P}, \quad v(X) = \sum_{x \in X} v(x)$$

Thus, we mark $v(\Omega)$ the sum of all the individual values. Given a partition $\mathcal{X}$ of Ω, the set of values $\{v(X)\}_{X\in\mathcal{X}}$ provides a *description* of the population attribute. It is *microscopic* when all the individual values are indicated: $\{v(x)\}_{x\in\Omega}$, and *fully aggregated* when only the sum of all values is indicated: $\{v(\Omega)\}$. In order to simplify the system descriptions, aggregation looses information regarding the individual values. Hence, two aspects of partitions should be evaluated: (1) How much a given partition reduces the description complexity? (2) How much, by doing so, it loses precious information about it?

- The gain measure estimates the reduction of complexity between two partitions: $\mathrm{gain}(\mathcal{X}) = C(\mathcal{X}_0) - C(\mathcal{X})$, where $\mathcal{X}_0$ is a partition of reference (for example the microscopic partition).
- In the same way, the loss measure estimate the information lost by aggregating $\mathcal{X}$ wrt a partition $\mathcal{X}_0$.

These two measures can then be combined in an *parametrized information criterion* PIC that expresses the trade-off between such the *gain* ratio and the *loss* ratio:

$$\mathrm{PIC}_\alpha(\mathcal{X}) = \alpha\left(\frac{\mathrm{gain}(\mathcal{X})}{\mathrm{gain}(\{\Omega\})}\right) - (1-\alpha)\left(\frac{\mathrm{loss}(\mathcal{X})}{\mathrm{loss}(\{\Omega\})}\right)$$

where $\alpha \in [0,1]$ is the parameter used to balance the trade-off. For $\alpha = 0$, maximizing the PIC is equivalent to minimizing the loss: the best partition is always the microscopic one. For $\alpha = 1$, it is the fully aggregated one. When α varies between 0 and 1, a whole class of nested partitions arises. The choice of this parameter is deliberately left to the user, so she can adapt the level description to her varying requirements: between the expected amount of details and the available computational resources.

Shannon Entropy as a Measure of Complexity

The Shannon entropy estimates the quantity of information needed to encode a given description. From Shannon's formula [15], we get (see [11] for details):

$$\mathrm{gain}(\mathcal{X}) = \sum_{X\in\mathcal{X}}\left(v(X)\log_2 v(X) - \sum_{x\in X} v(x)\log_2 v(x)\right)$$

The entropy is a classical complexity measure: the higher it is for a given aggregate, the more information we save.

Kullback-Leibler Divergence as a Measure of Information Loss

The Kullback-Leibler divergence estimates the information quantity wasted during the aggregation process. From Kullback's and Leibler's formula [8], we get:

$$\text{loss}(\mathcal{X}) = \sum_{X \in \mathcal{X}} \left(\sum_{x \in X} v(x) \times \log_2 \left(v(x) \times \frac{|X|}{v(P)} \right) \right)$$

Interestingly, an aggregate which internal distribution is very *heterogeneous* has a very high divergence, indicating an important information loss.

6 Application to News Analysis

Our global approach is currently being evaluated as part of a multidisciplinary research project, in collaboration with geographers and media experts [4]. It aims at designing tools for the visualization and the analysis of international relations within print media. The concept of *event* is at stake, in particular regarding its spatial and temporal granularity. Indeed, the description level of news carries a strong meaning for social scientists. The project aims at defining useful abstractions for the analysis, in space, in time, and wrt to the thematic dimension.

This section presents early applications of the best-partitions algorithm to illustrate its results and to convince of its interest for the analysis of MAS. We work with two populations:

- Ω_c contains $n_c = 168$ countries (the system agents)
- Ω_w contains $n_w = 90$ weeks (the system steps)

We work with 60,000 papers of *The Guardian*. For each couple of individuals (c, w) in $\Omega_c \times \Omega_w$, we count the number of papers published during the week w that cite the country c. This gives us the values $v(c, w)$ of an attribute that expresses the media weight of a country at a given time. These values are interpreted as the microscopic description of the MAS activity. The following experiments shows how it can efficiently be aggregated by our approach to help analyzing news.

6.1 Temporal Aggregation

The set of weeks Ω_w is naturally ordered by the "arrow of time". We mark $\mathfrak{P}_w$ the corresponding set of ordered partitions. Given a country $c \in \Omega_c$, the algorithm can be applied to the corresponding values $v(c, \Omega_w) = \{v(c, w) \backslash w \in \Omega_w\}$. Among

the $|\mathfrak{P}_w| = 2^{n_w-1} = 6.19 \times 10^{26}$ admissible partitions, the algorithm computes $\overline{\mathfrak{P}}$ in $\mathrm{time}(\mathfrak{P}_w) = \frac{1}{6}(n_w+1)n_w(n_w-1) = 1.21 \times 10^5$ steps

The following results concern $c =$ Tunisia. The left plot in Fig. 3 gives the gain and loss ratio of the best partition depending on the value of the α parameter. It presents the set of best partitions that one can choose depending on the expected level of details and the corresponding information loss. Fig. 4 explicits the microscopic descriptions ($\alpha = 0$) and two aggregated descriptions. For $\alpha = 0.4$, only 13 values are displayed, 9 of them being aggregated values. Some global patterns – as peaks and homogeneous periods – can instantaneously be spotted. For $\alpha = 0.6$, the distribution is roughly approximated by 3 aggregated values. It shows that the number of citations regarding Tunisia has been globally decreasing over the last two years. In both cases, the temporal order is preserved by the aggregation.

6.2 Spatial Aggregation

In a geographical analysis context, partitions can be defined according to world topological or socio-economical properties. We consider *WUTS* – a hierarchical organizations of Ω_c that defines 5-level nested partitions of world countries [6]. These aggregates are used by geographers to build global statistics about world areas, from the microscopic level of agents to the full aggregation. We mark $\mathfrak{P}_c$ the corresponding hierarchical set of partitions. Given a week $w \in \Omega_w$, the algorithm can be applied to the corresponding values $v(\Omega_c, w) = \{v(c, w) \backslash c \in \Omega_c\}$. Among the $|\mathfrak{P}_c| = 3.8 \times 10^{12}$ admissible partitions, the algorithm computes $\overline{\mathfrak{P}}$ in $\mathrm{time}(\mathfrak{P}_c) = 64$ comparisons (the number of non-individual aggregates in the *WUTS* hierarchy).

The following results concern $w =$ the week from the 7th to the 13th of January, 2013. Fig. 3 presents the gain and the loss of the best partitions depending on the value of the α parameter. Fig. 5 explicits two of these partitions, for $\alpha = 0.4$ and $\alpha = 0.5$. The best-partitions algorithm leads to multi-resolution maps of the worlds that are coherent with its topological properties. For example, on the second map, 6 values are represented, 3 of which are aggregated: for Latin American, Euro-African, and Asian-Pacific countries. These three aggregates do correspond to commonly used high-level abstractions for geographical analysis.

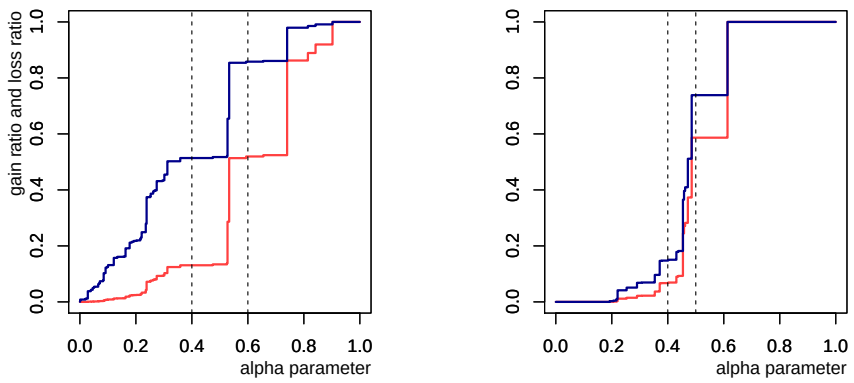

Figure 3: The gain and loss ratio (resp. black and gray lines) of best partitions, for temporal and spatial aggregation (resp. left and right plots). [Dashed vertical lines correspond to the value of the alpha parameter for aggregations displayed resp. in Fig. 4 and Fig. 5.]

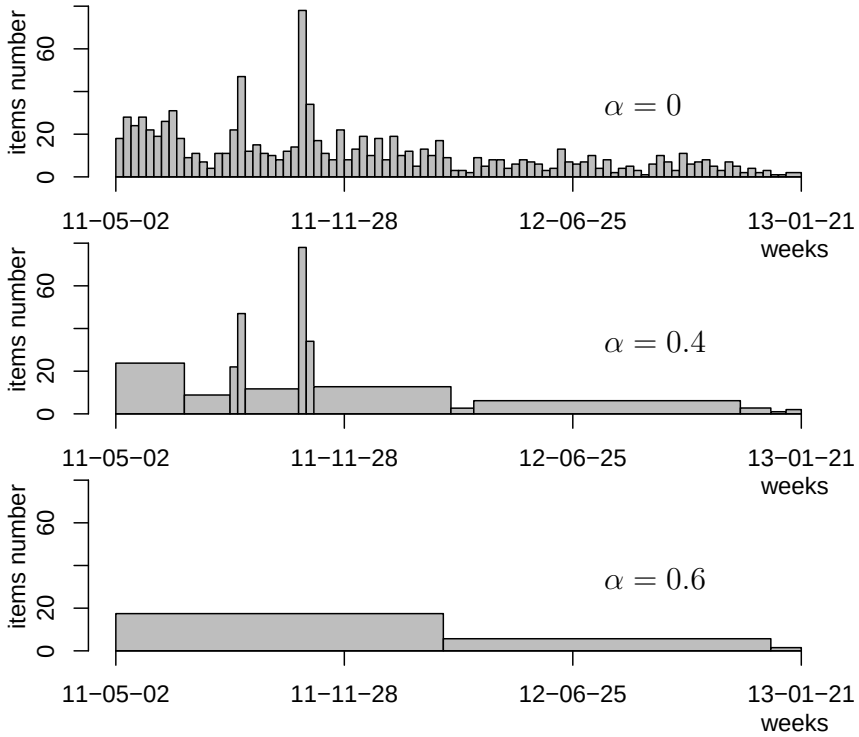

Figure 4: Temporal aggregation of the Tunisia citations, from the weeks level. [The best-partitions algorithm computes the aggregations that preserves the natural order of time, while maximizing our information-theoretic measure.]

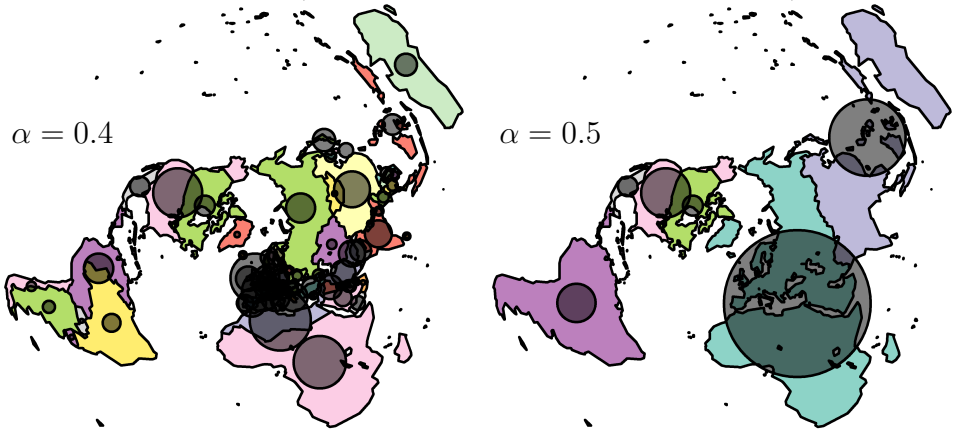

Figure 5: Spatial aggregation of a week in January, from the countries level. [The size of circles correspond to the number of citations regarding world areas.]

7 Conclusion and Perspectives

In this report, we propose an algorithm to efficiently solve the best-partitions problem on the basis of two essential assumptions: (1) the measure to be optimized has the *sum property* [2] and (2) the semantics of aggregated dimensions can be used to reduce the set of admissible partitions. Given that, the best-partitions algorithm is able to efficiently build meaningful aggregated descriptions for MAS analysis. The algorithm is illustrated on a dataset of newspaper articles. The temporal dimension is aggregated in a polynomial time and the resulting best-partitions are consistent with the temporal order. The spatial dimension is aggregated in a linear time, thanks to a hierarchical organization inherited from geographers. It leads to multi-resolution maps that are consistent with the geographical constraints.

We are currently exploiting this approach further for performance visualization of large-scale distributed systems [10]. This kind of application shows that the best-partitions algorithm can be scaled up to 1 million agents on the spatial dimension. To continue this research effort, other interesting constraints for MAS analysis should be investigated, along with their associated algorithmic complexity. This approach may also be generalized from partitions to all *incomplete* and/or *redundant* sets of parts. This would allow a greater variety of admissible macroscopic descriptions.

References

[1] L. Búrdalo, A. Terrasa, V. Julián, and A. García-Fornes. A Tracing System Architecture for Self-adaptive Multiagent Systems. In Y. Demazeau, F. Dignum, J. M. Corchado, and J. B. Pérez, editors, *Advances in Practical Applications of Agents and Multiagent Systems*, volume 70 of *Advances in Intelligent and Soft Computing*, pages 205–210. Springer Berlin, Heidelberg, 2010.

[2] I. Csiszár. Axiomatic Characterizations of Information Measures. *Entropy*, 10(3):261–273, 2008.

[3] N. Elmqvist and J.-D. Fekete. Hierarchical Aggregation for Information Visualization: Overview, Techniques, and Design Guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439–454, May/June 2010.

[4] P. GEOMEDIA. Geomedia: Observatoire des flux géomédiatiques internationaux, 2011.

[5] J. Gil-Quijano, T. Louail, and G. Hutzler. From Biological to Urban Cells: Lessons from Three Multilevel Agent-Based Models. In N. Desai, A. Liu, and M. Winikoff, editors, *Proceedings of the 13th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA-2010)*, volume 7057 of *Lecture Notes in Artificial Intelligence*, pages 620–635, Kolkata, Inde, November 2012. Springer-Verlag Berlin, Heidelberg.

[6] C. Grasland and C. Didelon. Europe in the World – Final Report (Vol. 1). Technical report, ESPON project 3.4.1, December 2007.

[7] P. Iravani. Multi-level Network Analysis of Multi-agent Systems. In *RoboCup 2008: Robot Soccer World Cup XII*, volume 5399 of *Lecture Notes in Artificial Intelligence*, pages 495–506. Springer-Verlag Berlin, Heidelberg, 2009.

[8] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[9] R. Lamarche-Perrin, Y. Demazeau, and J.-M. Vincent. How to Build the Best Macroscopic Description of your Multi-agent System? Application to News Analysis of International Relations. In Y. Demazeau and T. Ishida, editors, *Advances in Practical Applications of Agents and Multiagent Systems*, volume 73 of *Lecture Notes in Computer Science/Lecture Notes in Artificial Intelligence*. Springer-Verlag Berlin, Heidelberg, 2013. Forthcoming.

[10] R. Lamarche-Perrin, L. M. Shnorr, J.-M. Vincent, and Y. Demazeau. Evaluating Trace Aggregation Through Entropy Measures for Optimal Performance Visualization of Large Distributed Systems. Research Report RR-LIG-XXX, Laboratoire d'Informatique de Grenoble, Grenoble, France, 2012.

[11] R. Lamarche-Perrin, J.-M. Vincent, and Y. Demazeau. Informational Measures of Aggregation for Complex Systems Analysis. Research Report RR-LIG-026, Laboratoire d'Informatique de Grenoble, Grenoble, France, 2012.

[12] D. Nieuwenborgh, M. Vos, S. Heymans, and D. Vermeir. Hierarchical Decision Making in Multi-agent Systems Using Answer Set Programming. In K. Inoue, K. Satoh, and F. Toni, editors, *Computational Logic in Multi-Agent Systems*, volume 4371 of *Lecture Notes in Computer Science*, pages 20–40. Springer Berlin, Heidelberg, 2007.

[13] W. Peng, W. Krueger, A. Grushin, P. Carlos, V. Manikonda, and M. Santos. Graph-based methods for the analysis of large-scale multiagent systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, volume 1 of *AAMAS'09*, pages 545–552, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.

[14] S. F. Railsback, S. L. Lytinen, and S. K. Jackson. Agent-based Simulation Platforms: Review and Development Recommendations. *Simulation*, 82(9):609–623, September 2006.

[15] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423,623–656, 1948.

[16] A. Sharpanskykh and J. Treur. Group Abstraction for Large-Scale Agent-Based Social Diffusion Models with Unaffected Agents. In D. Kinny, J. Y.-j. Hsu, G. Governatori, and A. K. Ghose, editors, *Agents in Principle, Agents in Practice*, volume 7047 of *Lecture Notes in Computer Science*, pages 129–142. Springer Berlin, Heidelberg, 2011.

[17] J. Tonn and S. Kaiser. ASGARD A Graphical Monitoring Tool for Distributed Agent Infrastructures. In Y. Demazeau, F. Dignum, J. M. Corchado, and J. B. Pérez, editors, *Advances in Practical Applications of Agents and Multiagent Systems*, volume 70 of *Advances in Intelligent and Soft Computing*, pages 163–173. Springer Berlin, Heidelberg, 2010.

[18] M. H. Van Liedekerke and N. M. Avouris. Debugging multi-agent systems. *Information and Software Technology*, 37(2):103–112, 1995.