NTU

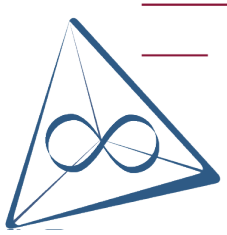
Singapore, 10 December 2015

Data Aggregation and Multiscale Analysis of Complex Systems

Robin Lamarche-Perrin

Yves Demazeau

Jean-Marc Vincent



Max Planck Institute for
Mathematics
in the Sciences

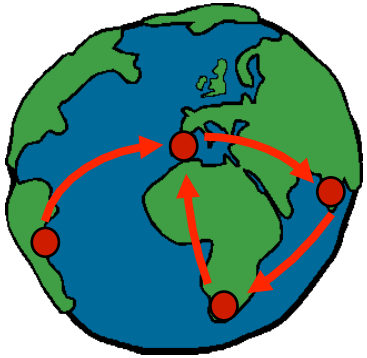
The Analysis of Large-scale MAS

How to provide a macroscopic overview of microscopic processes?

Analysis of International Relations

Geographer

International System



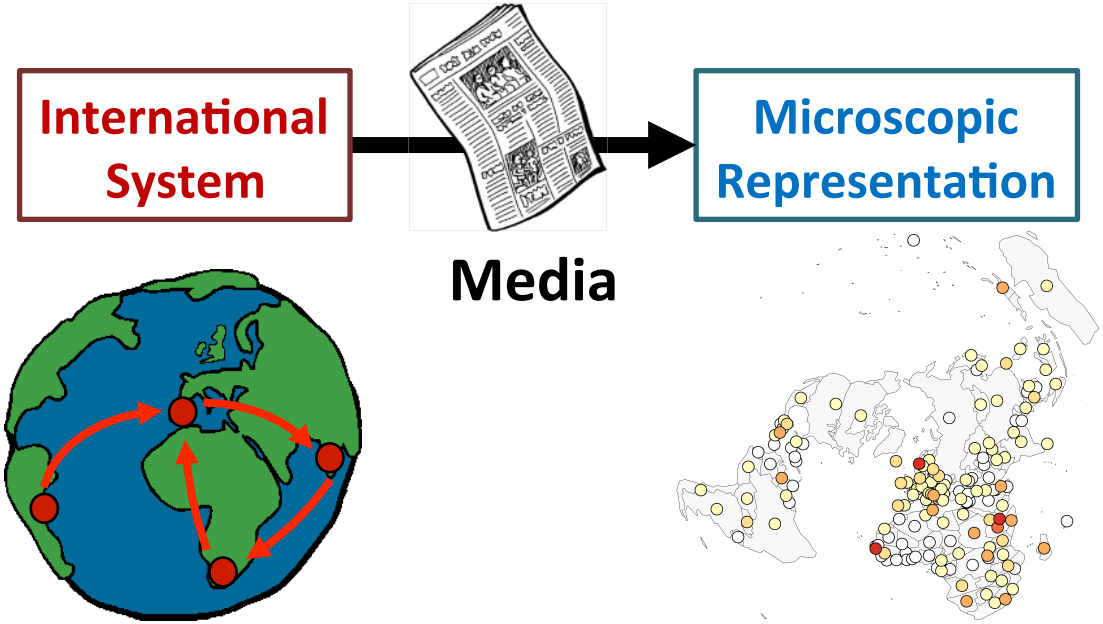
World observation

Analysis of International Relations

Hypothesis: media constitute an adequate instrument to observe the national level

[Grasland *et al.*, 2011]

Geographer

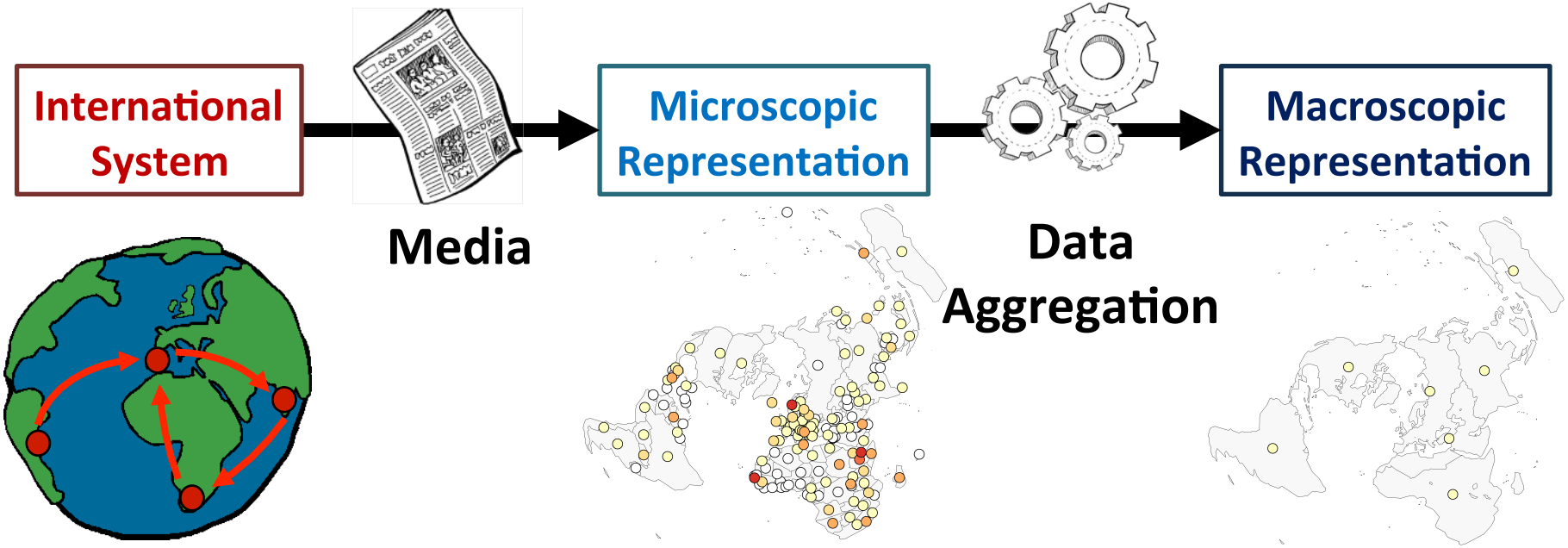


Analysis of International Relations

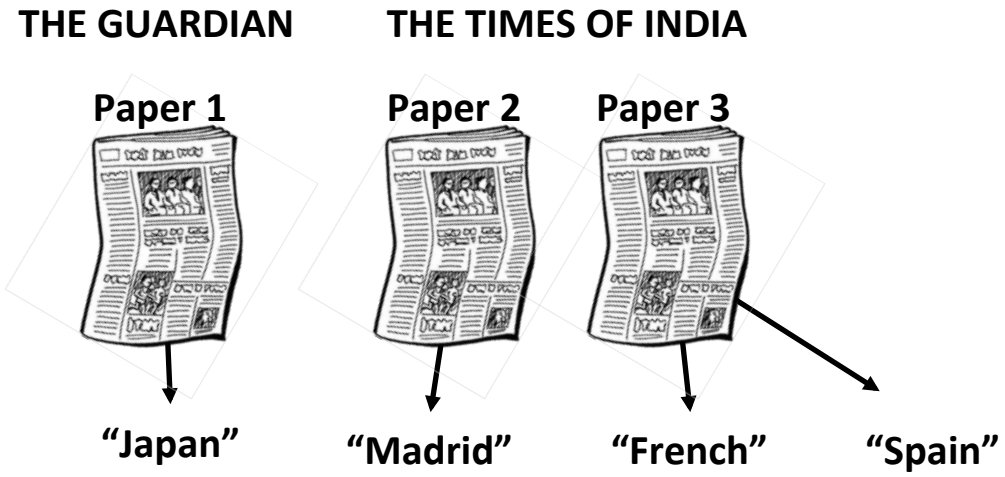
Hypothesis: media constitute an adequate instrument to observe the national level

[Grasland *et al.*, 2011]

Geographer



Data from Print Media



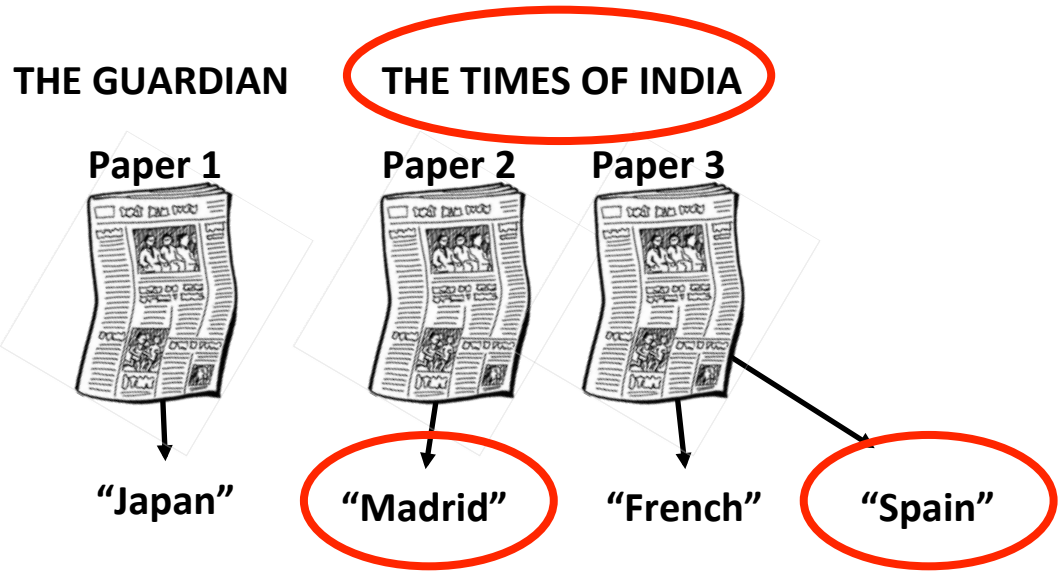
The GEOMEDIA Database
(ANR CORPUS GUI-AAP-04)

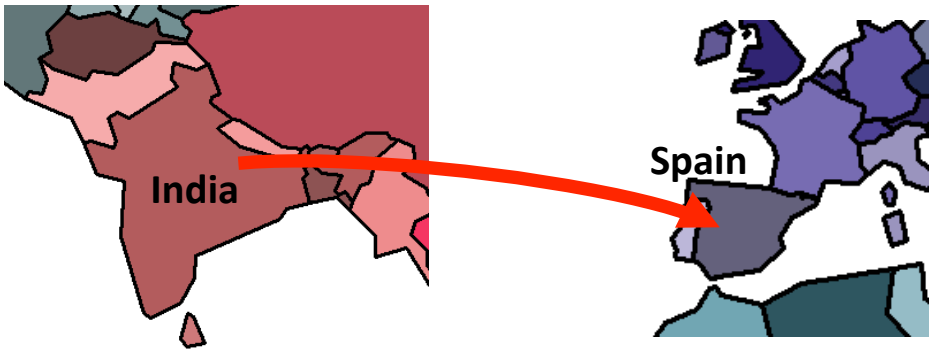
150 newspapers

1,944,000 papers

GEOGRAPHIC INFORMATION
193 countries (UN members)

Data from Print Media





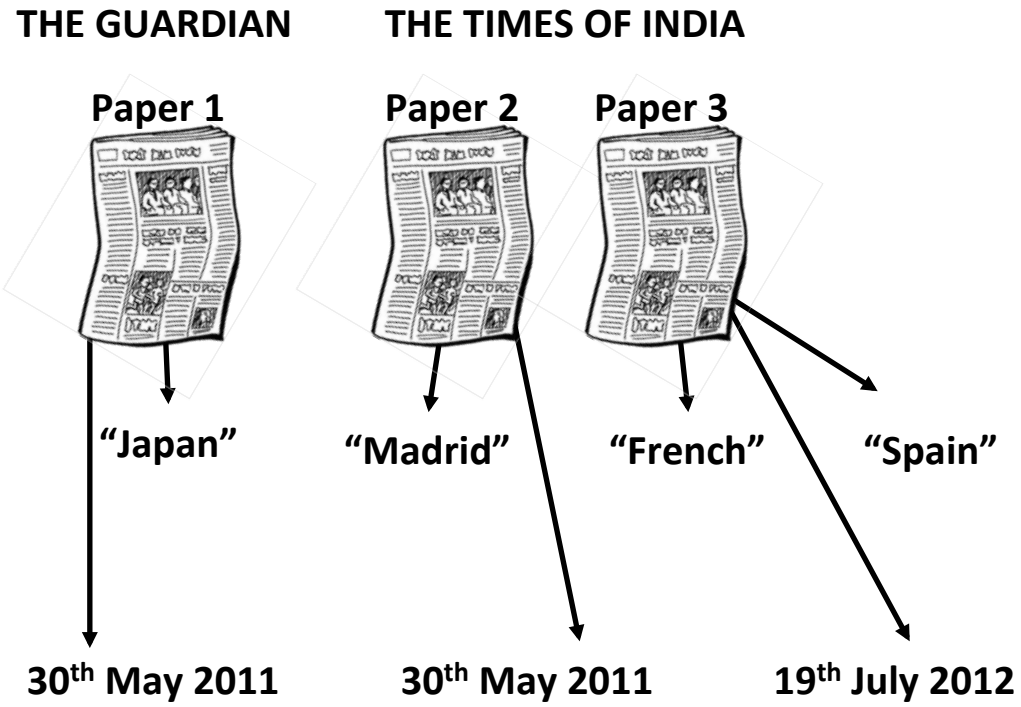
The GEOMEDIA Database
(ANR CORPUS GUI-AAP-04)

150 newspapers

1,944,000 papers

GEOGRAPHIC INFORMATION
193 countries (UN members)

Data from Print Media



The GEOMEDIA Database
(ANR CORPUS GUI-AAP-04)

150 newspapers

1,944,000 papers

GEOGRAPHIC INFORMATION

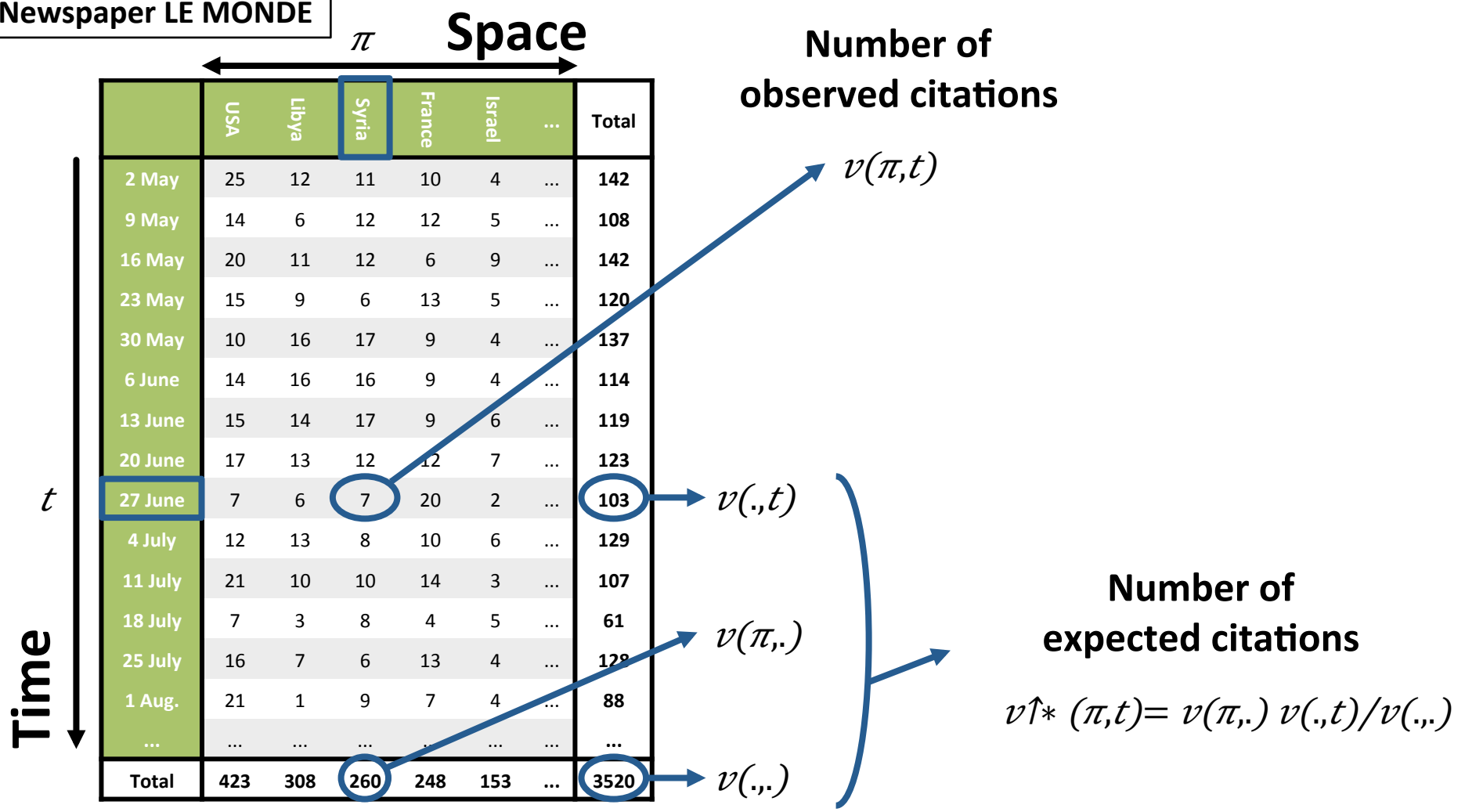
193 countries (UN members)

TEMPORAL INFORMATION

889 days / 127 weeks
(from the 3rd May 2011 to today)

Microscopic Representation

Newspaper LE MONDE



	USA	Libya	Syria	France	Israel	...	Total
2 May	25	12	11	10	4	...	142
9 May	14	6	12	12	5	...	108
16 May	20	11	12	6	9	...	142
23 May	15	9	6	13	5	...	120
30 May	10	16	17	9	4	...	137
6 June	14	16	16	9	4	...	114
13 June	15	14	17	9	6	...	119
20 June	17	13	12	12	7	...	123
27 June	7	6	7	20	2	...	103
4 July	12	13	8	10	6	...	129
11 July	21	10	10	14	3	...	107
18 July	7	3	8	4	5	...	61
25 July	16	7	6	13	4	...	128
1 Aug.	21	1	9	7	4	...	88
...	...	...	...	...	...	...	...
Total	423	308	260	248	153	...	3520

Knowledge Representation (micro)

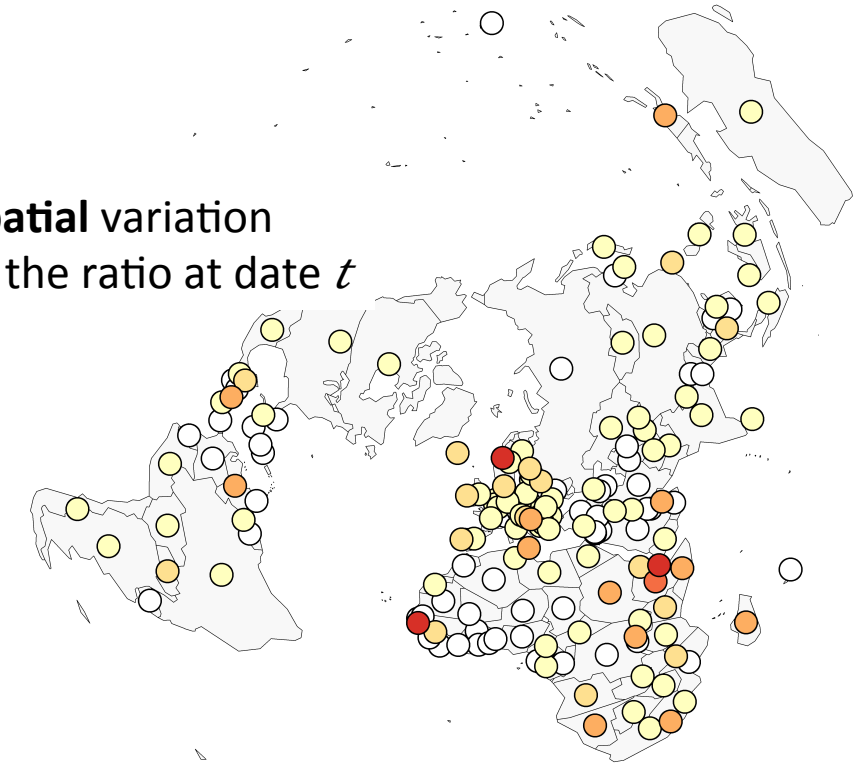
Space

	USA	Libya	Syria	France	Israel	…	Total
2 May	25	12	11	10	4	…	**142**
9 May	14	6	12	12	5	…	**108**
16 May	20	11	12	6	9	…	**142**
23 May	15	9	6	13	5	…	**120**
30 May	10	16	17	9	4	…	**137**
6 June	14	16	16	9	4	…	**114**
13 June	15	14	17	9	6	…	**119**
20 June	17	13	12	12	7	…	**123**
t 27 June	7	6	7	20	2	…	**103**
4 July	12	13	8	10	6	…	**129**
11 July	21	10	10	14	3	…	**107**
18 July	7	3	8	4	5	…	**61**
25 July	16	7	6	13	4	…	**128**
1 Aug.	21	1	9	7	4	…	**88**
…	…	…	…	…	…	…	…
Total	**423**	**308**	**260**	**248**	**153**	…	**3520**

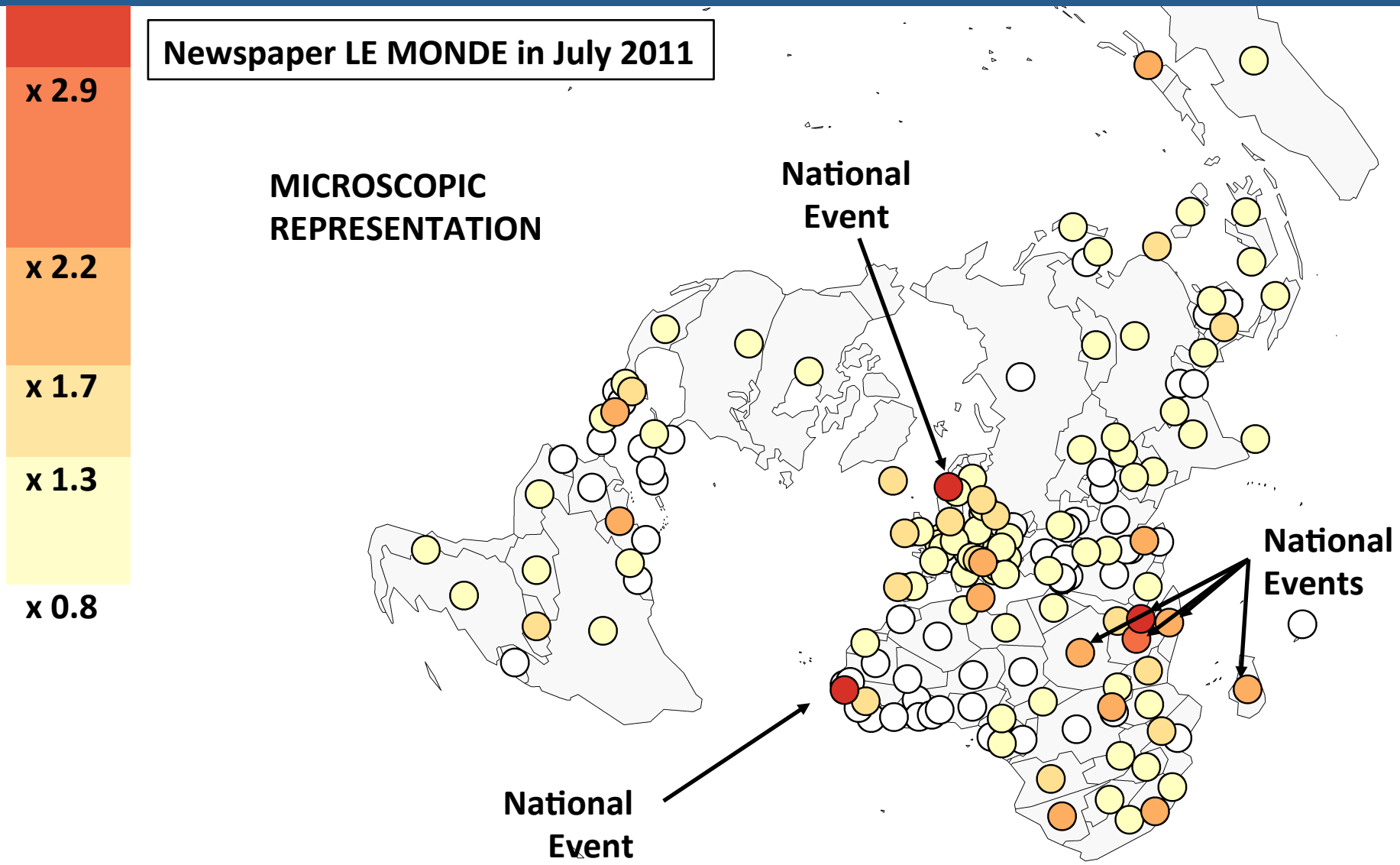
Time

Observed-to-expected ratio of citation number

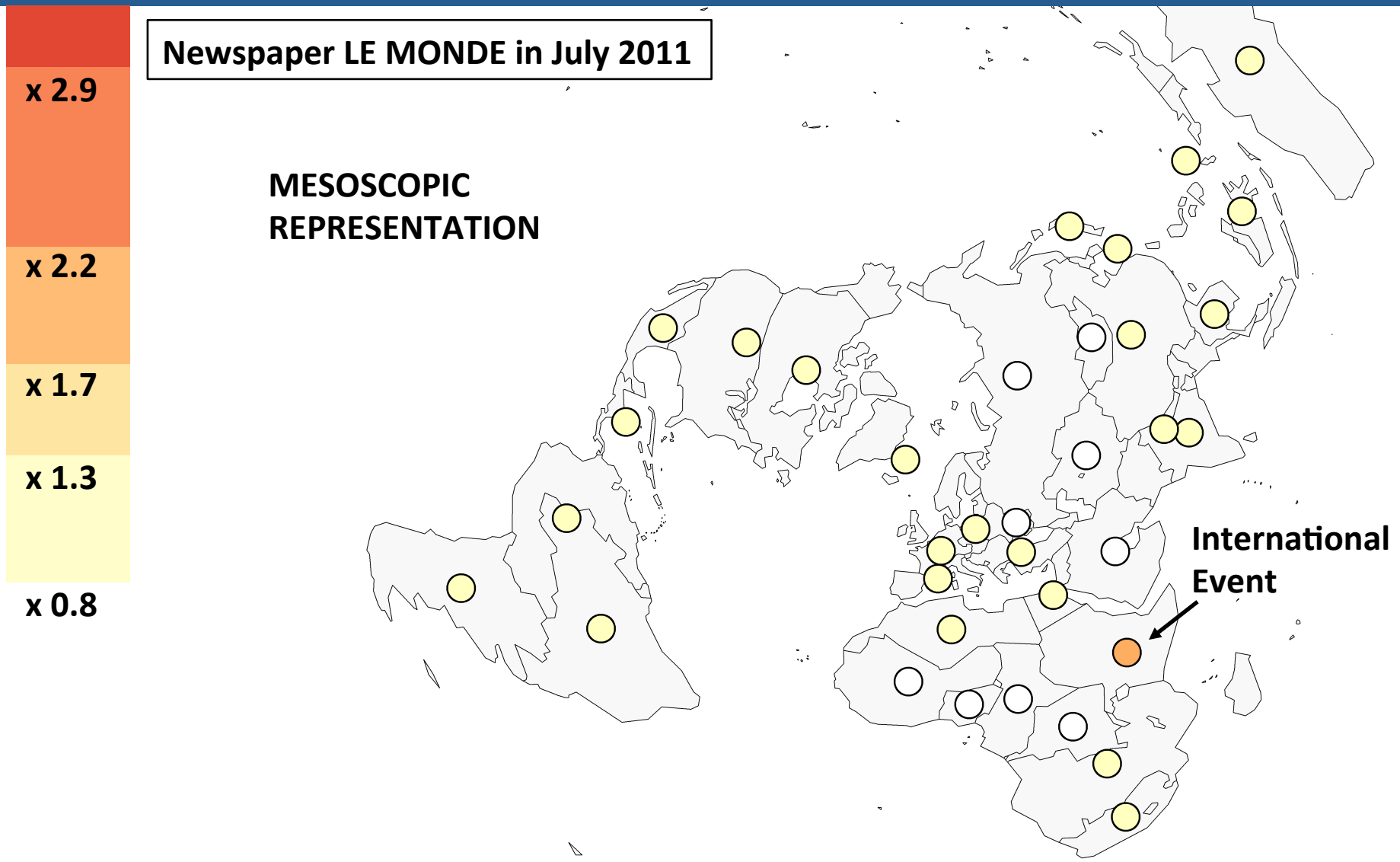
Spatial variation of the ratio at date t



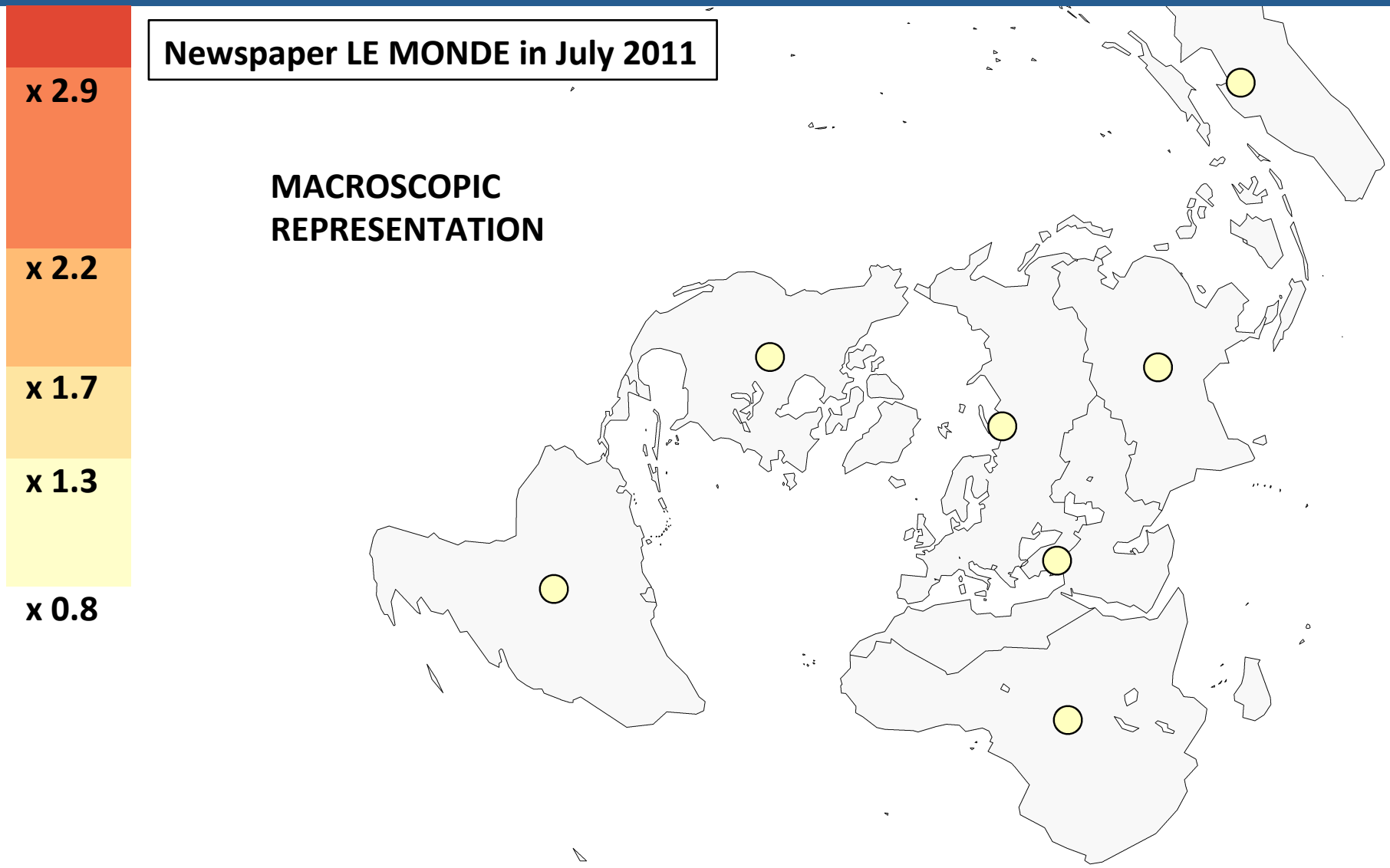
Knowledge Representation (micro)



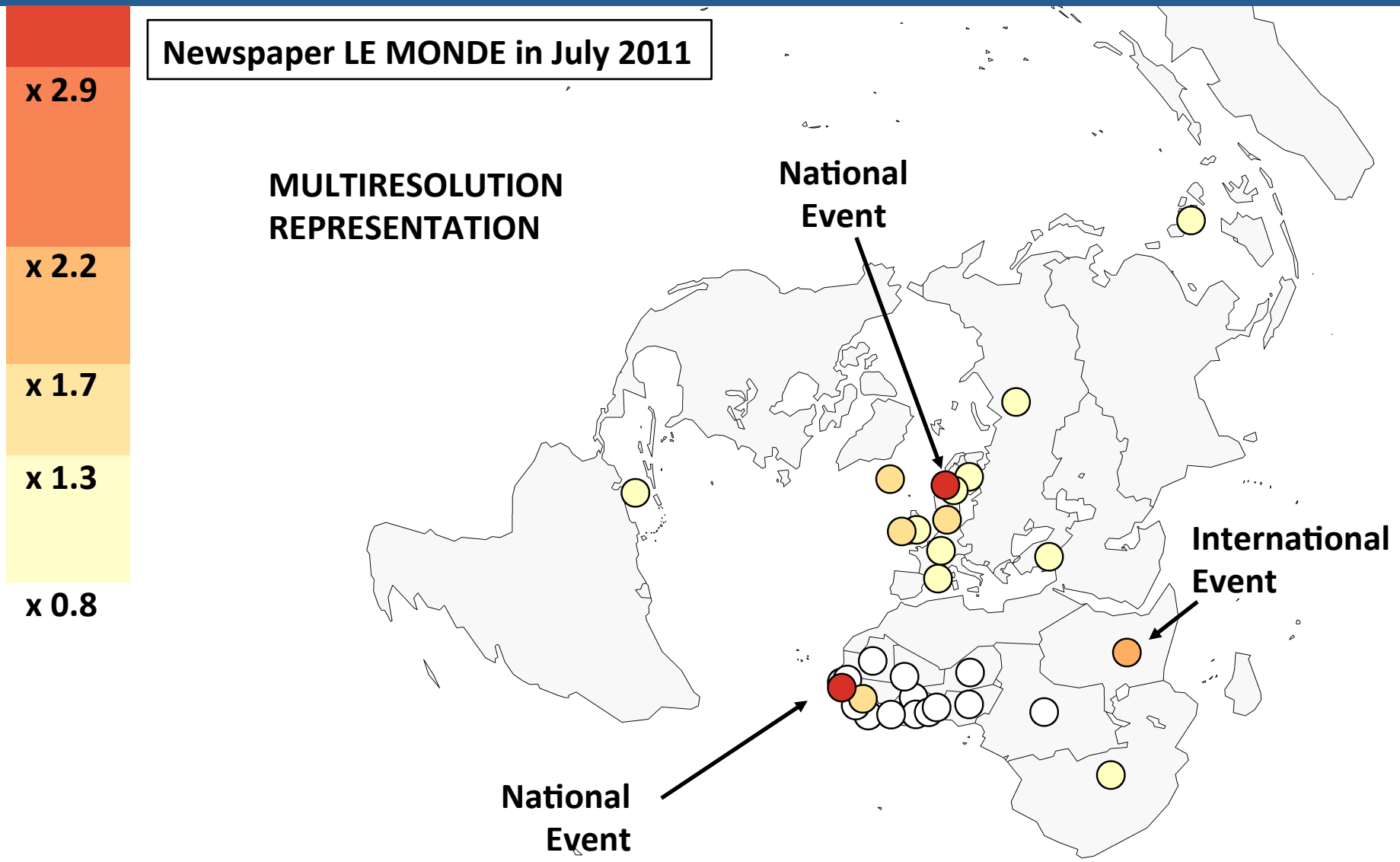
Knowledge Representation (meso)



Knowledge Representation (macro)



Macroscopic Representation



Lamarche-Perrin Approach

To characterize the aggregation process

→ The algebra of possible partitions

To preserve the system's semantics

→ A constrained partitioning method

To aggregate according to several dimension

→ Some constraints expressing the system's topology

To evaluate and compare the representations

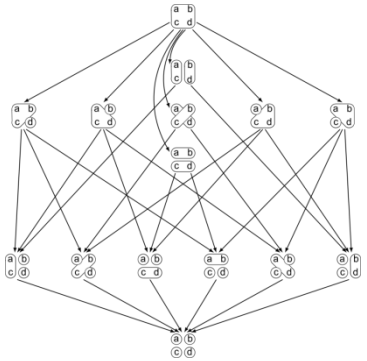
→ Some measures of complexity and information

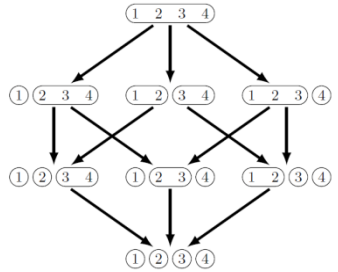
To offer several granularity levels

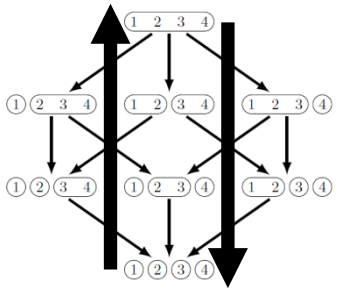
→ The optimization of a compromise

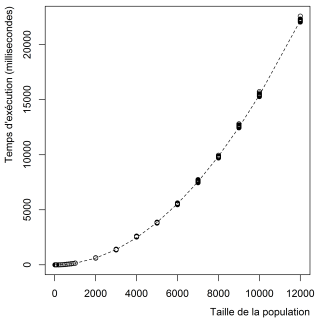
To compute the best representations

→ A generic algorithm of constrained optimization





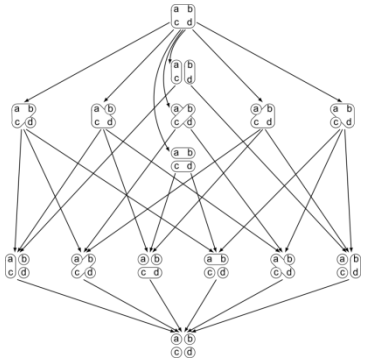




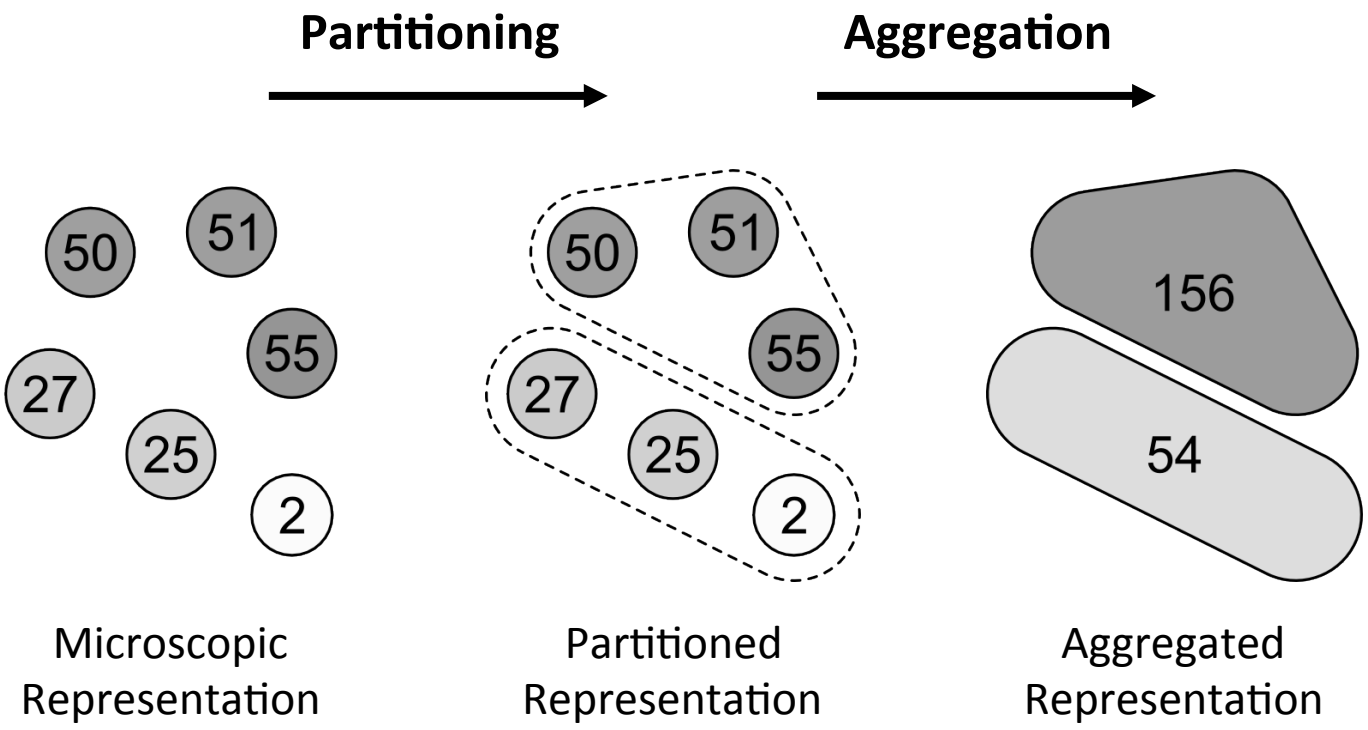
Lamarche-Perrin Approach

To characterize the aggregation process

→ The algebra of possible partitions



The Aggregation Process



Knowledge Aggregation (spatial)

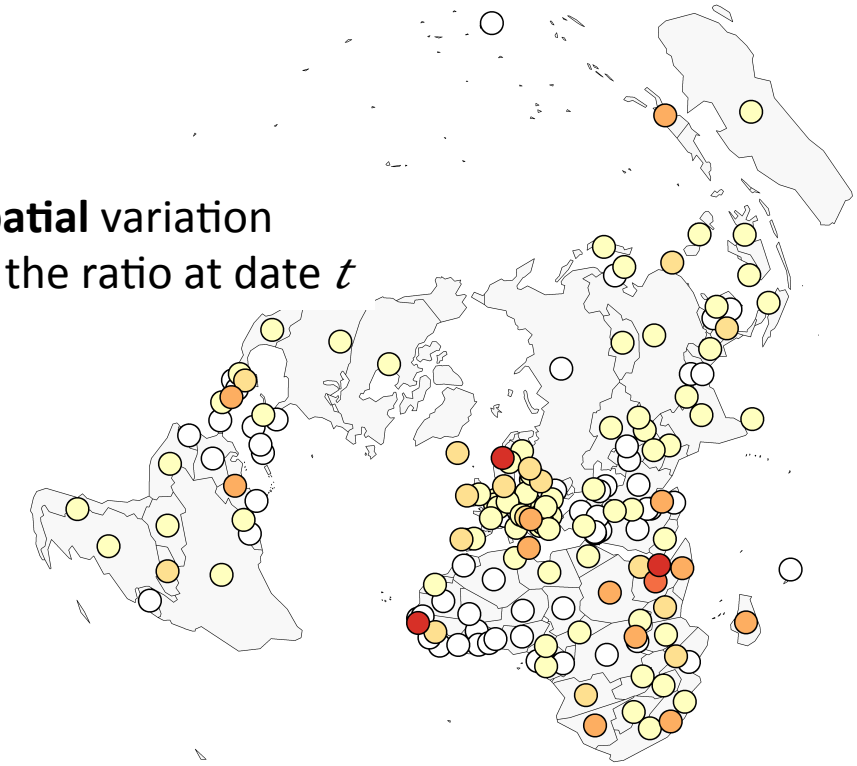
Space

	USA	Libya	Syria	France	Israel	…	Total
2 May	25	12	11	10	4	…	**142**
9 May	14	6	12	12	5	…	**108**
16 May	20	11	12	6	9	…	**142**
23 May	15	9	6	13	5	…	**120**
30 May	10	16	17	9	4	…	**137**
6 June	14	16	16	9	4	…	**114**
13 June	15	14	17	9	6	…	**119**
20 June	17	13	12	12	7	…	**123**
t 27 June	7	6	7	20	2	…	**103**
4 July	12	13	8	10	6	…	**129**
11 July	21	10	10	14	3	…	**107**
18 July	7	3	8	4	5	…	**61**
25 July	16	7	6	13	4	…	**128**
1 Aug.	21	1	9	7	4	…	**88**
…	…	…	…	…	…	…	…
Total	**423**	**308**	**260**	**248**	**153**	…	**3520**

Time

Observed-to-expected ratio of citation number

Spatial variation of the ratio at date t

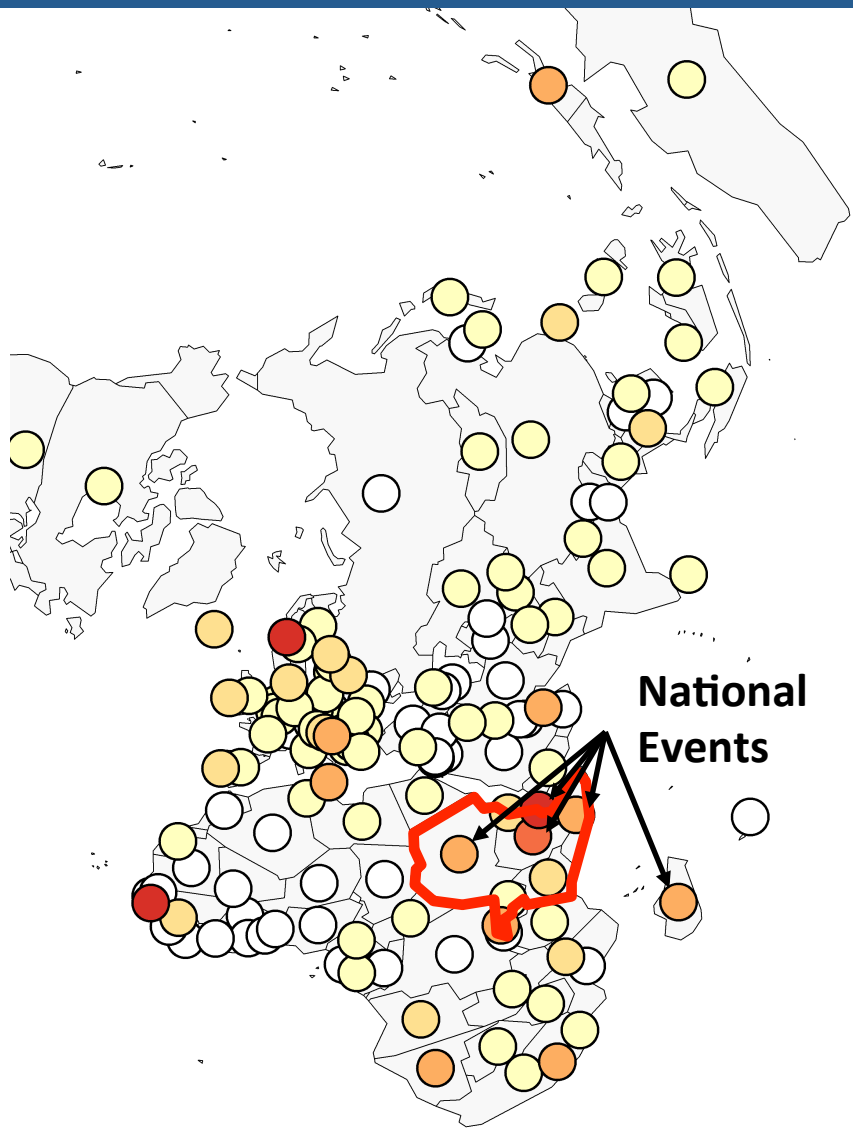


Knowledge Aggregation (spatial)

$\pi\downarrow 1 \pi\downarrow 2 \pi\downarrow 3$ **Space**

	USA	Libya	Syria	France	Israel	…	Total
2 May	25	12	11	10	4	…	**142**
9 May	14	6	12	12	5	…	**108**
16 May	20	11	12	6	9	…	**142**
23 May	15	9	6	13	5	…	**120**
30 May	10	16	17	9	4	…	**137**
6 June	14	16	16	9	4	…	**114**
13 June	15	14	17	9	6	…	**119**
20 June	17	13	12	12	7	…	**123**
27 June	7	6	7	20	2	…	**103**
4 July	12	13	8	10	6	…	**129**
11 July	21	10	10	14	3	…	**107**
18 July	7	3	8	4	5	…	**61**
25 July	16	7	6	13	4	…	**128**
1 Aug.	21	1	9	7	4	…	**88**
…	…	…	…	…	…	…	…
Total	**423**	**308**	**260**	**248**	**153**	…	**3520**

t **Time**



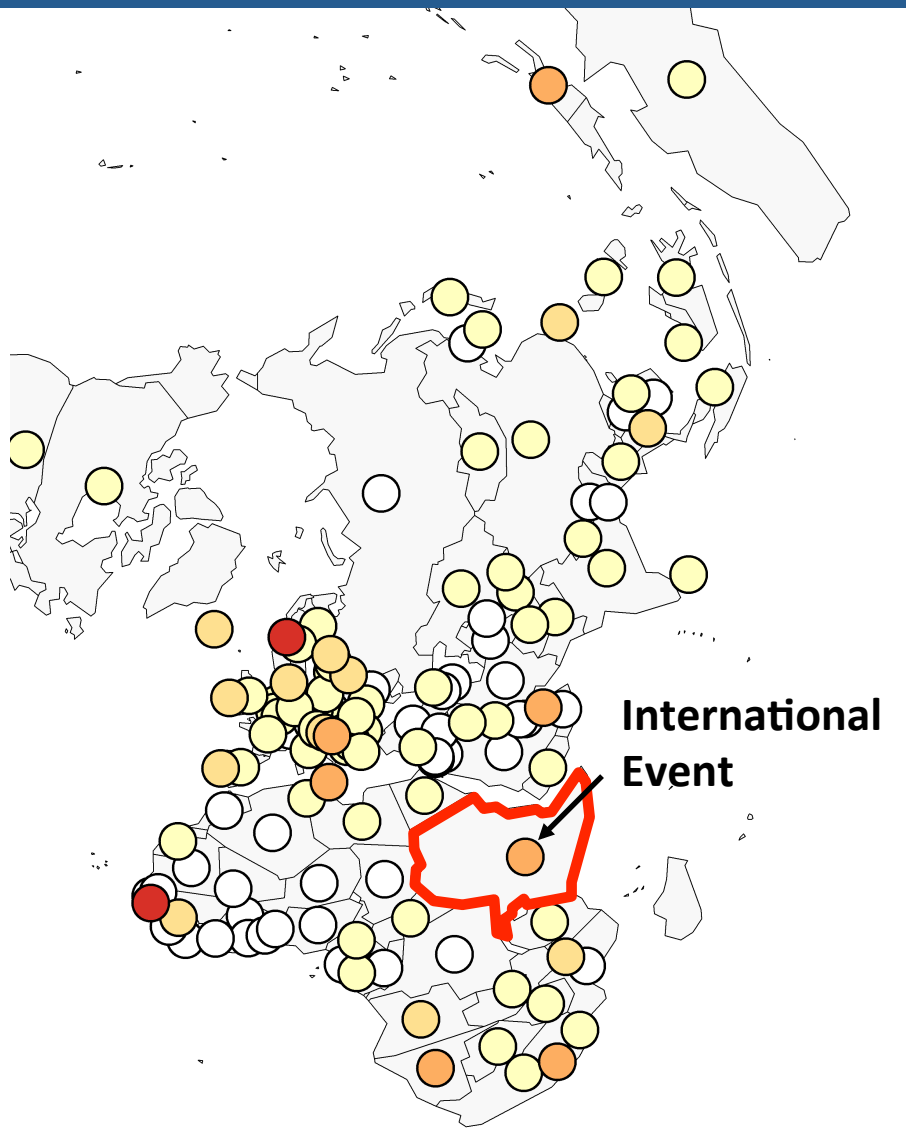
Knowledge Aggregation (spatial)

$\pi\downarrow 1 \pi\downarrow 2 \pi\downarrow 3$ **Space**

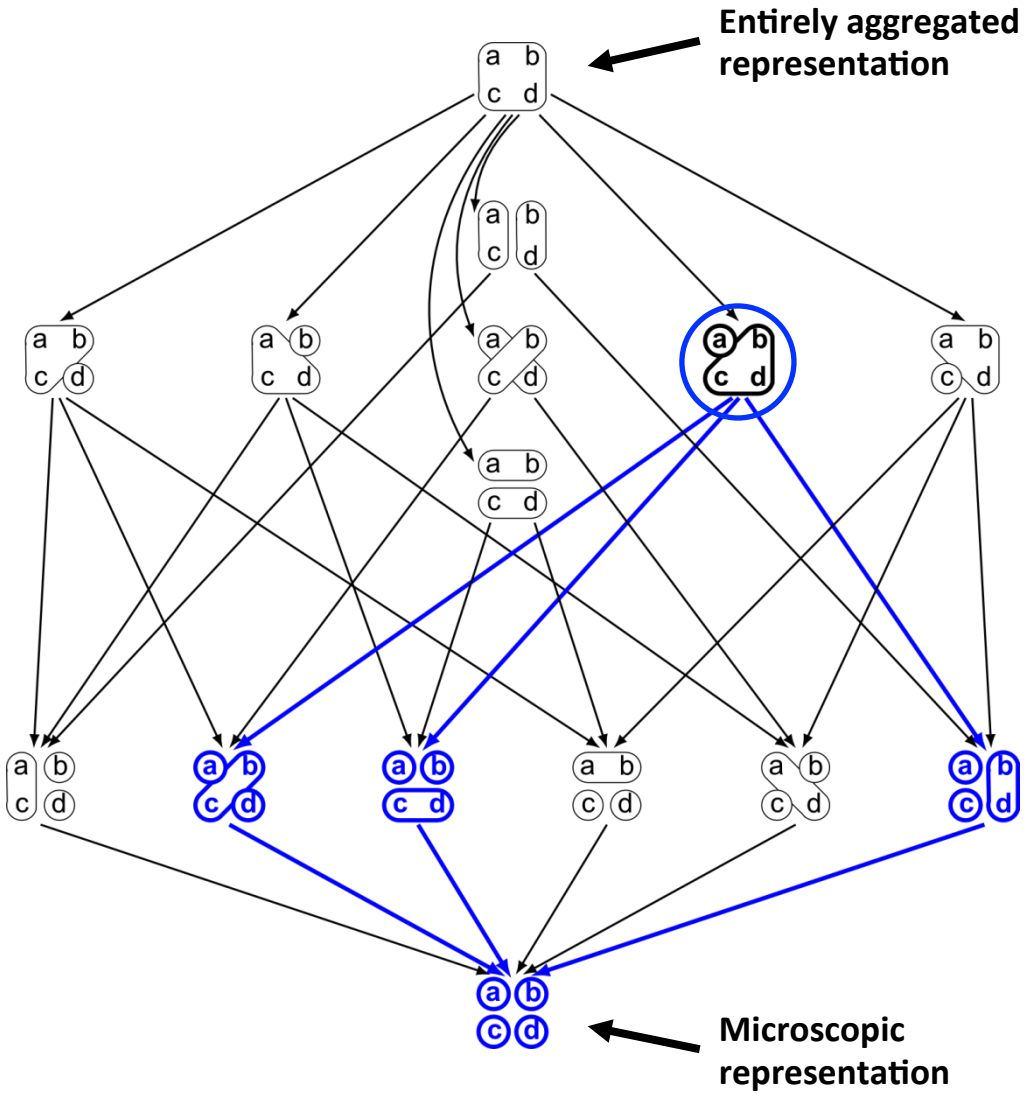
	USA	Aggregate	Israel	...	Total
2 May	25	13+11+10	4	...	142
9 May	14	6+12+12	5	...	108
16 May	20	11+12+6	9	...	142
23 May	15	9+6+13	5	...	120
30 May	10	16+17+9	4	...	137
6 June	14	16+16+9	4	...	114
13 June	15	14+17+9	6	...	119
20 June	17	13+12+12	7	...	123
27 June	7	6+7+20	2	...	103
4 July	12	13+8+10	6	...	129
11 July	21	10+10+14	3	...	107
18 July	7	3+8+4	5	...	61
25 July	16	7+6+13	4	...	128
1 Aug.	21	1+9+7	4	...	88
...	...	...	...	...	...
Total	423	308+260+248	153	...	3520

t

Time



Set of Possible Representations

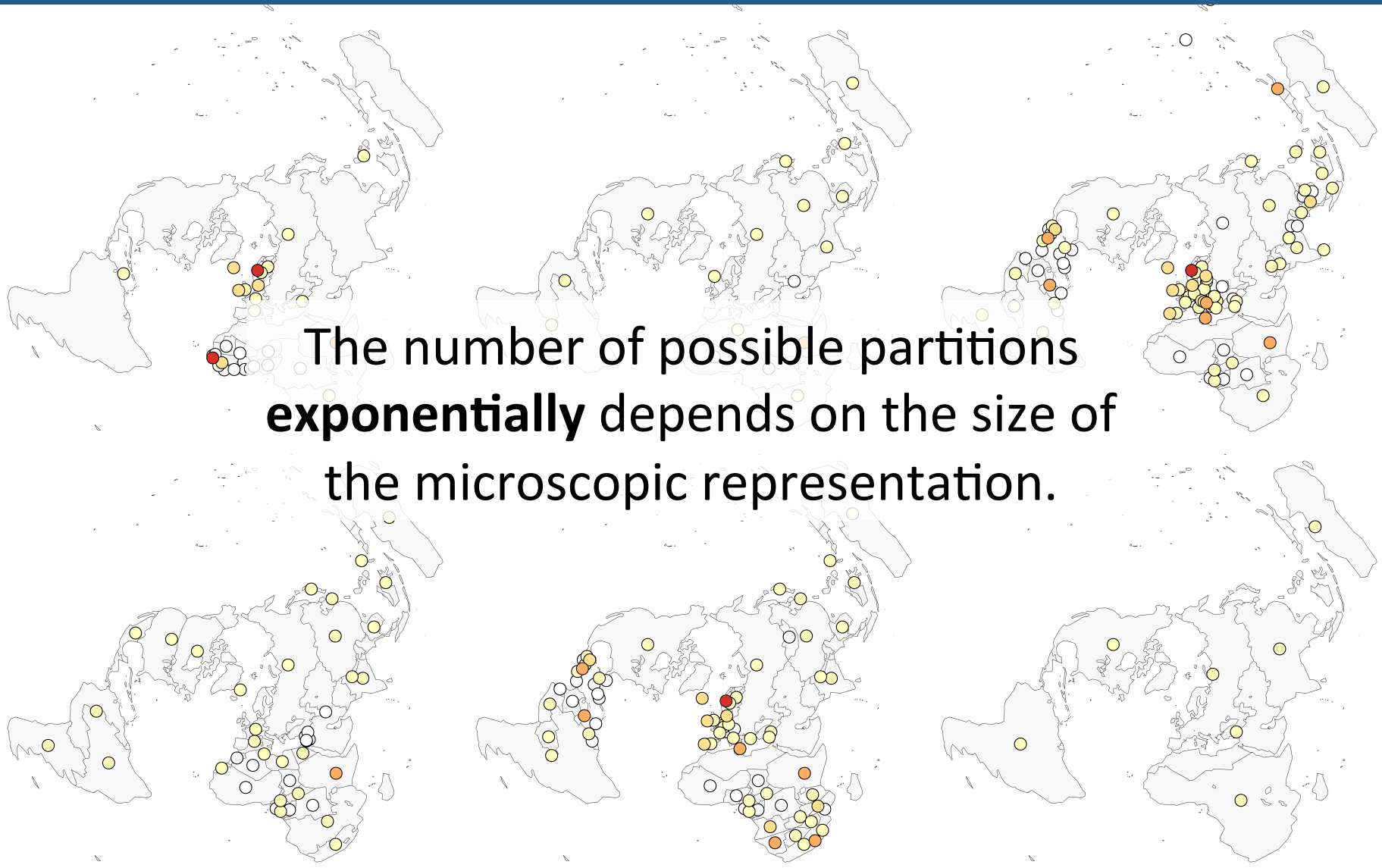


Algebraic Structure

A partial order on the set of possible partitions

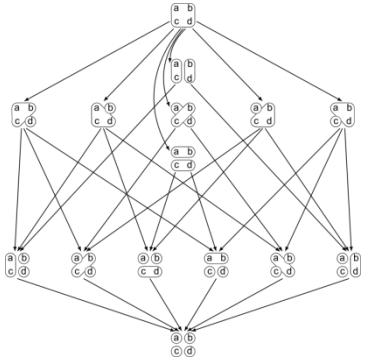
→ **the refinement relation**

Set of Possible Representations



The number of possible partitions **exponentially** depends on the size of the microscopic representation.

Lamarche-Perrin Approach



To characterize the aggregation process

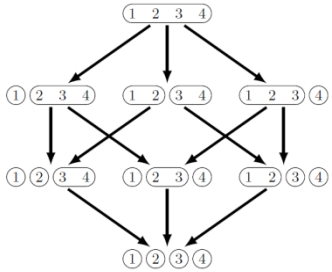
→ The algebra of possible partitions

To preserve the system's semantics

→ A constrained partitioning method

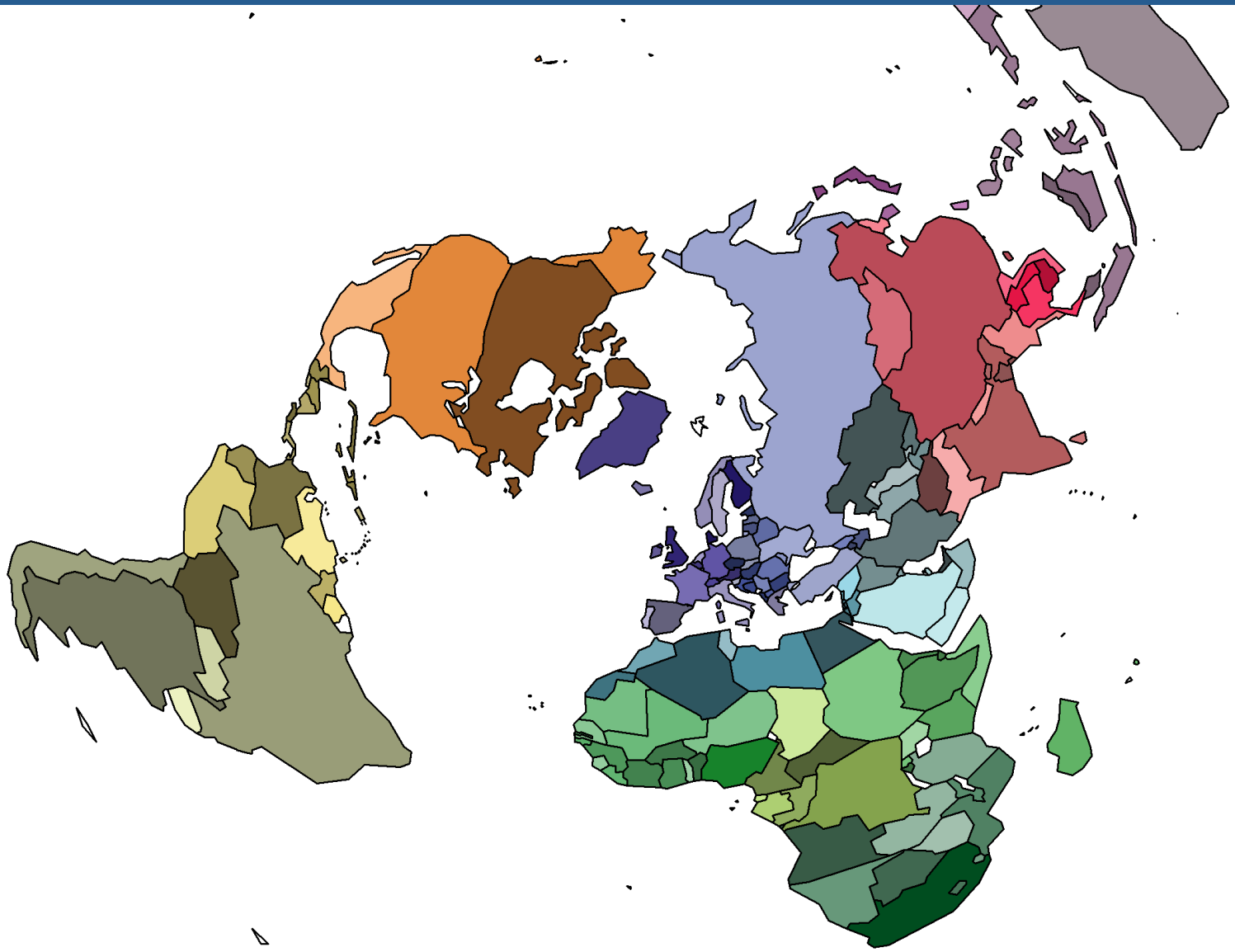
To aggregate according to several dimension

→ Some constraints expressing the system's topology



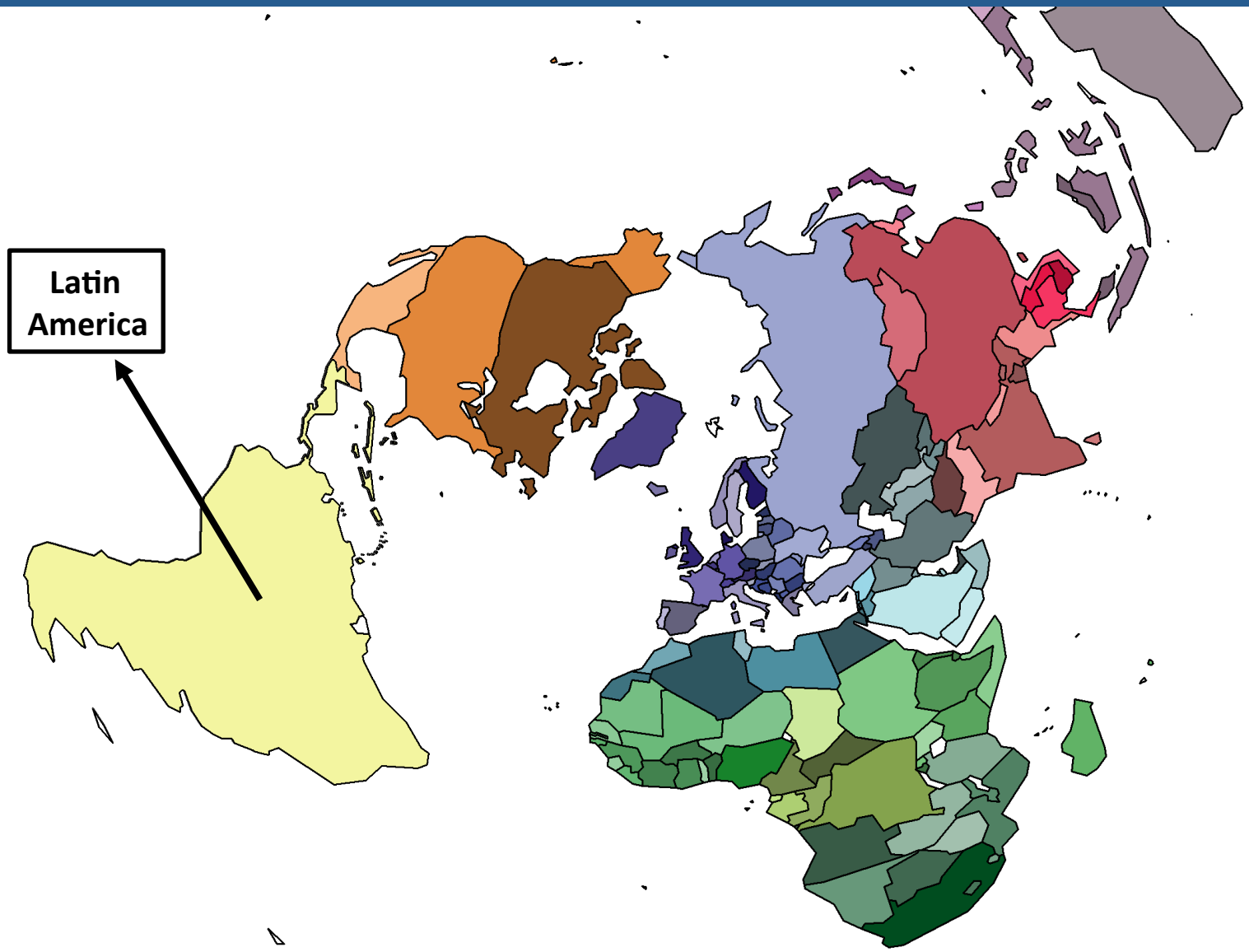
Meaningful Representations

Geographer

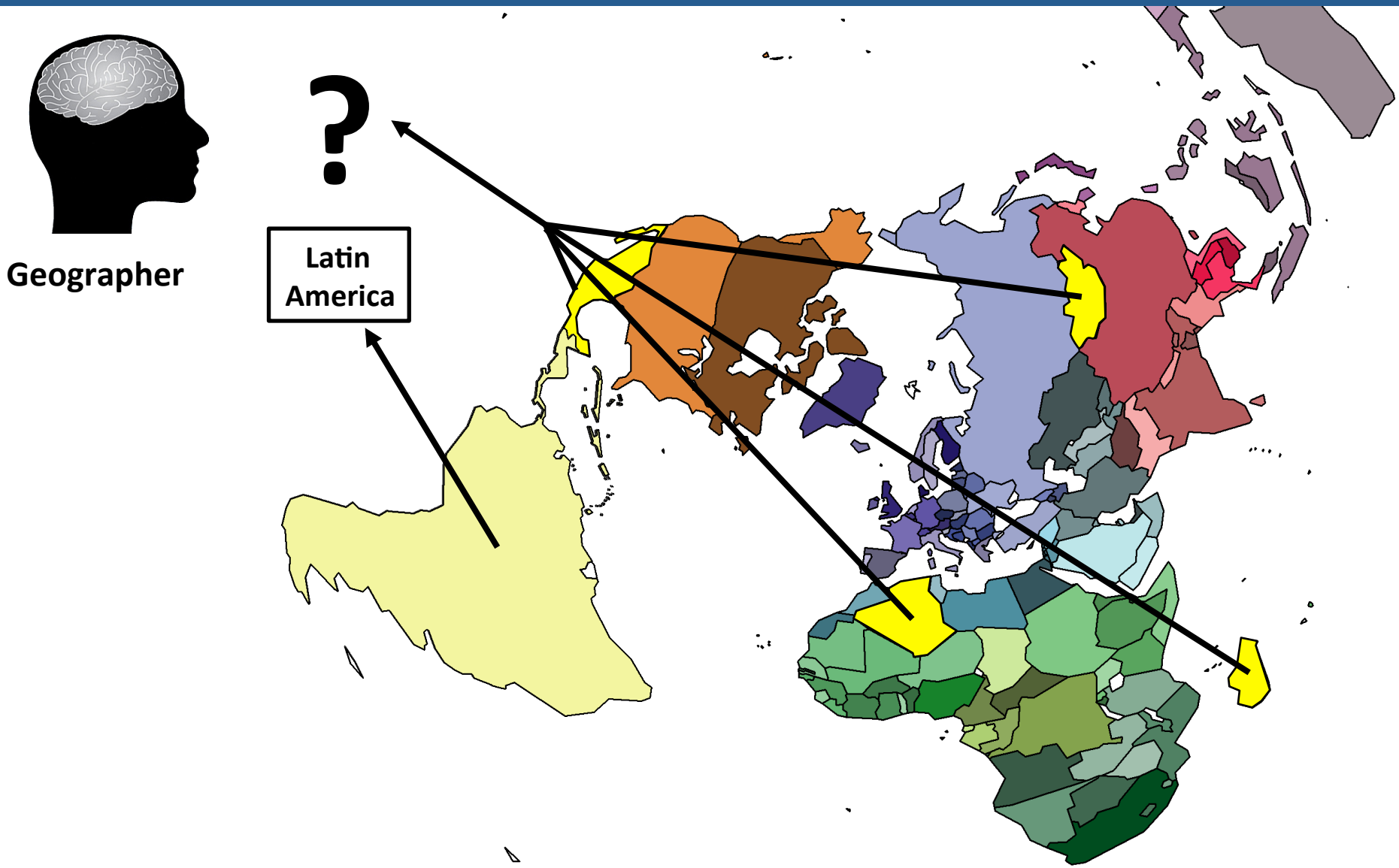


Meaningful Representations

Geographer



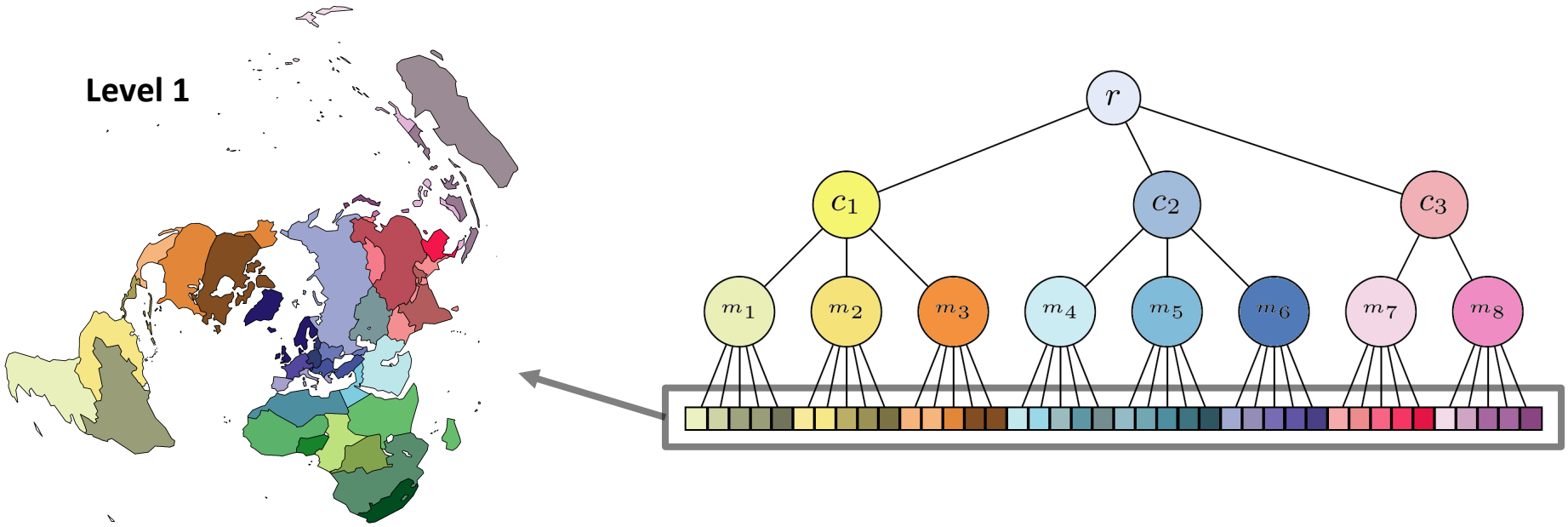
Meaningful Representations



The WUTS Hierarchy

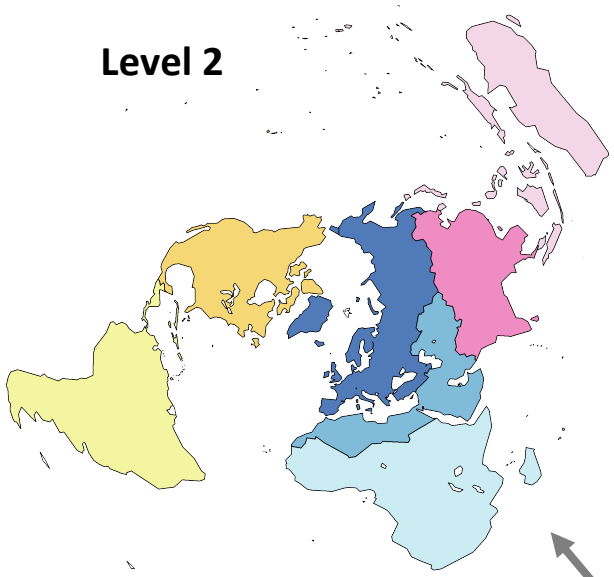
[Grasland and Didelon, 2007]

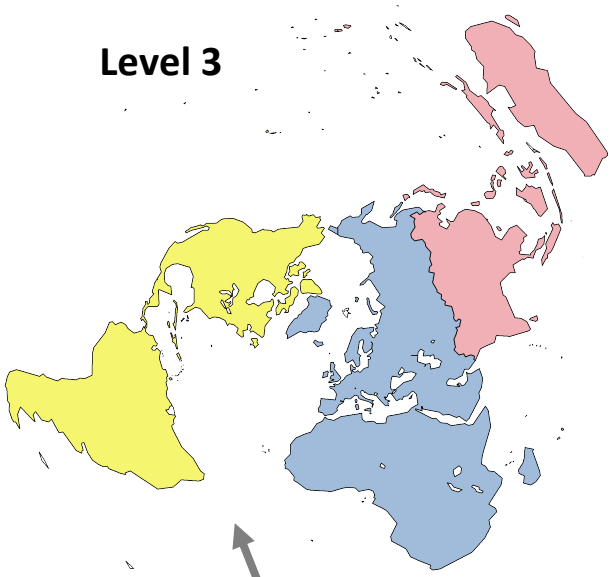
Admissible Parts:
set of countries that are politically, culturally and economically consistent

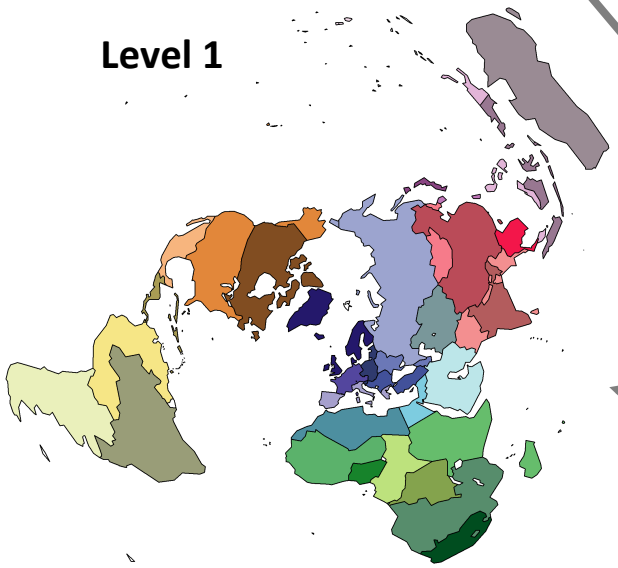


The WUTS Hierarchy

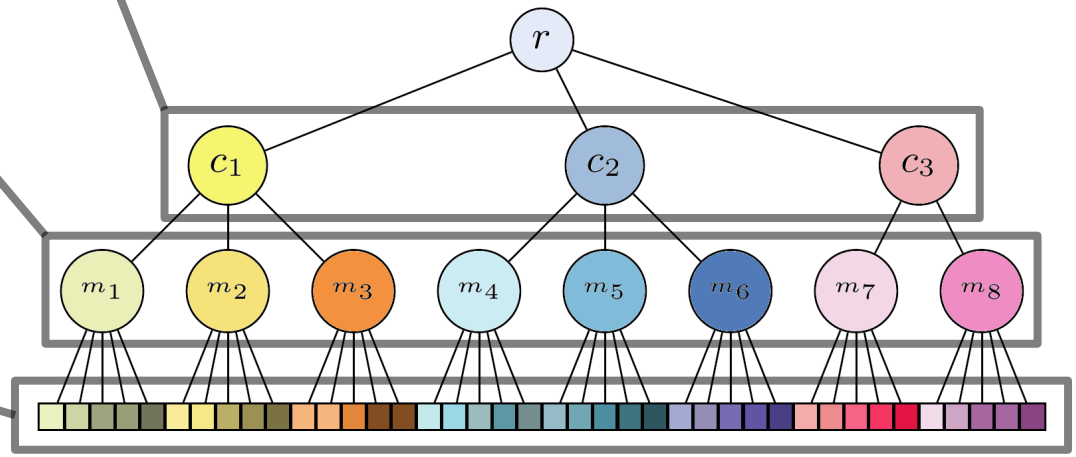
[Grasland and Didelon, 2007]





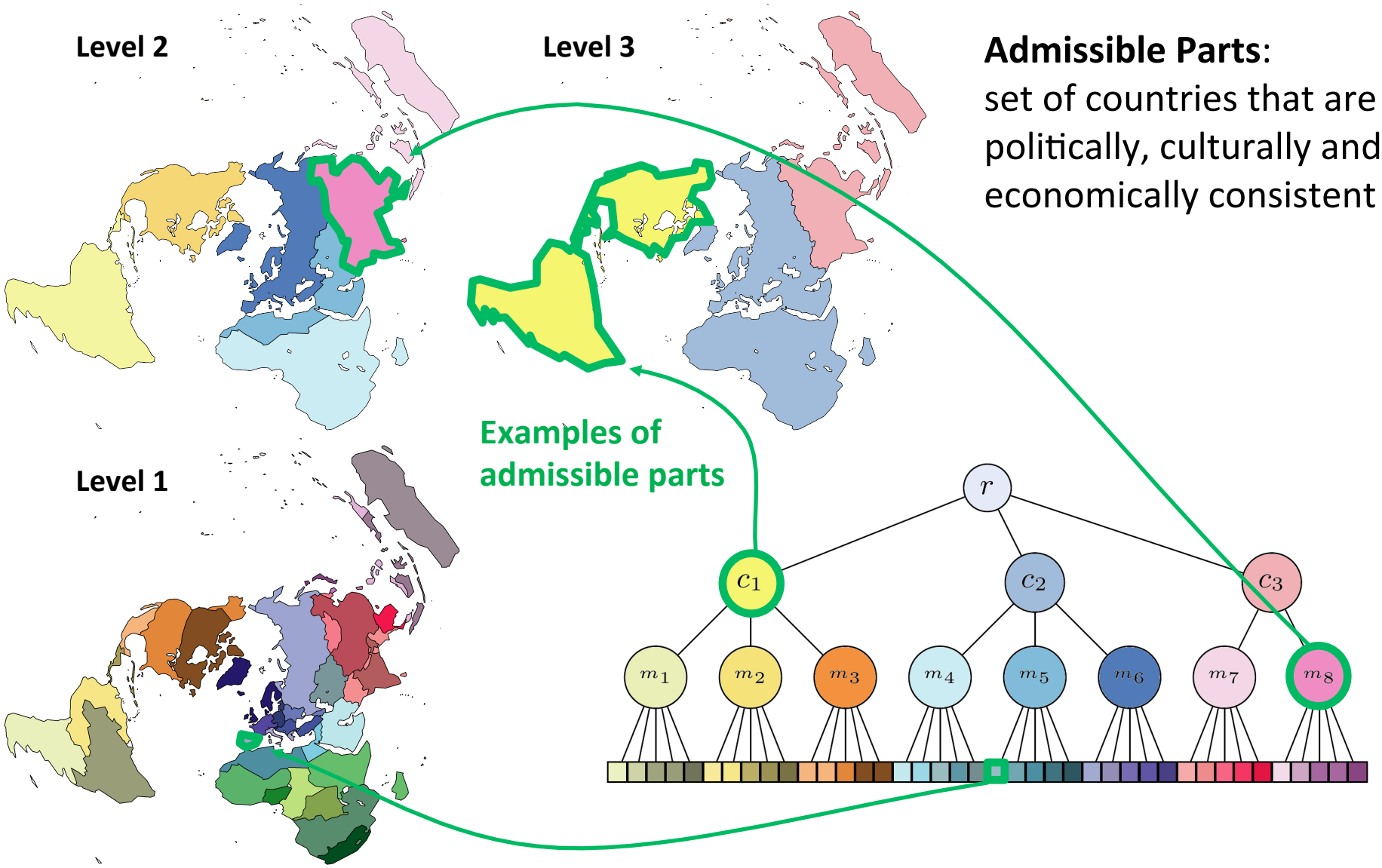


Admissible Parts:
set of countries that are politically, culturally and economically consistent



The WUTS Hierarchy

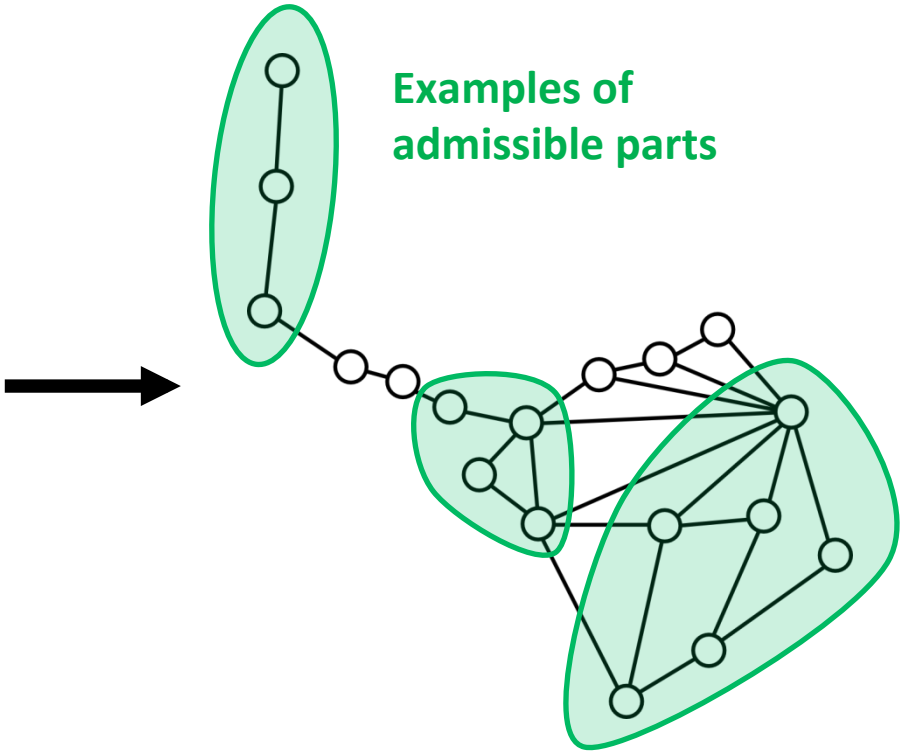
[Grasland and Didelon, 2007]



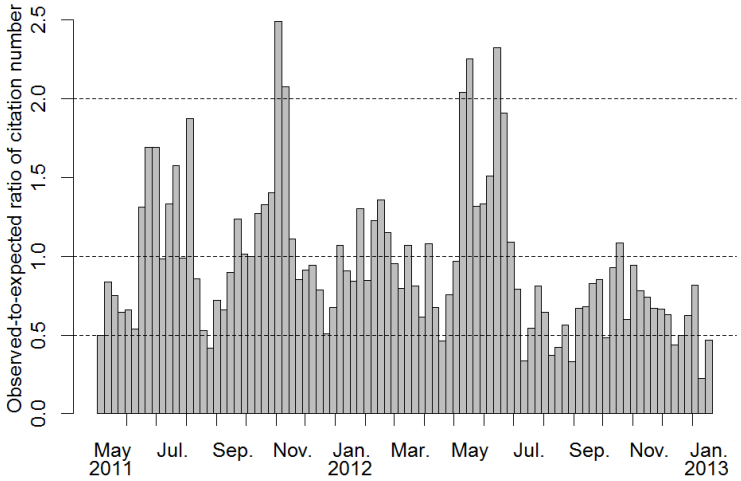
Admissible Parts:
set of countries that are politically, culturally and economically consistent

Preserving the Neighborhood Relation

Admissible Parts: set of connected countries regarding the adjacency graph

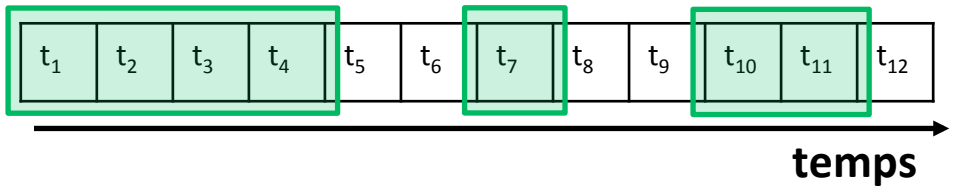


Preserving the Order of Time

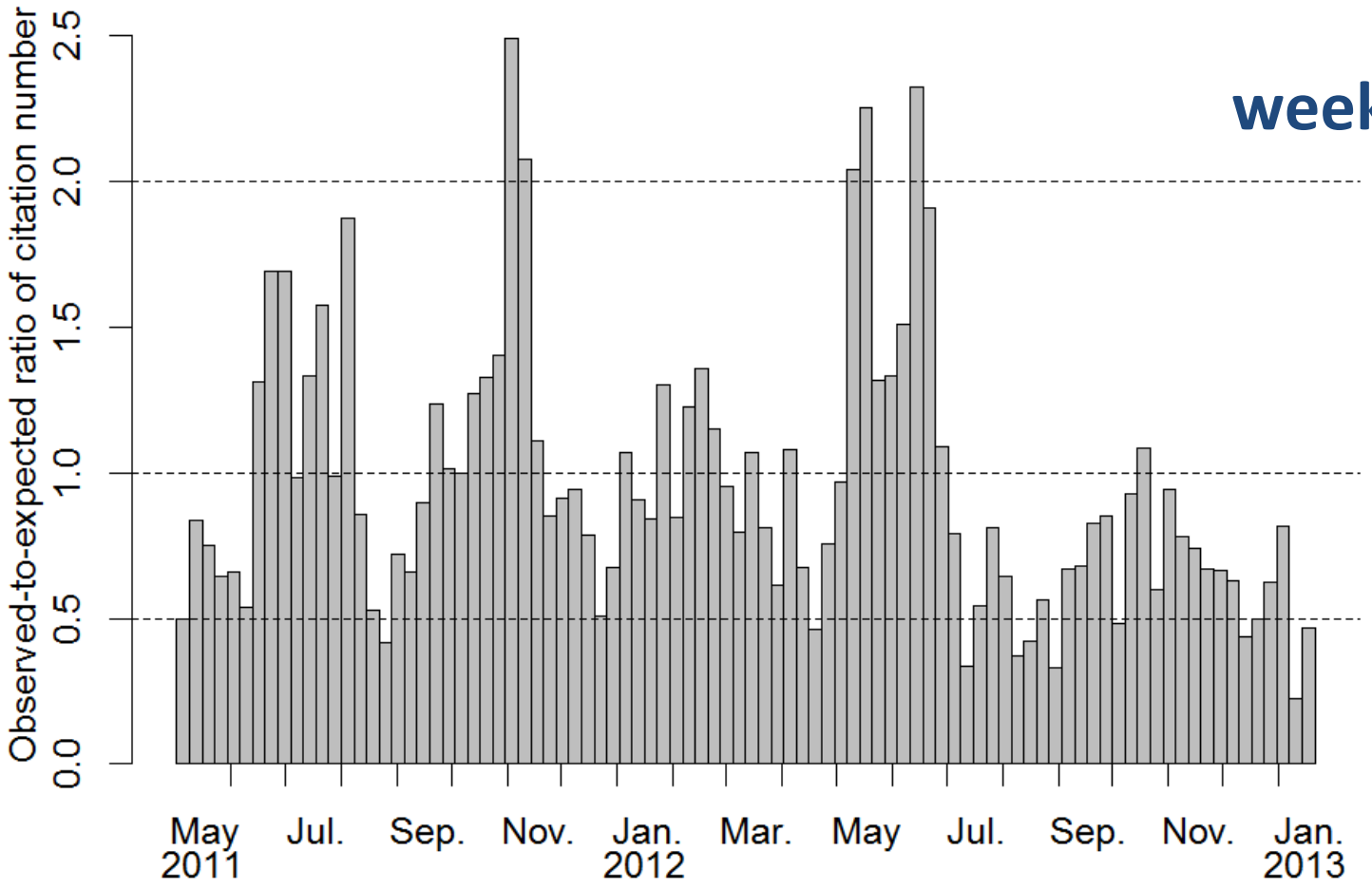


Admissible Parts:
time intervals

Examples of admissible parts

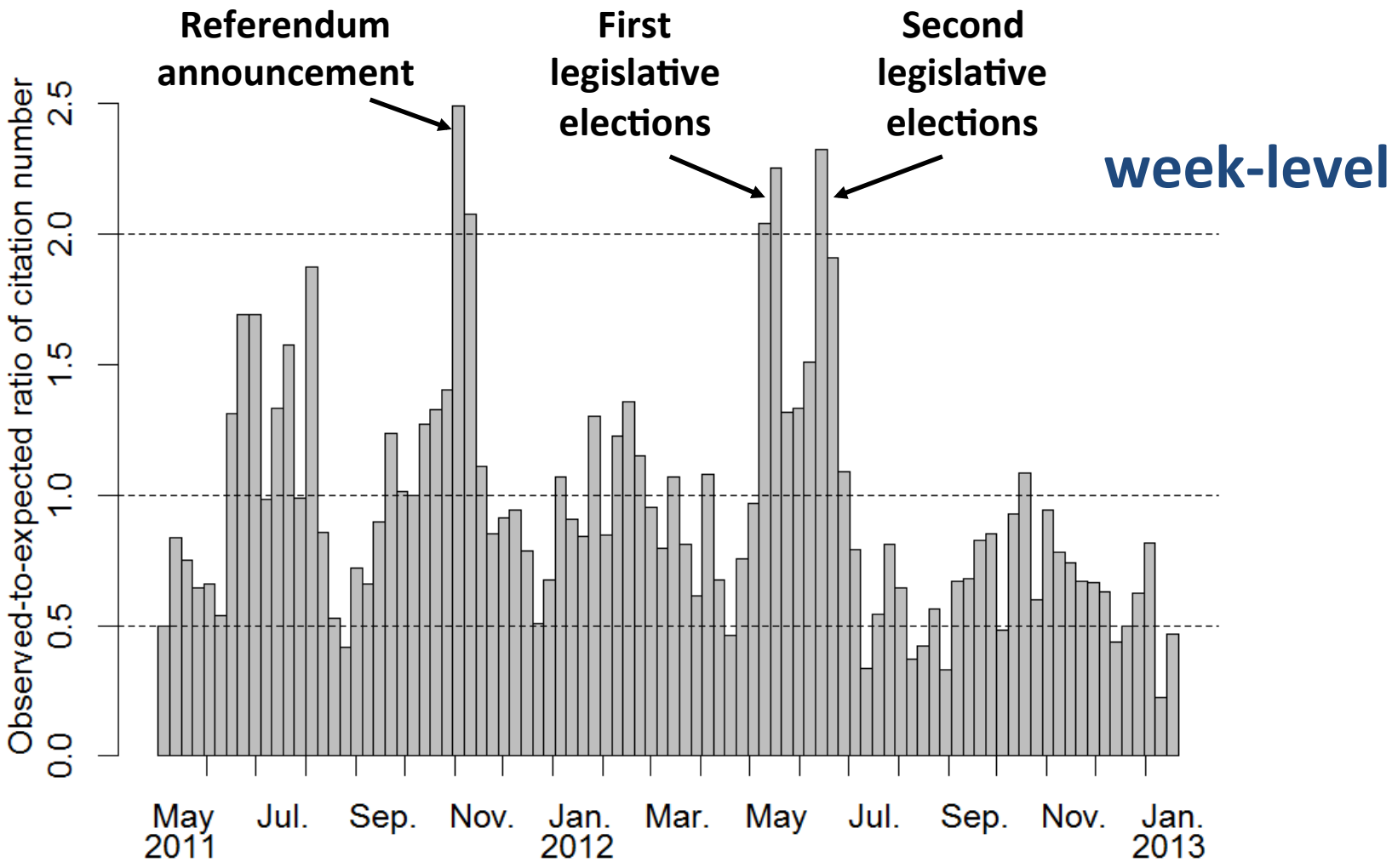


Citation of Greece by the Guardian

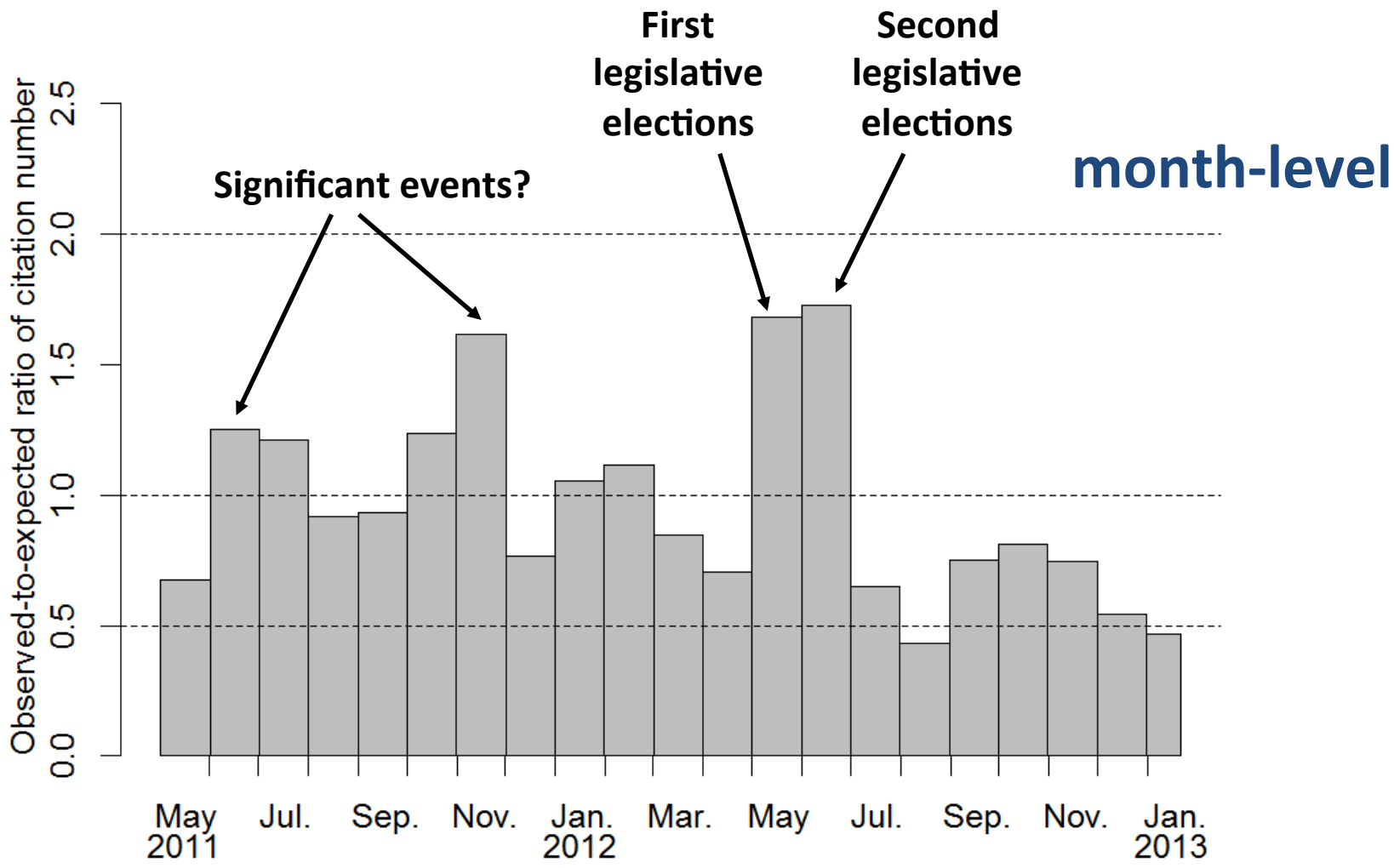


week-level

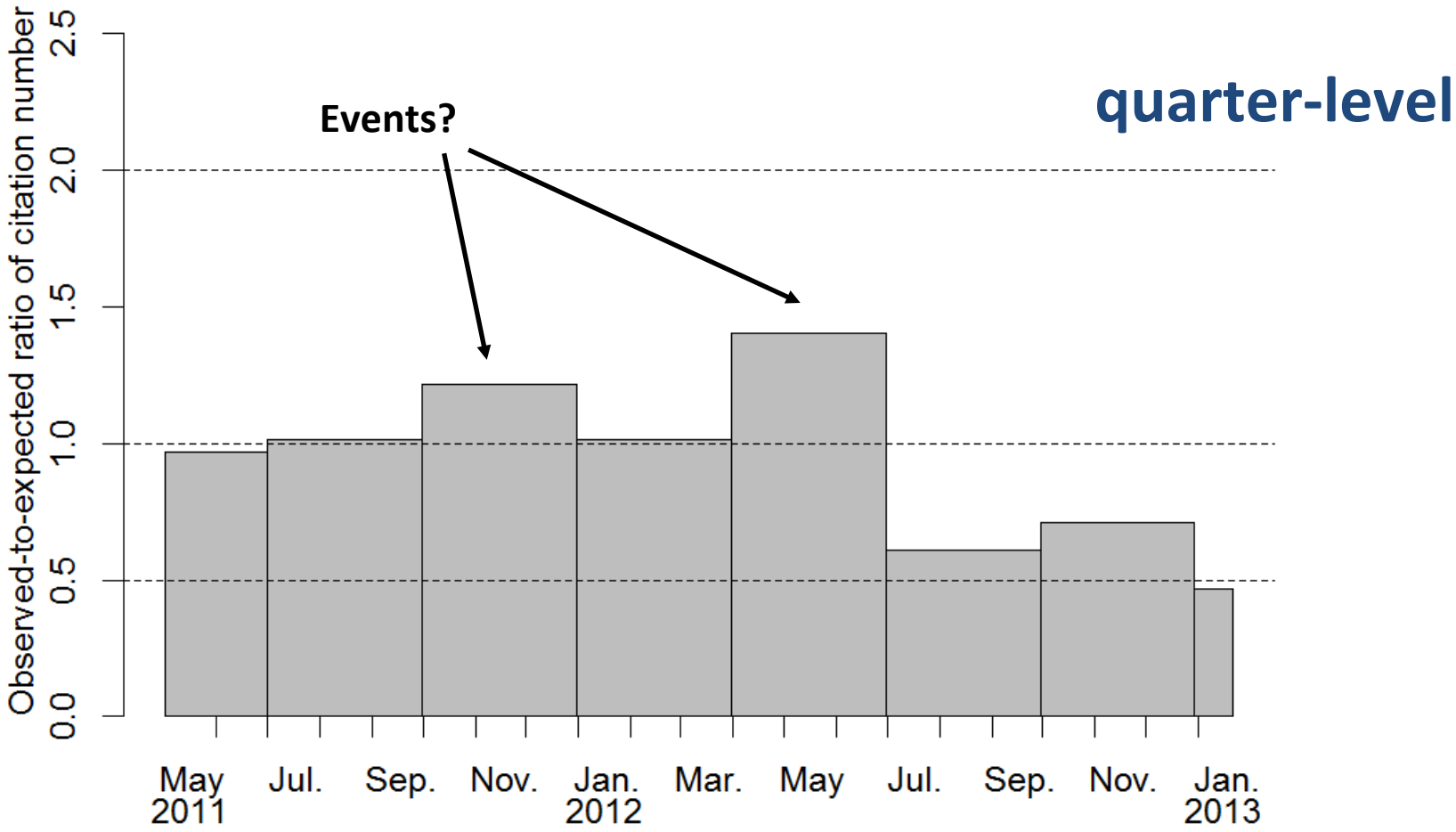
Citation of Greece by the Guardian



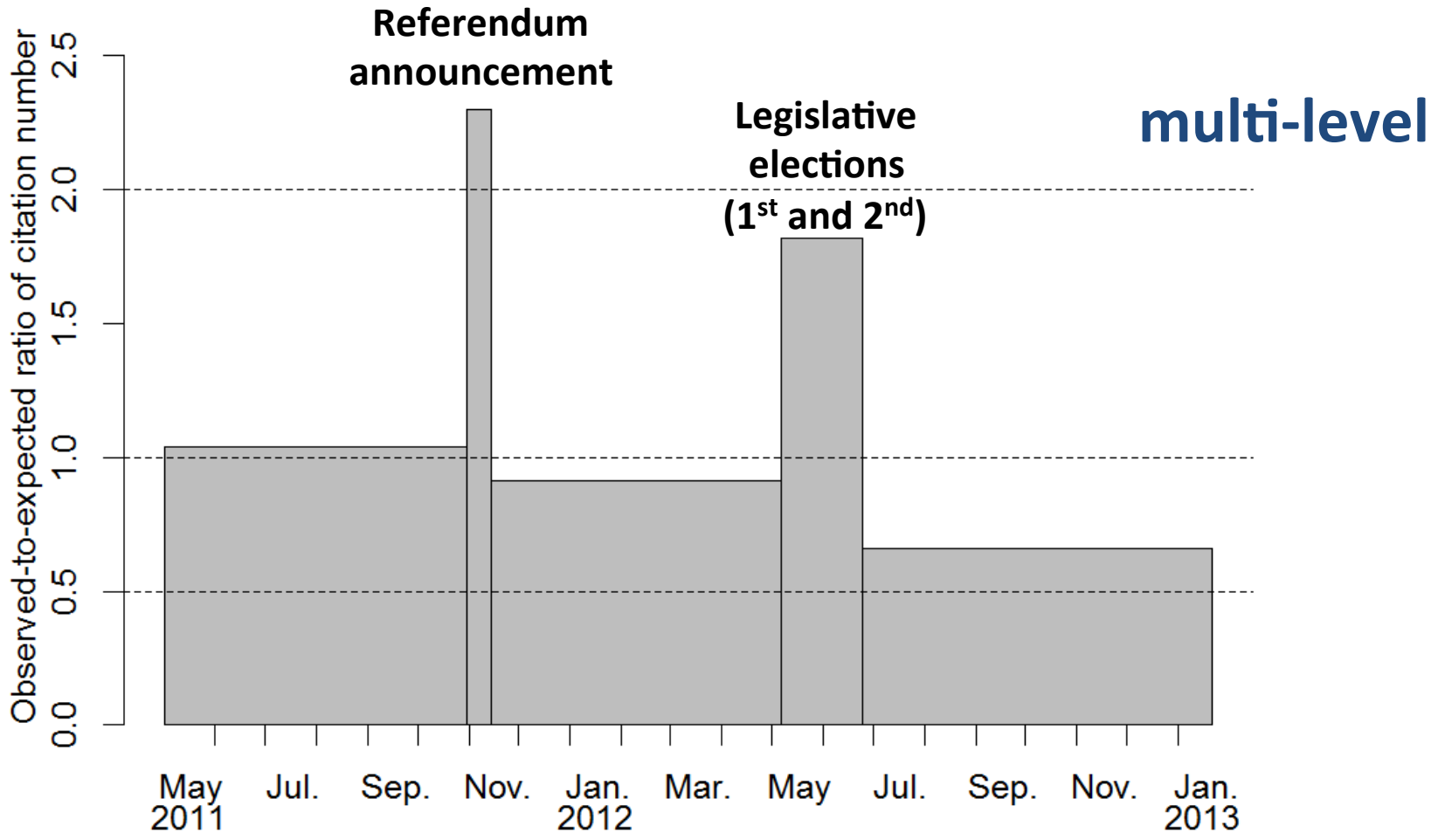
Citation of Greece by the Guardian



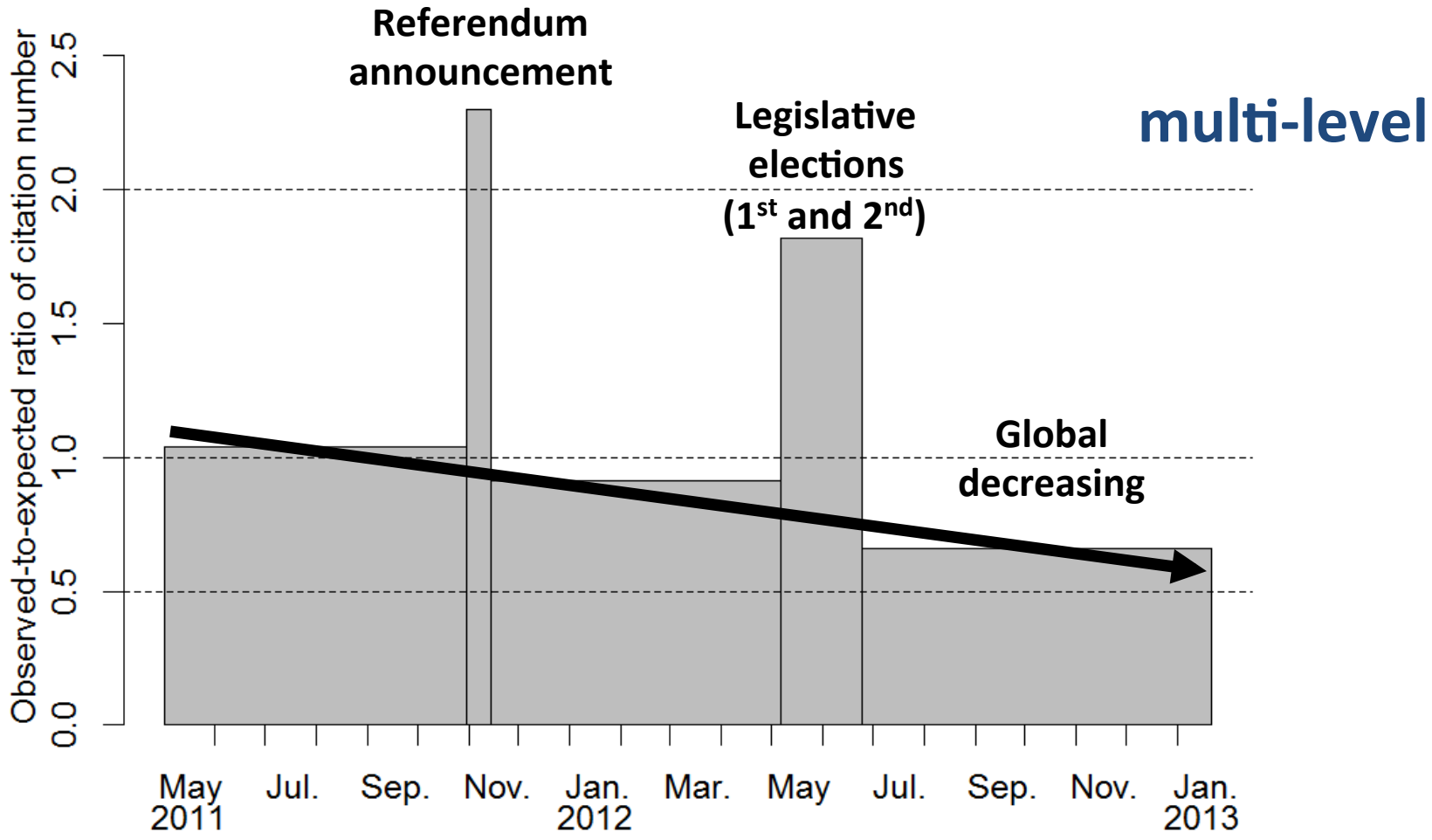
Citation of Greece by the Guardian



Citation of Greece by the Guardian

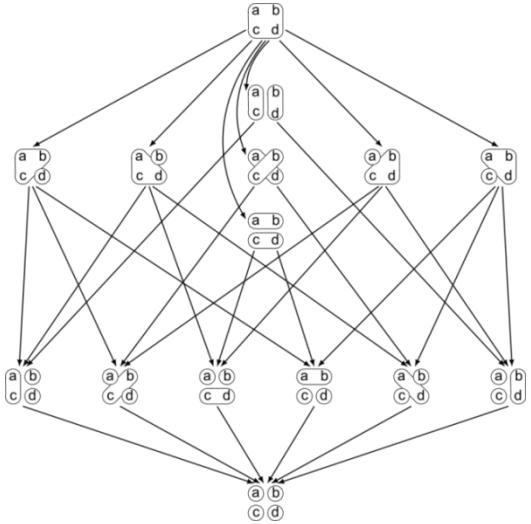


Citation of Greece by the Guardian

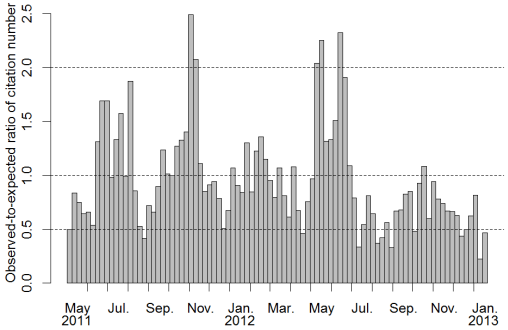


Problems and Objectives

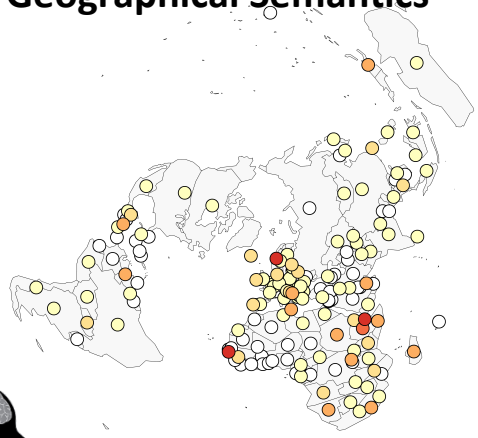
Set of <u>possible</u> partitions



Temporal Semantics



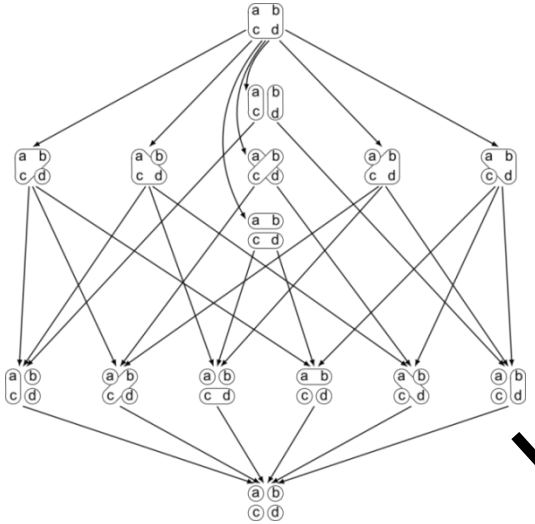
Geographical Semantics



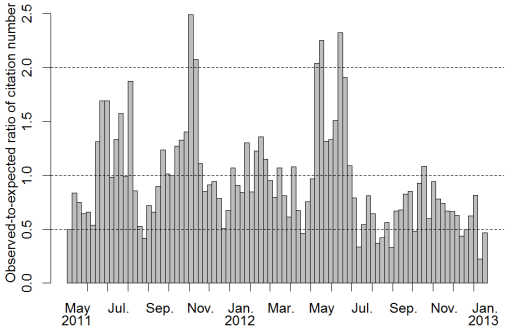
Geographer

Problems and Objectives

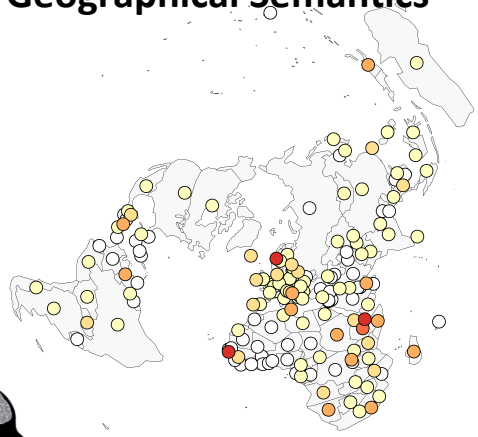
Set of possible partitions



Temporal Semantics



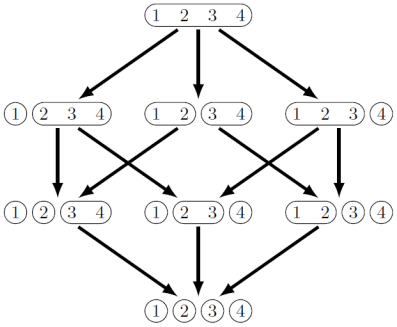
Geographical Semantics

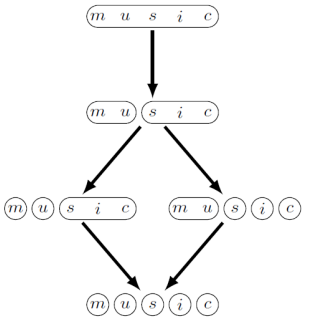


Constraints

Geographer

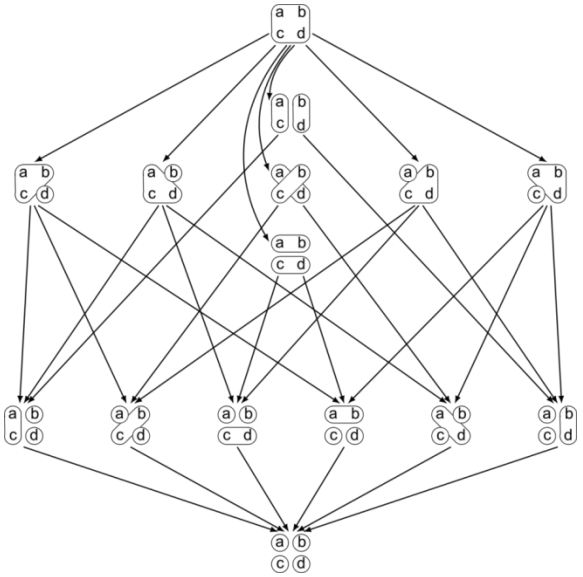
Set of admissible partitions



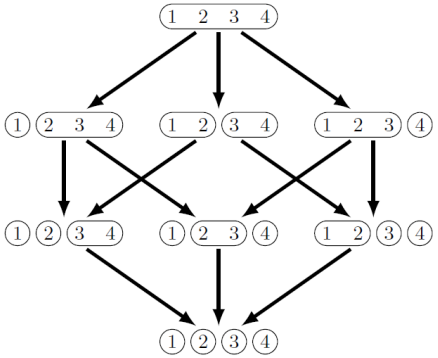


Complexity of Algebraic Structures

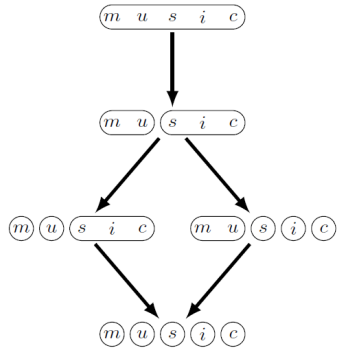
Non-constrained partitions



Admissible partitions according to a **total order**



Admissible partitions according to a **hierarchy**



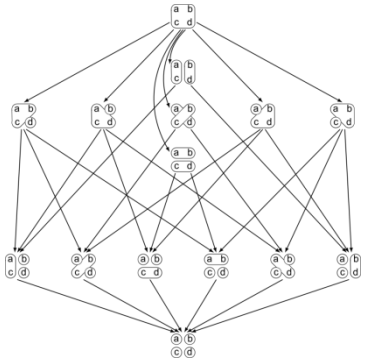
Less constrained

More constrained

More complex

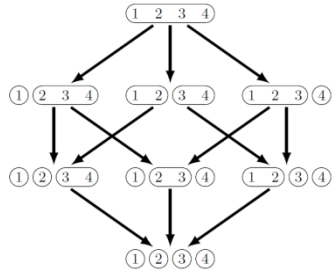
Less complex

Lamarche-Perrin Approach



To characterize the aggregation process

→ The algebra of possible partitions

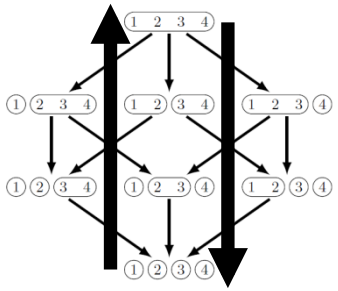


To preserve the system's semantics

→ A constrained partitioning method

To aggregate according to several dimension

→ Some constraints expressing the system's topology



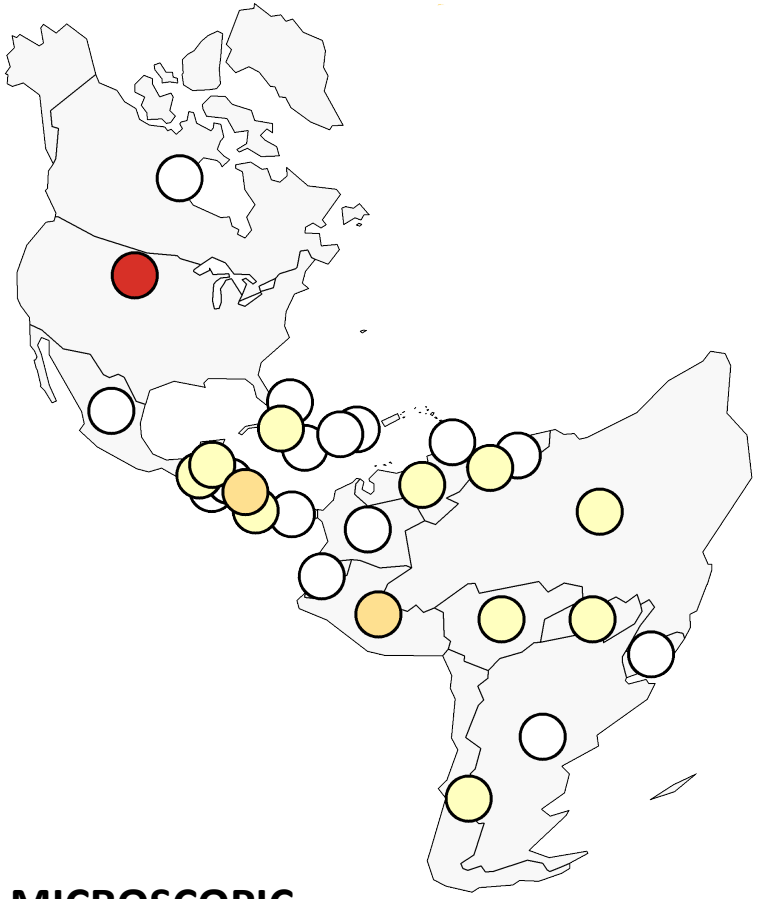
To evaluate and compare the representations

→ Some measures of complexity and information

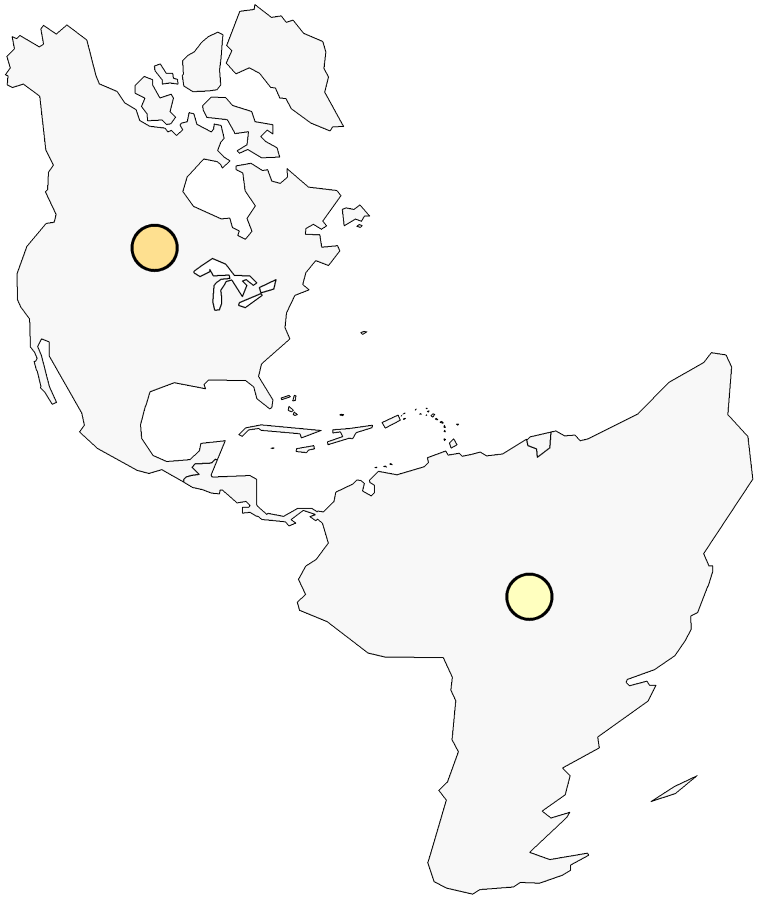
To offer several granularity levels

→ The optimization of a compromise

Complexity and Information

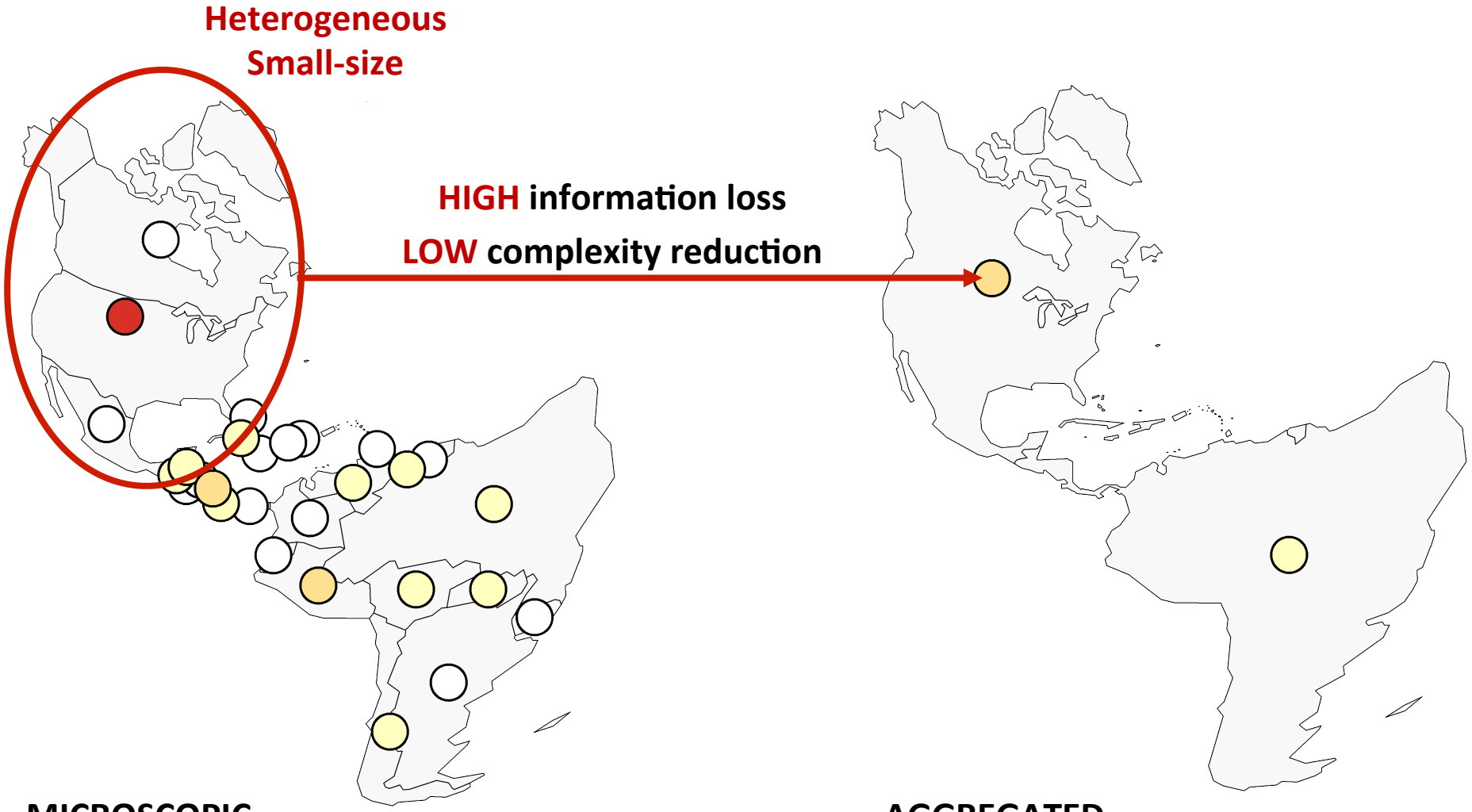


MICROSCOPIC REPRESENTATION

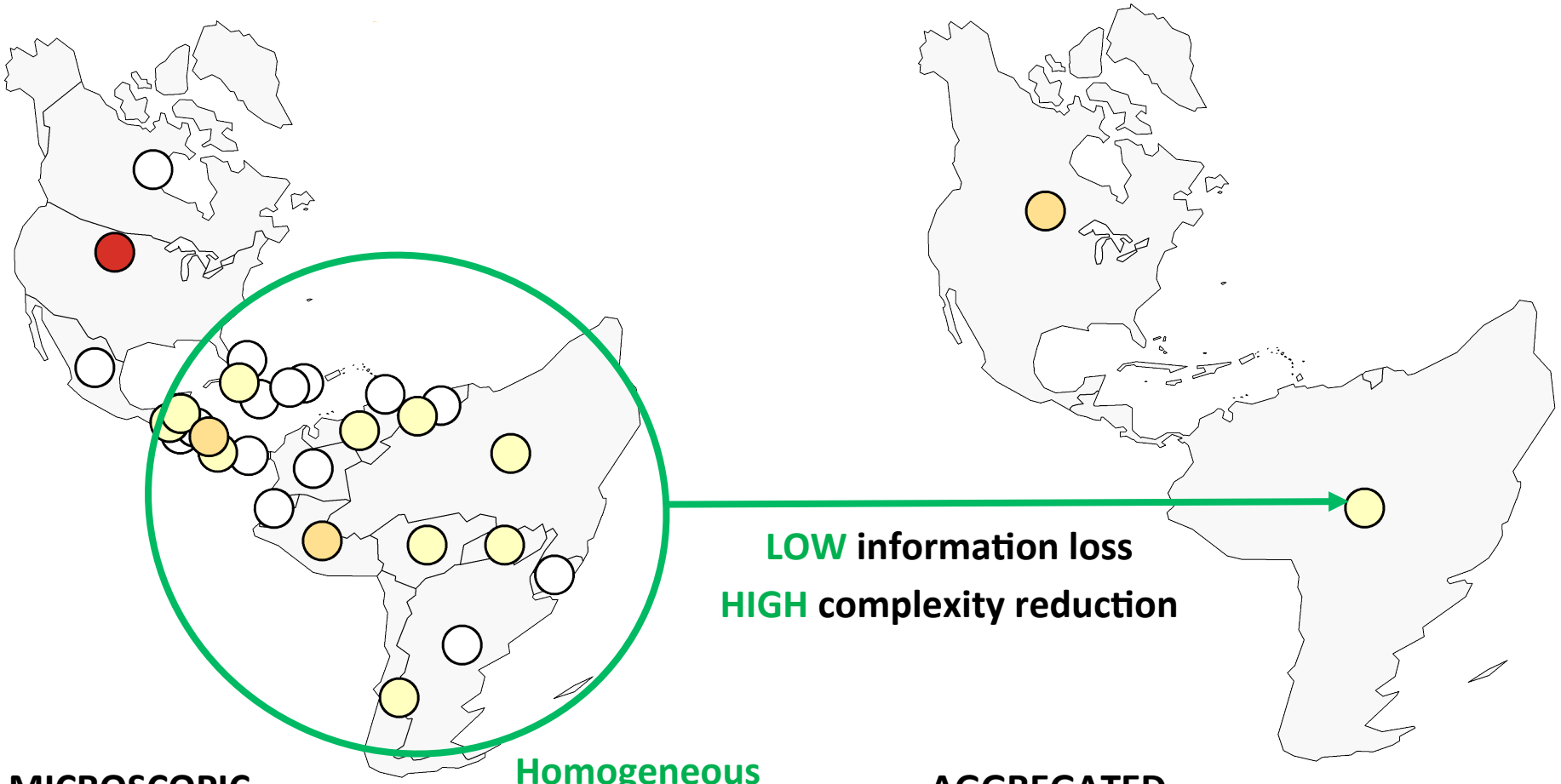


AGGREGATED REPRESENTATION

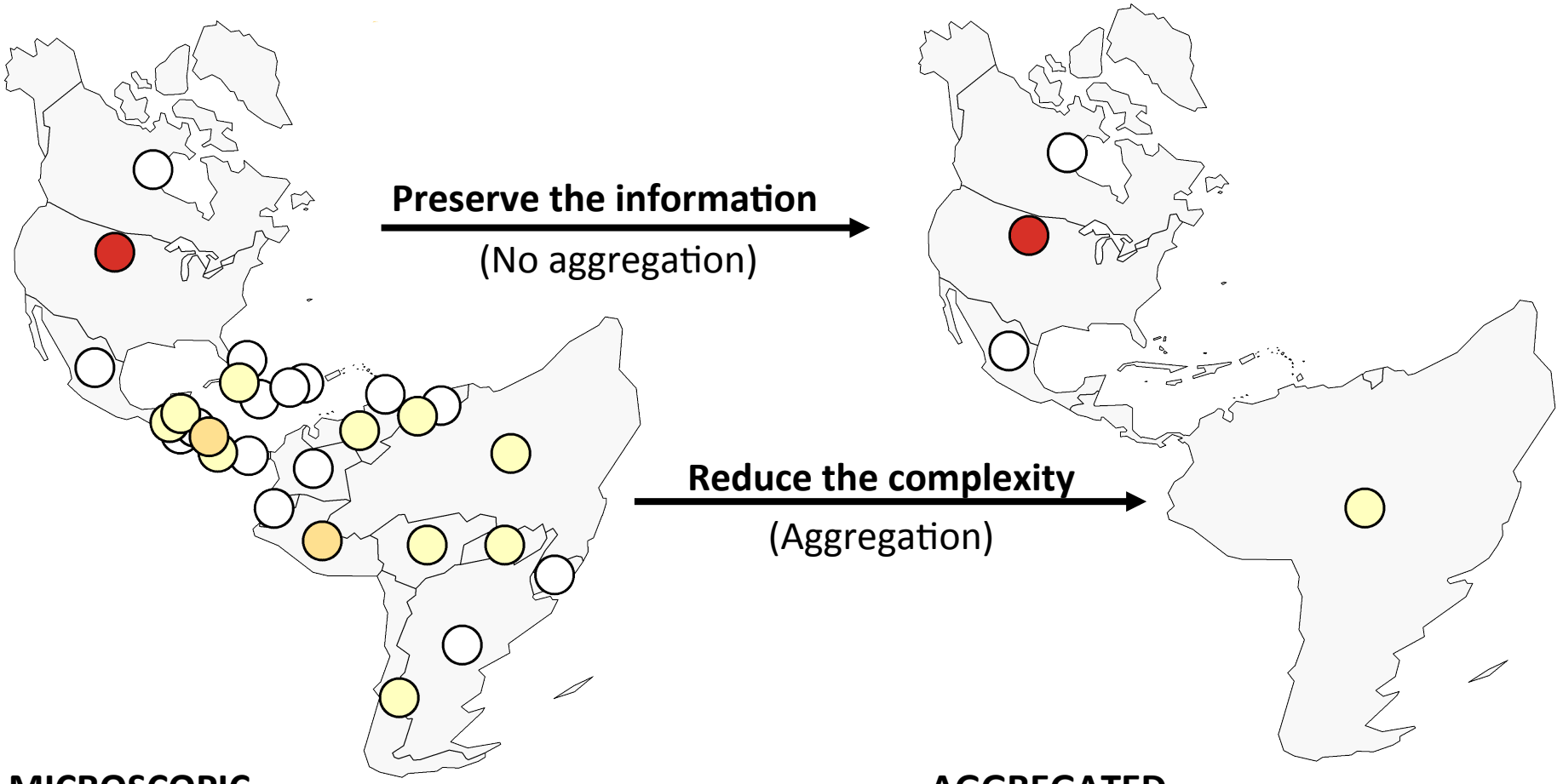
Complexity and Information



Complexity and Information



Complexity and Information



MICROSCOPIC REPRESENTATION

AGGREGATED REPRESENTATION

Quality Measures

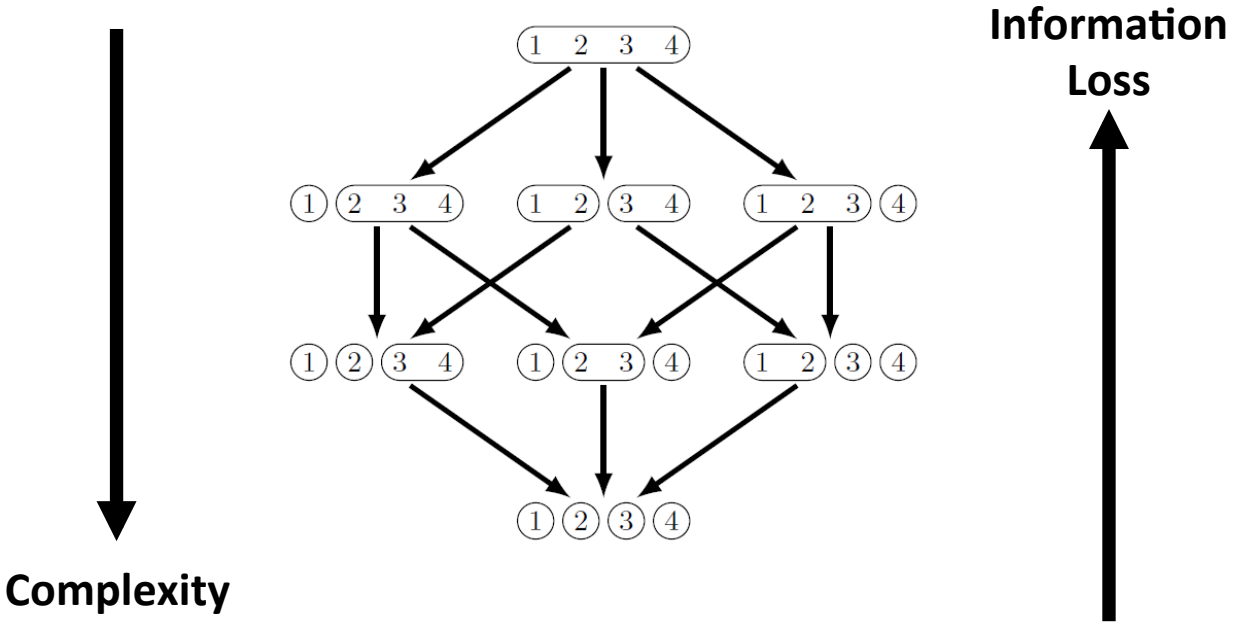
[Lamarche-Perrin *et al.*, ECCS 2012]

Complexity depends on the **tasks** we want to fulfill and the **description tools** that are available to do so

[Bonabeau and Dessalles, 1997]

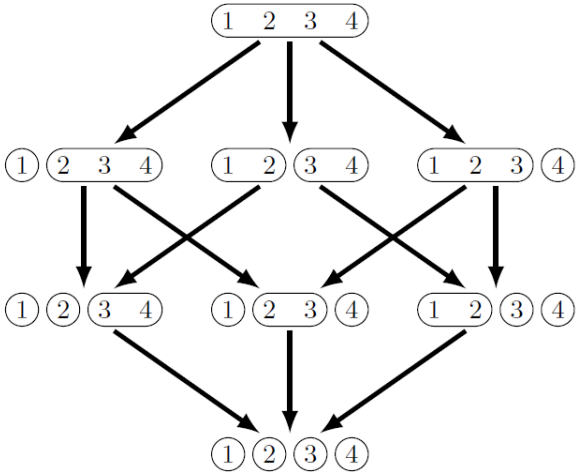
Information loss is measured by the **KL-divergence** between two probabilistic distributions

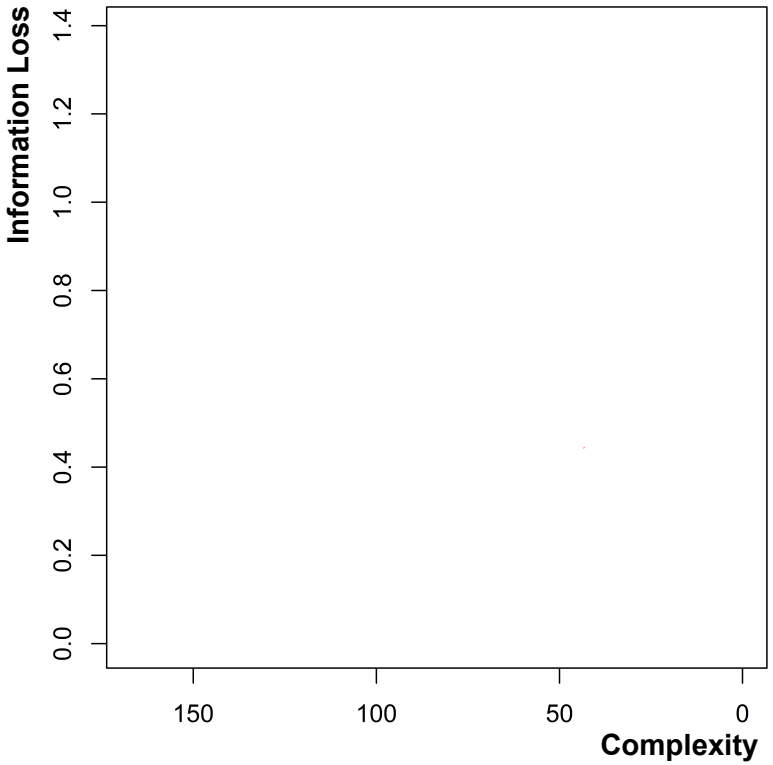
[Kullback et Leibler, 1951]



Optimizing the Partition Qualities

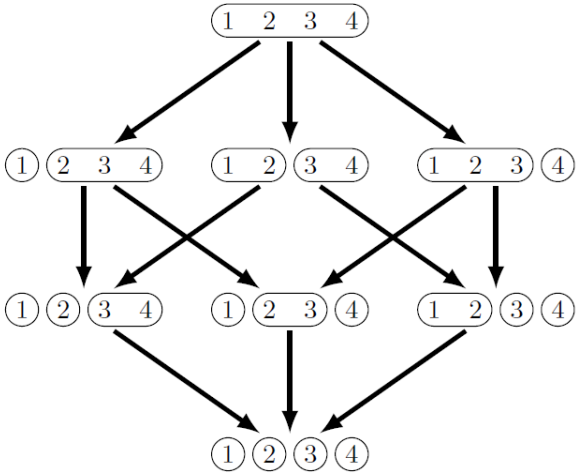
Two criteria that should be optimized

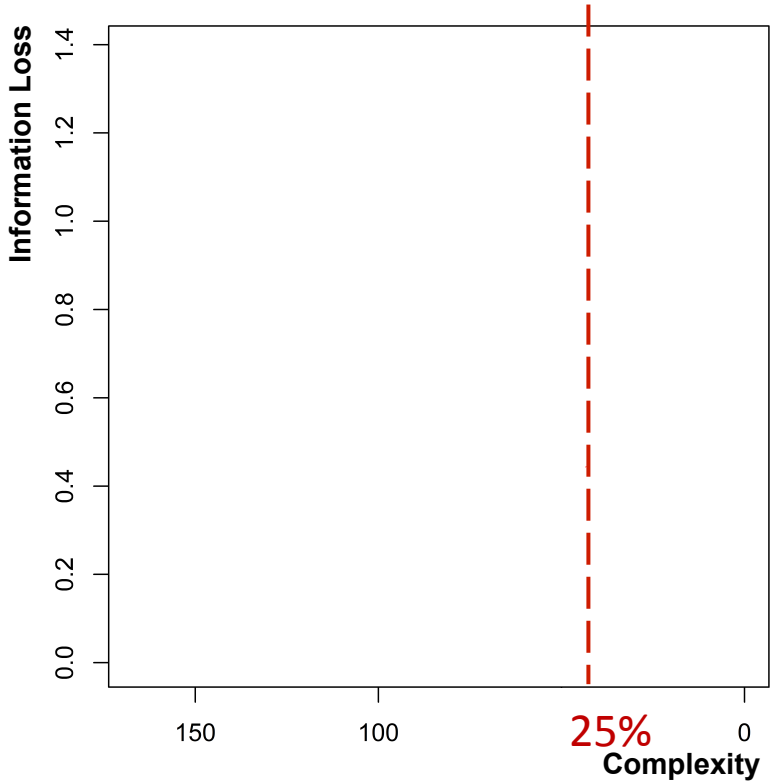




Optimizing the Partition Qualities

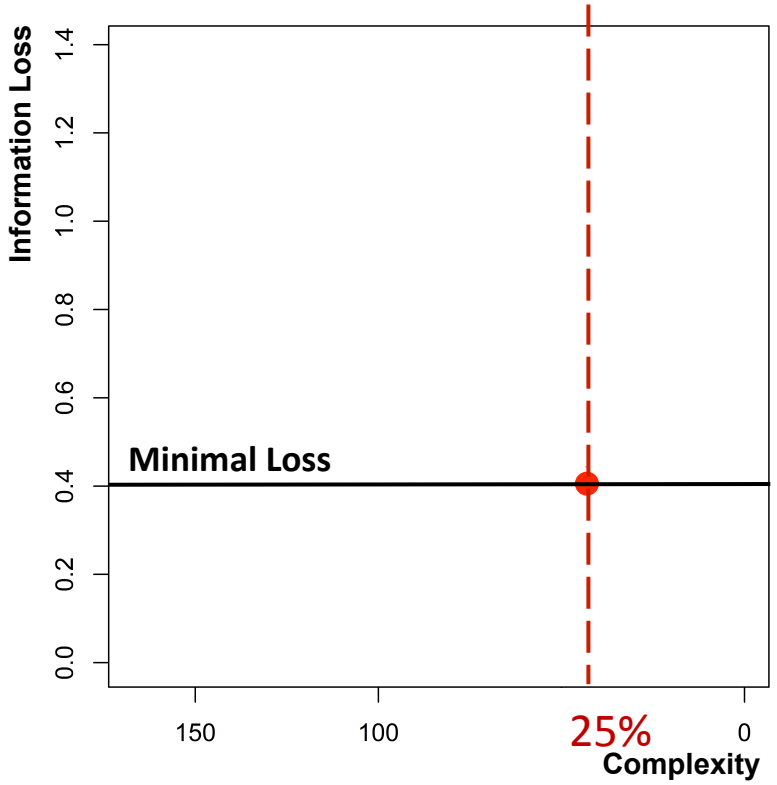
Two criteria that should be optimized

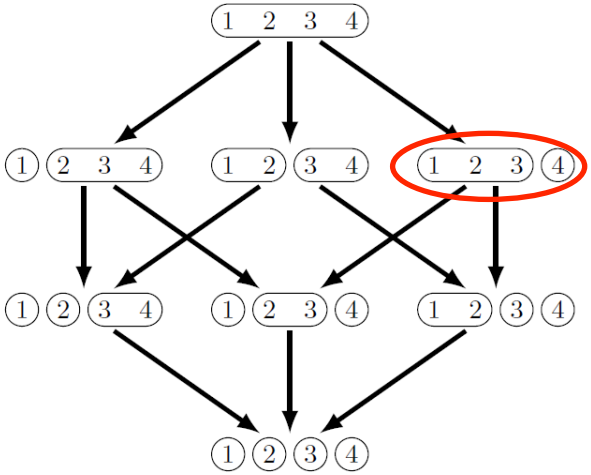


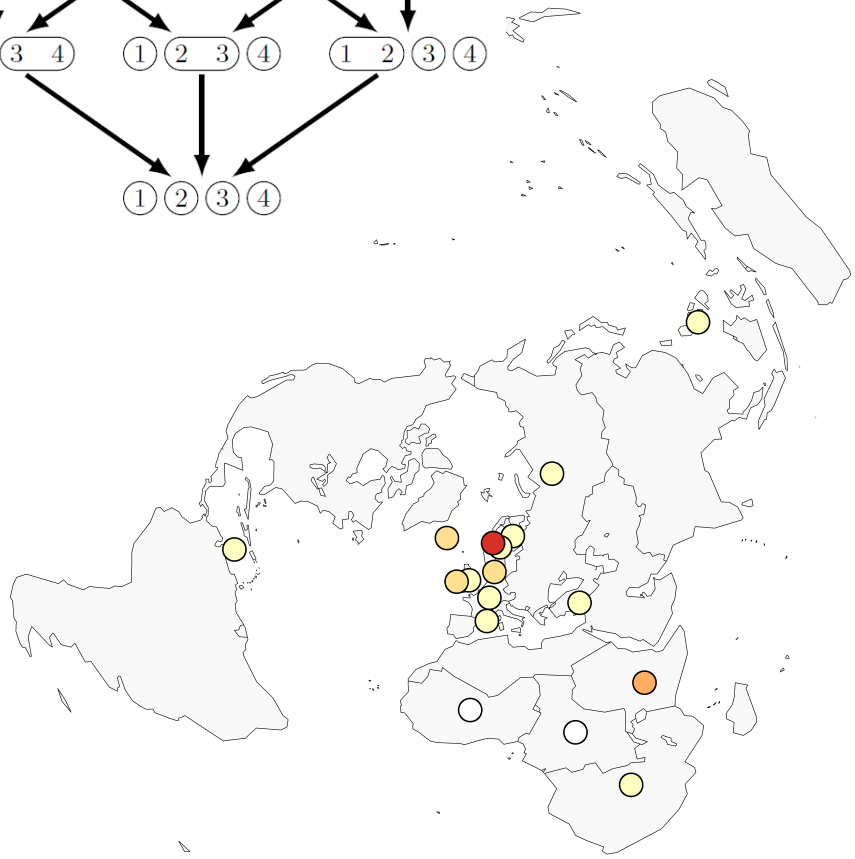


Optimizing the Partition Qualities

Two criteria that should be optimized

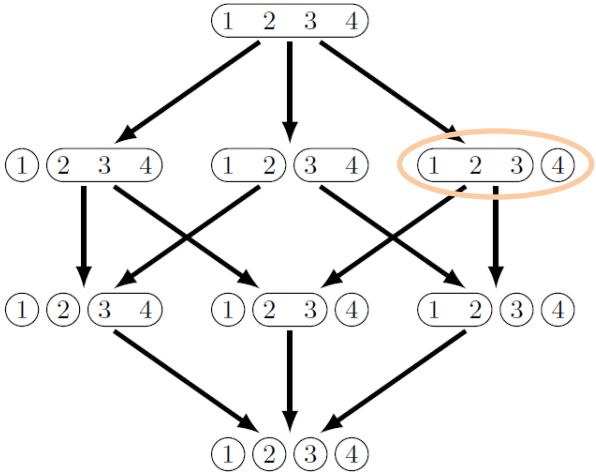


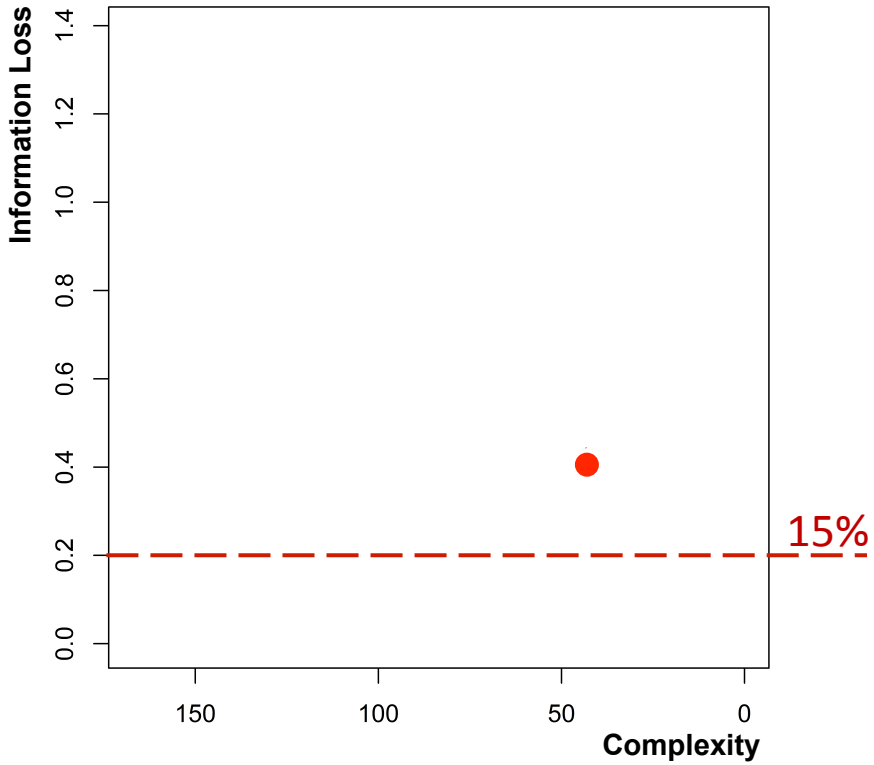




Optimizing the Partition Qualities

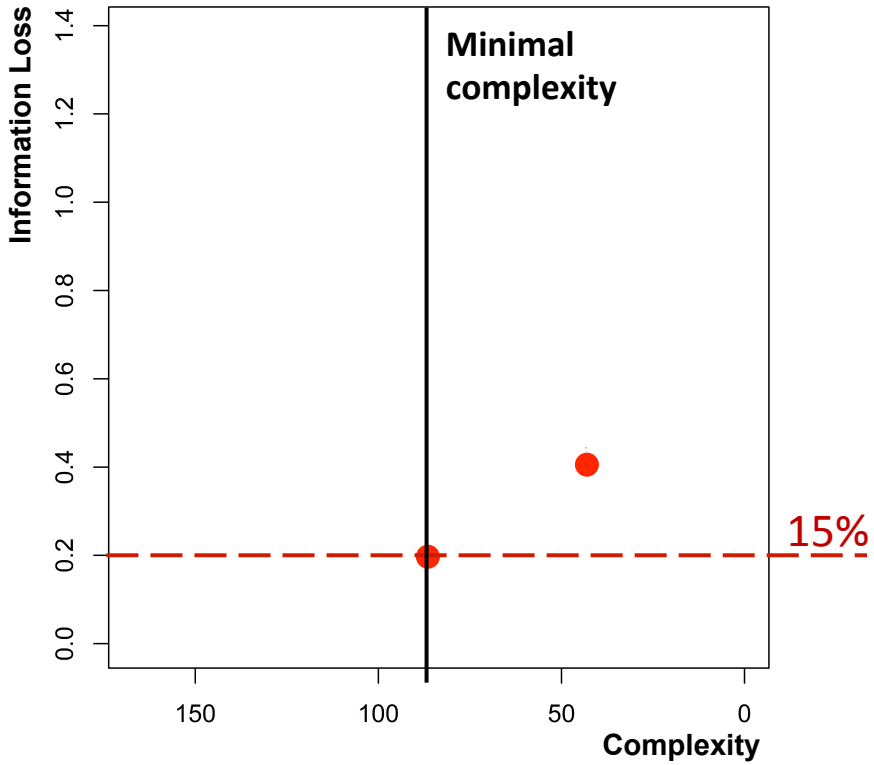
Two criteria that should be optimized

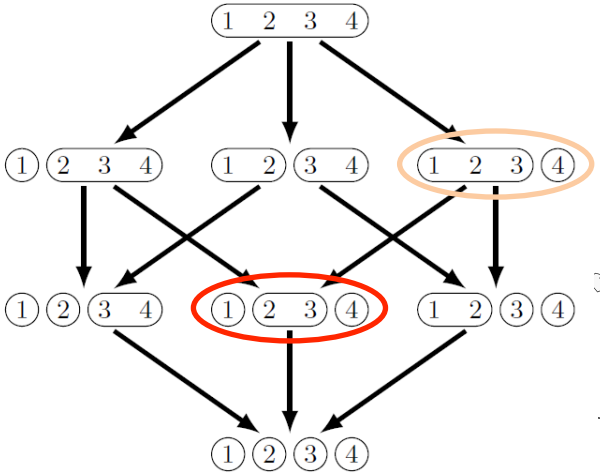


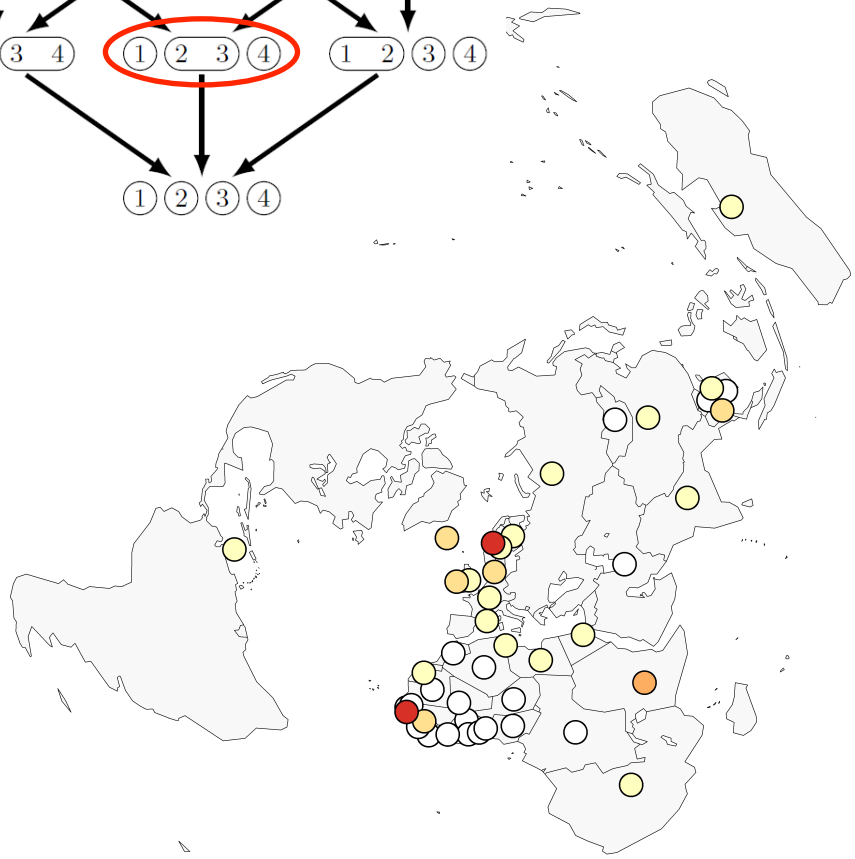


Optimizing the Partition Qualities

Two criteria that should be optimized

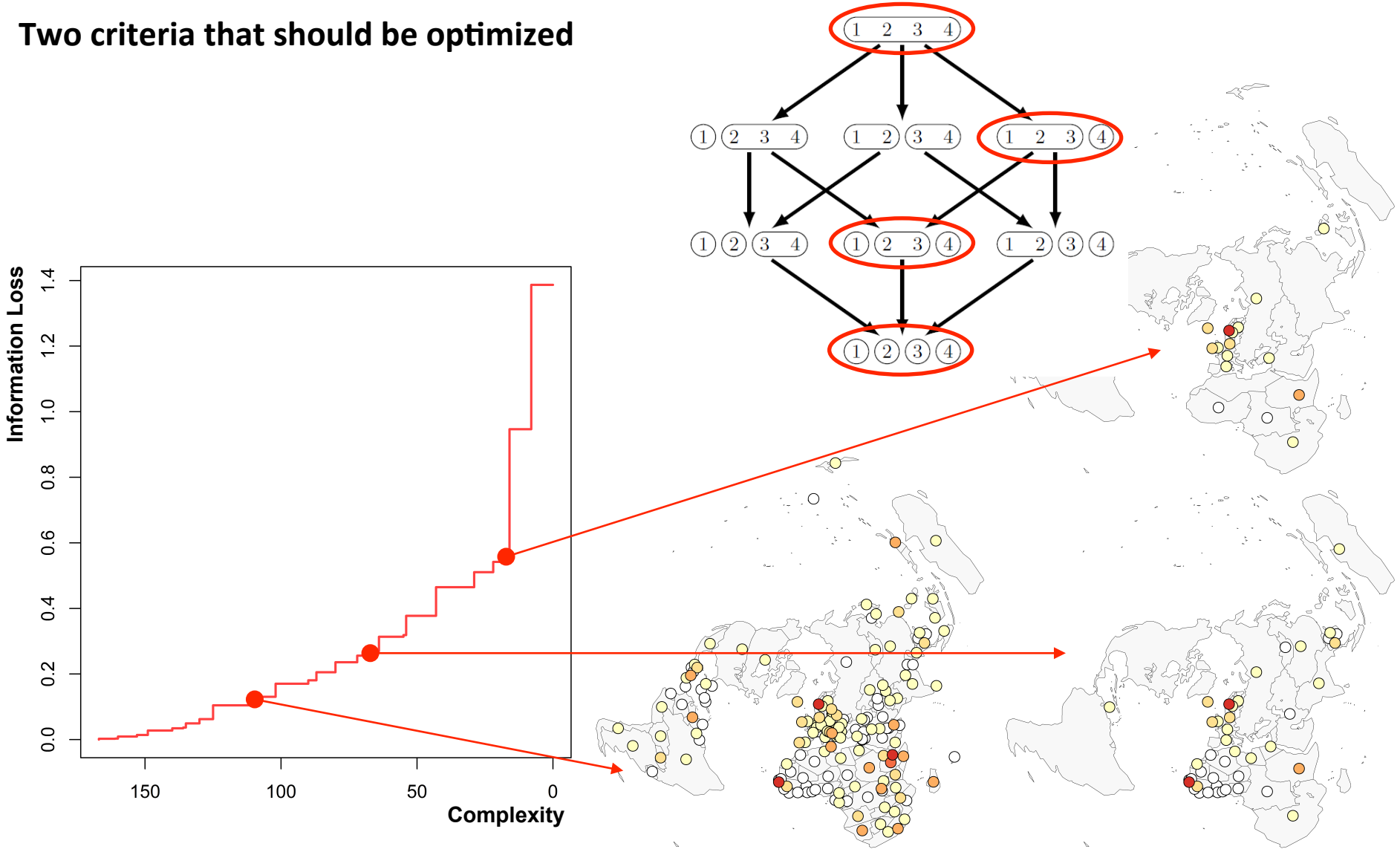






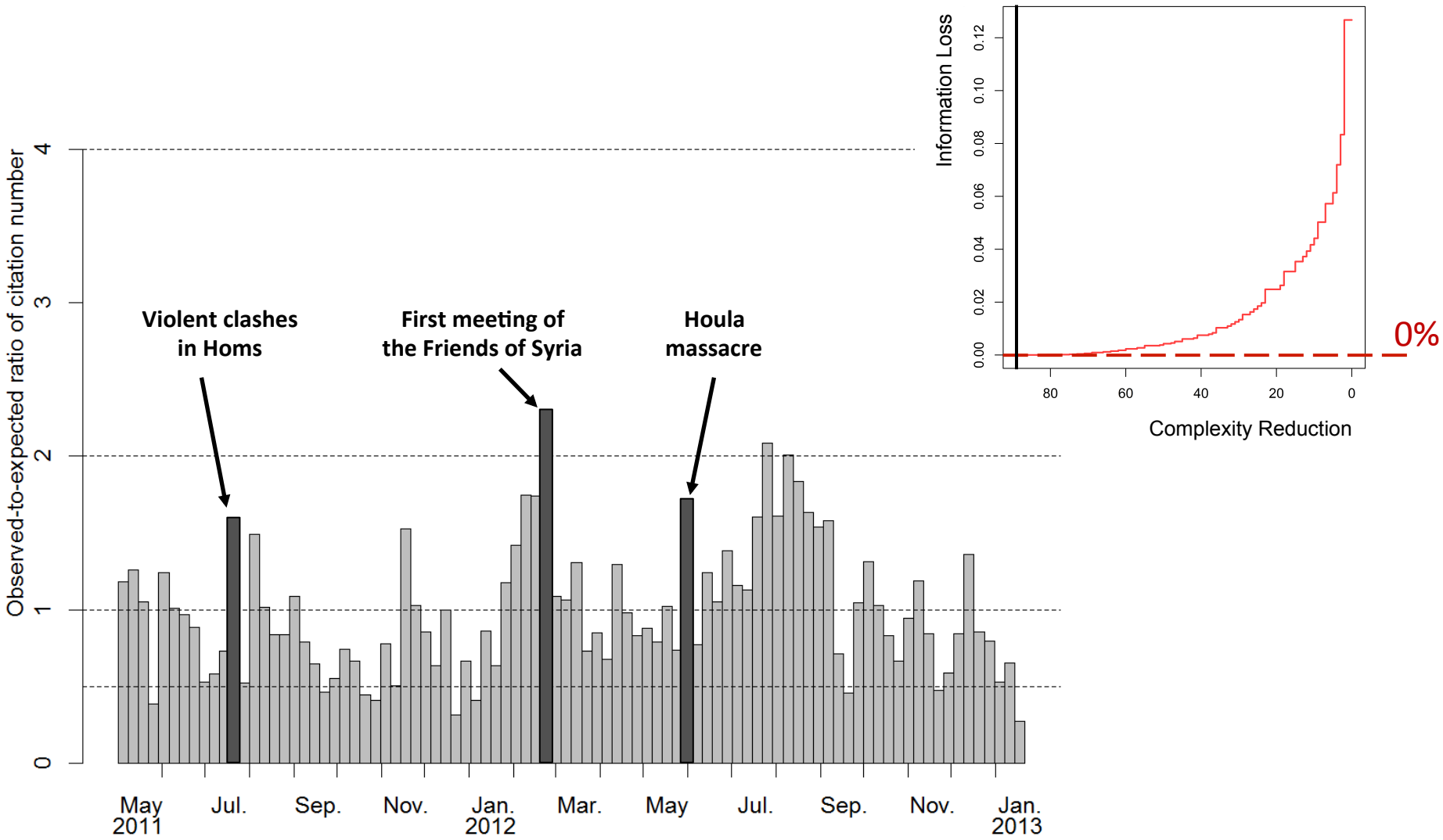
Optimizing the Partition Qualities

Two criteria that should be optimized



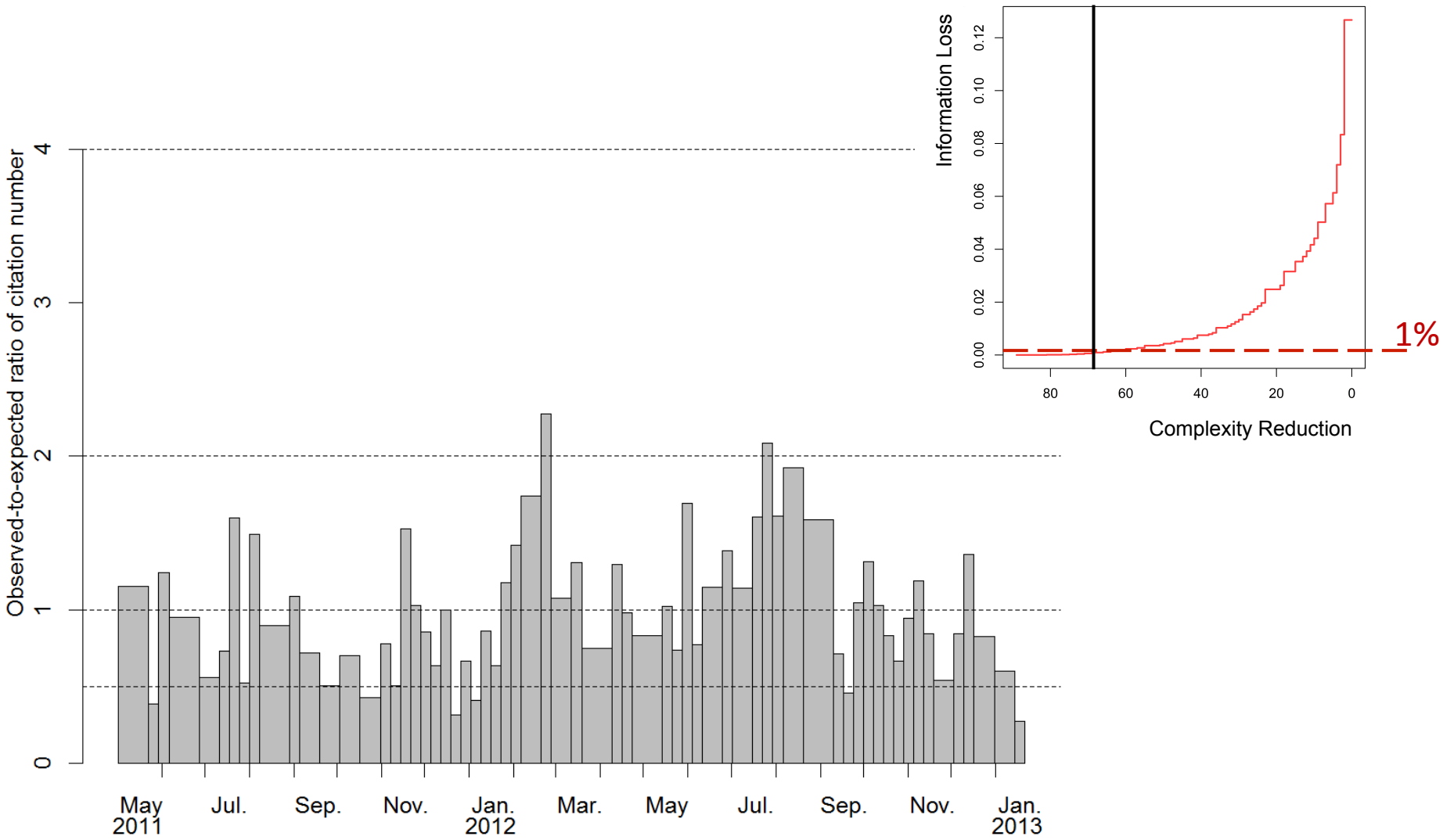
The Syrian civil war according to LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]



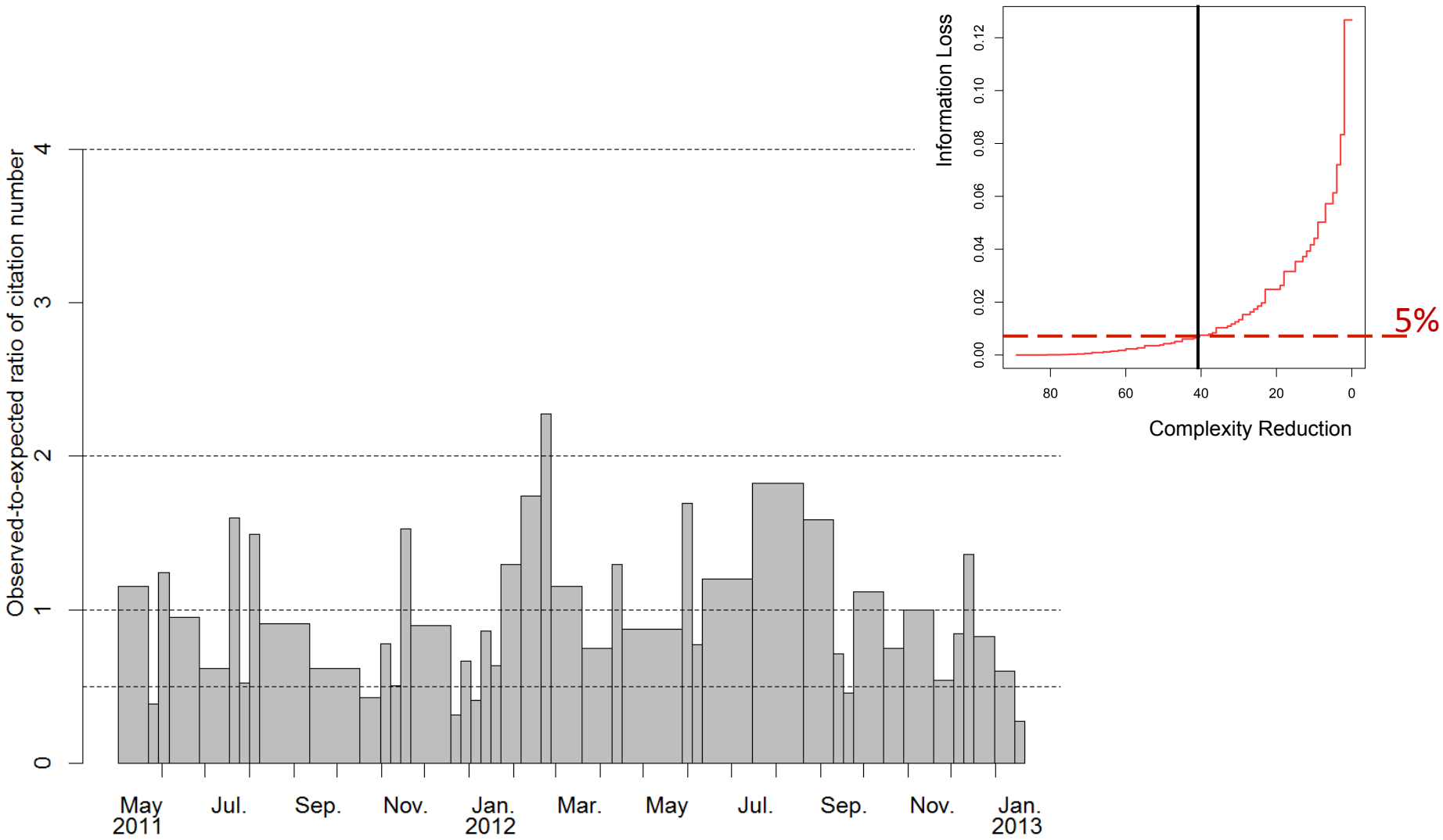
The Syrian civil war according to LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]



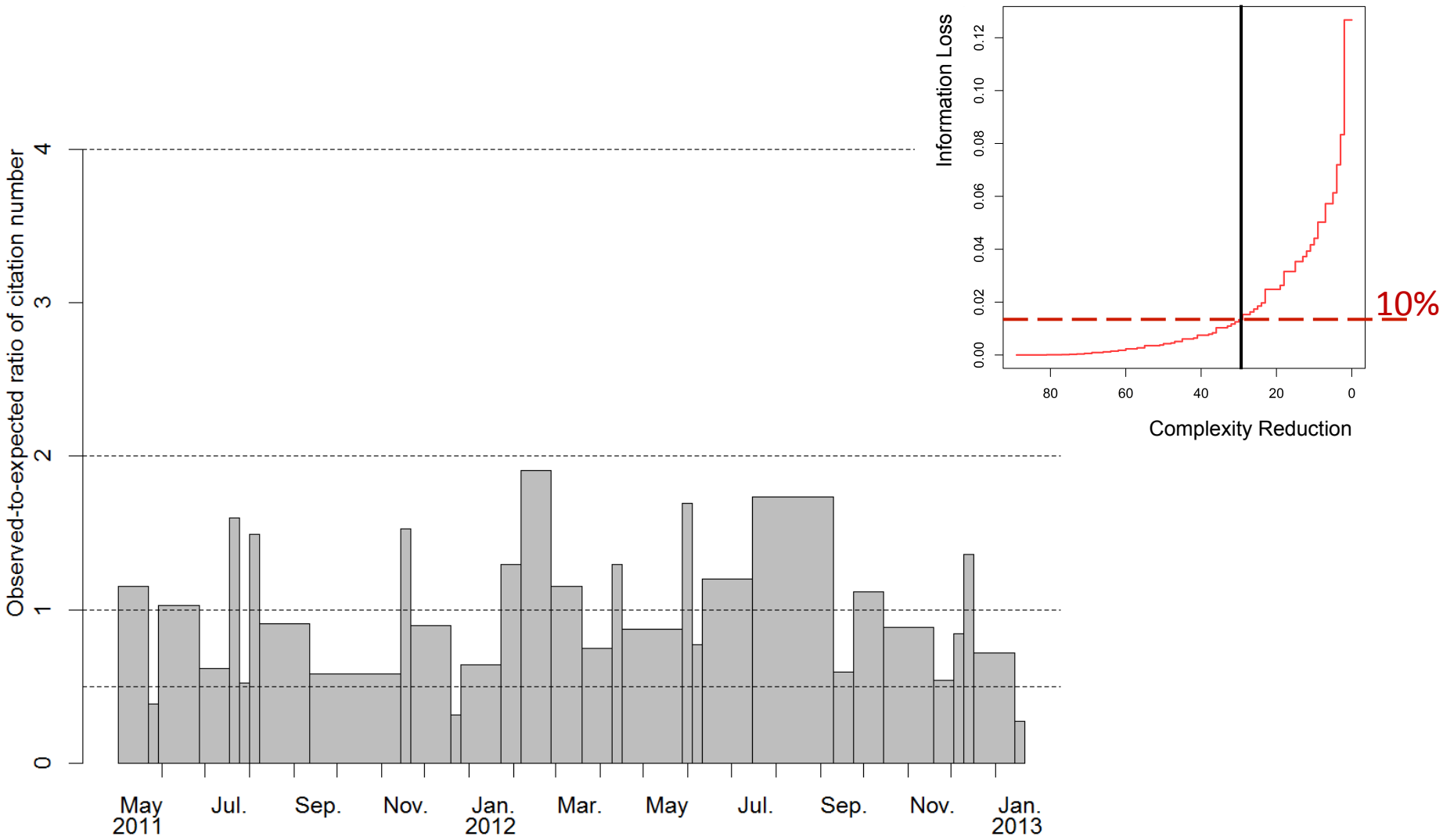
The Syrian civil war according to LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]



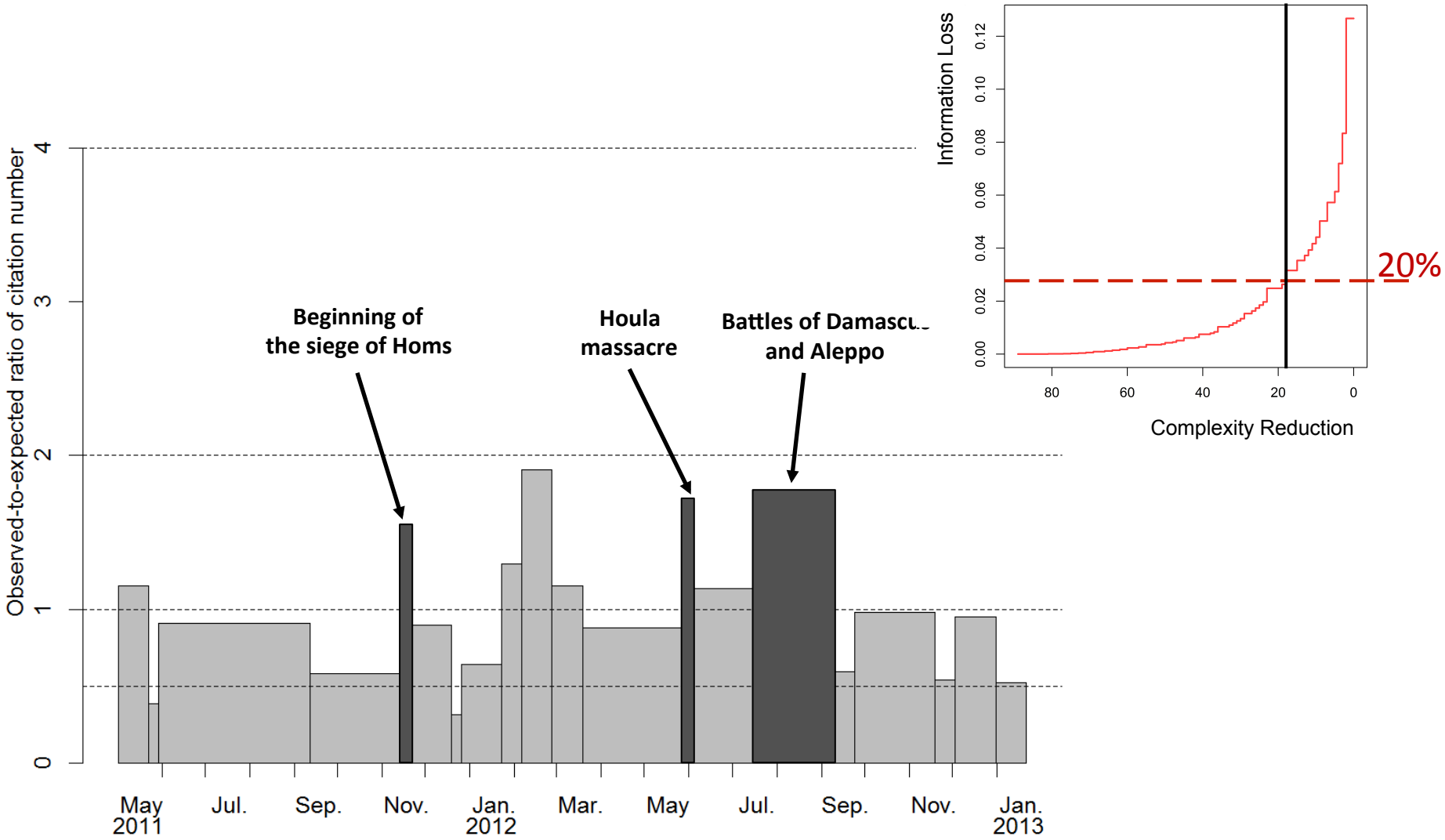
The Syrian civil war according to LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]



The Syrian civil war according to LE MONDE

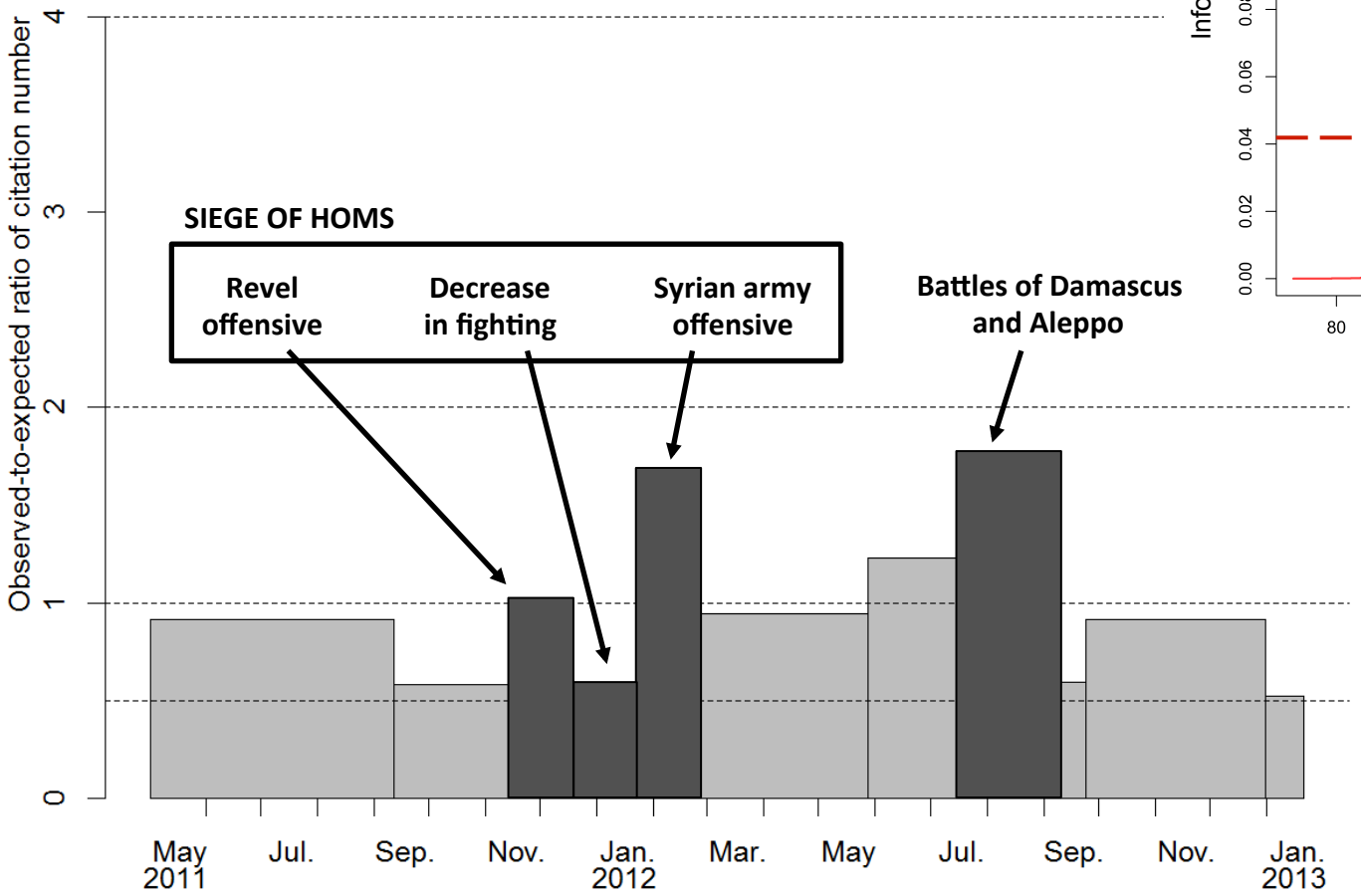
[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

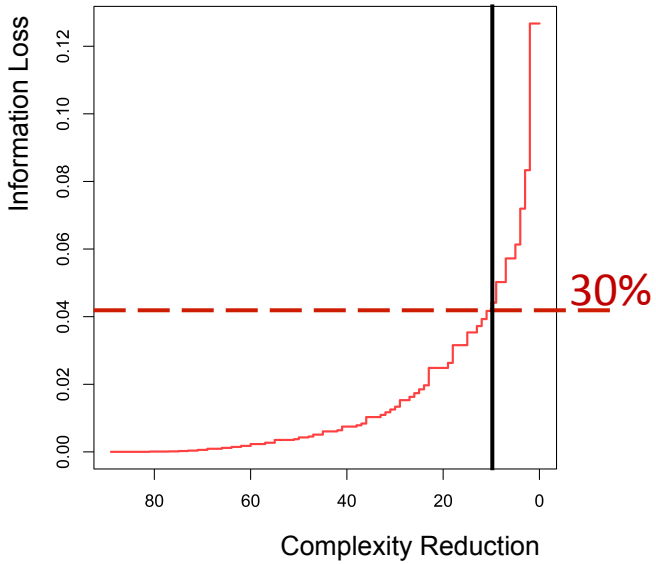


The Syrian civil war according to LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

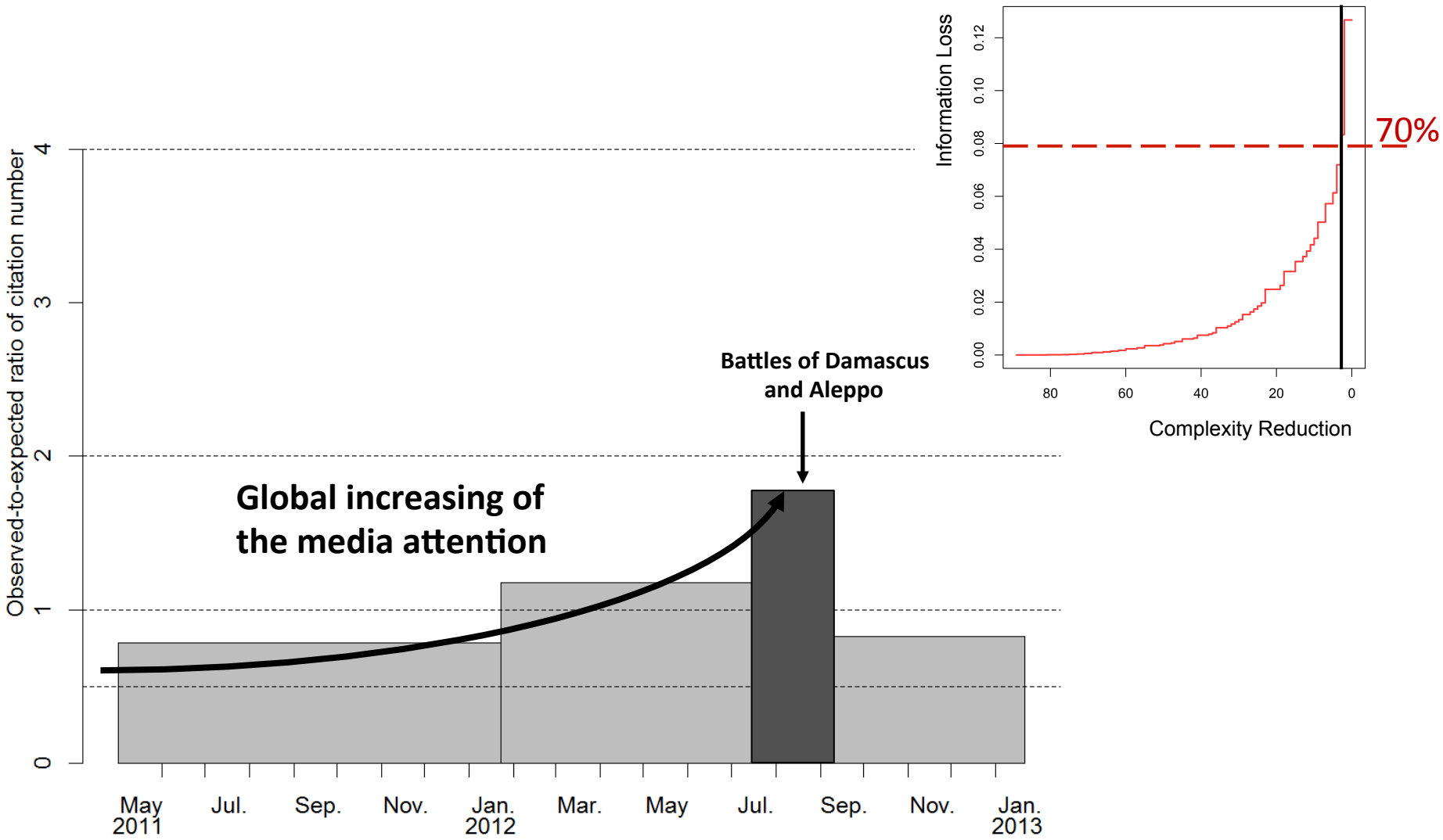
Source: Wikipedia
Timeline of the Syrian civil war
Siege of Homs





The Syrian civil war according to LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

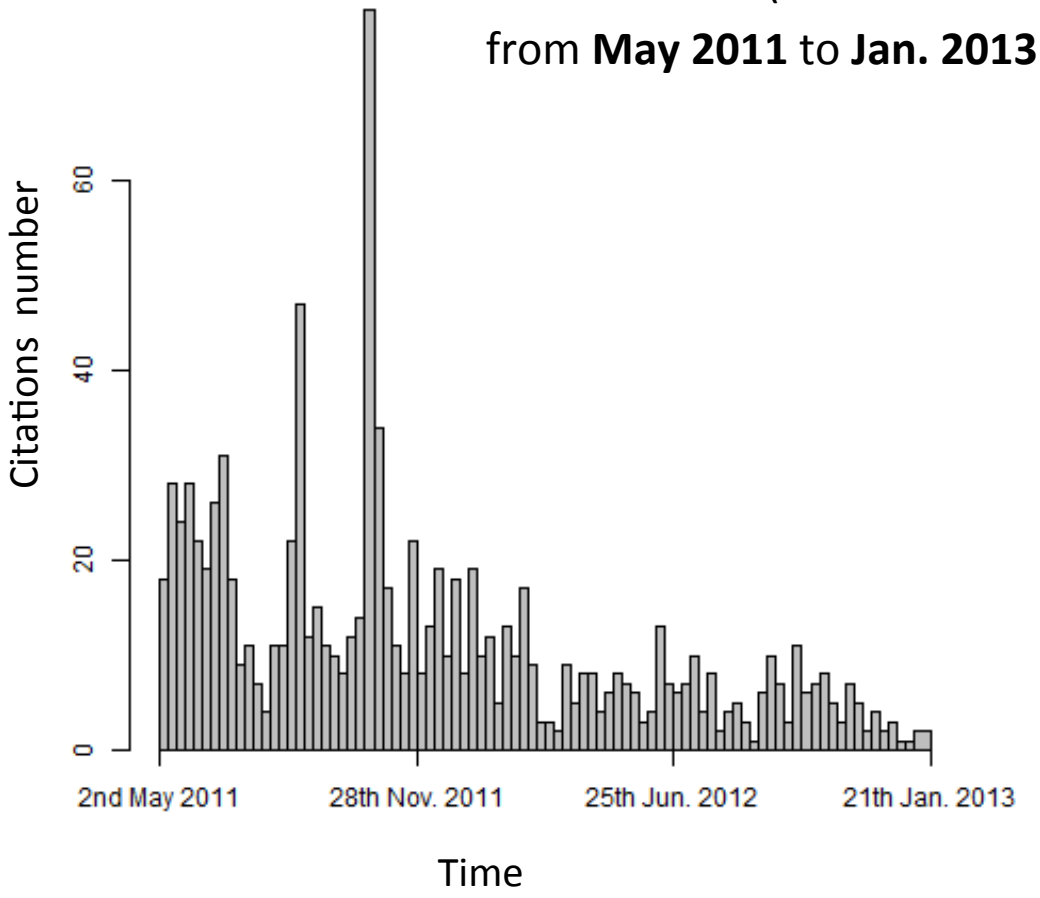


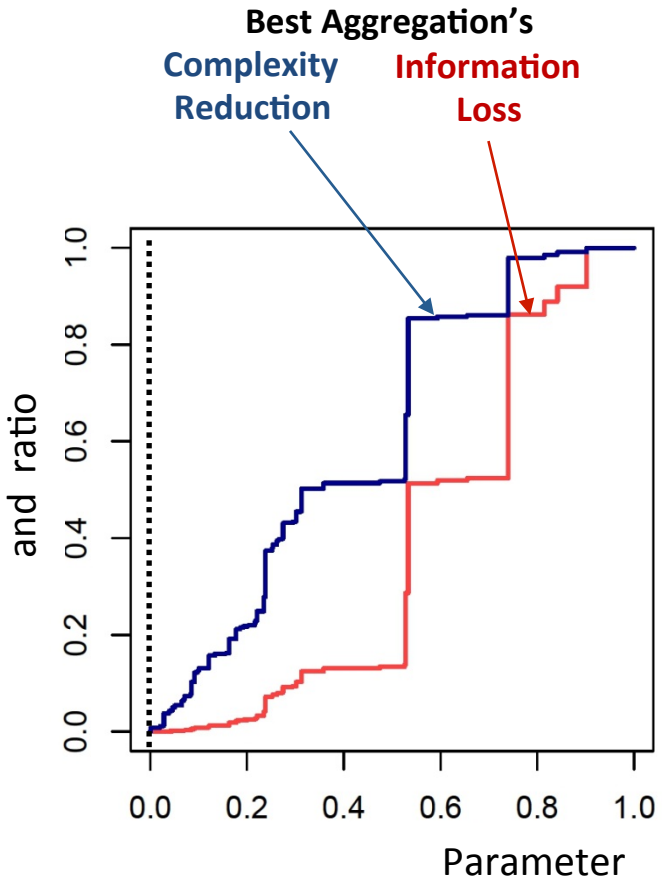
Best Temporal Aggregation

Citations of "Tunisia"

The Guardian (60.000 articles)

from **May 2011** to **Jan. 2013**



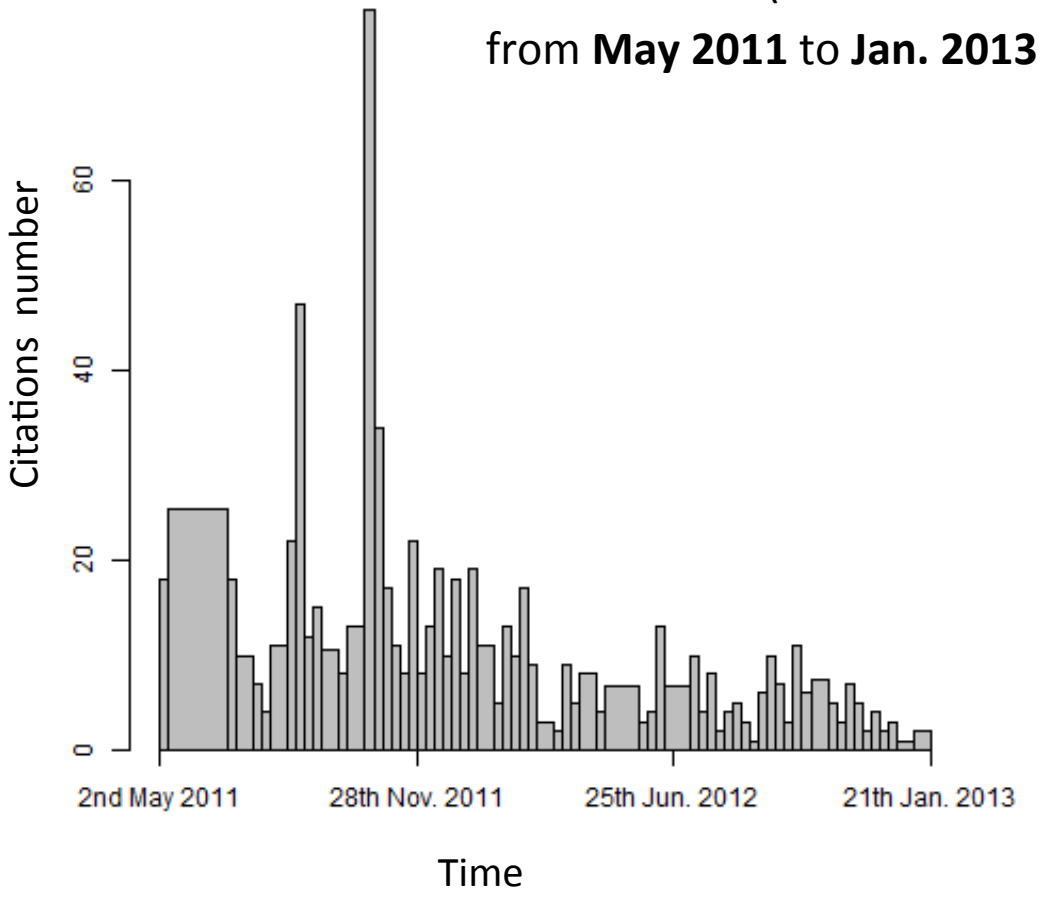


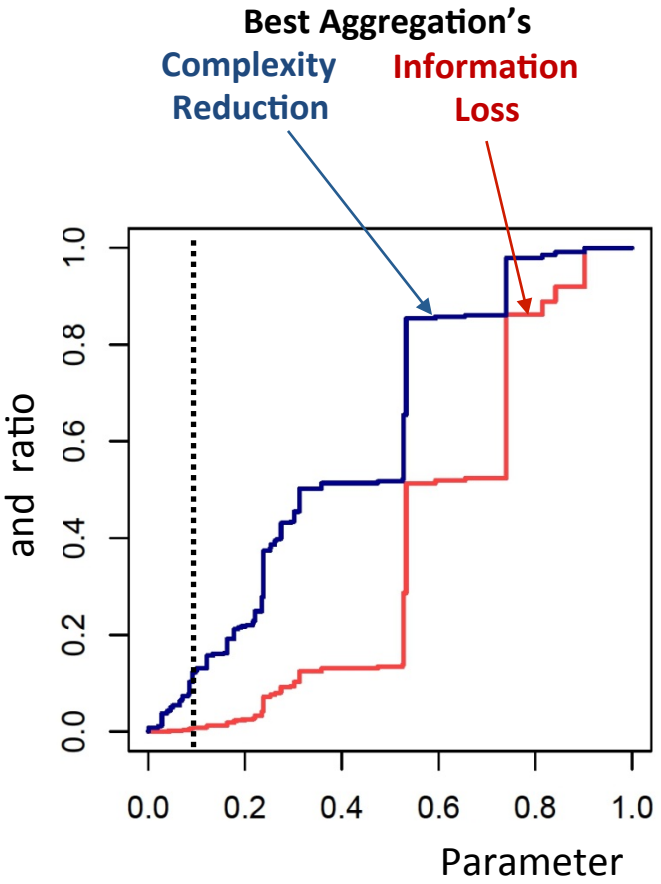
Best Temporal Aggregation

Citations of "Tunisia"

The Guardian (60.000 articles)

from **May 2011** to **Jan. 2013**



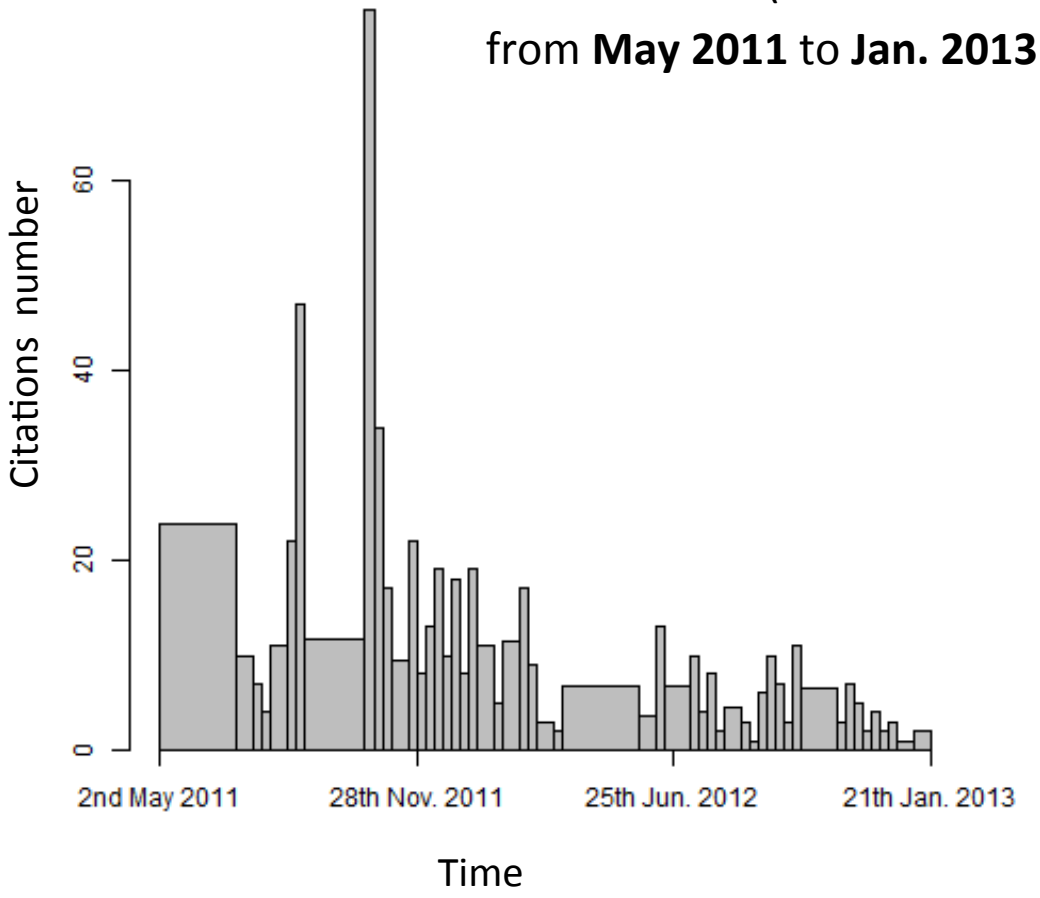


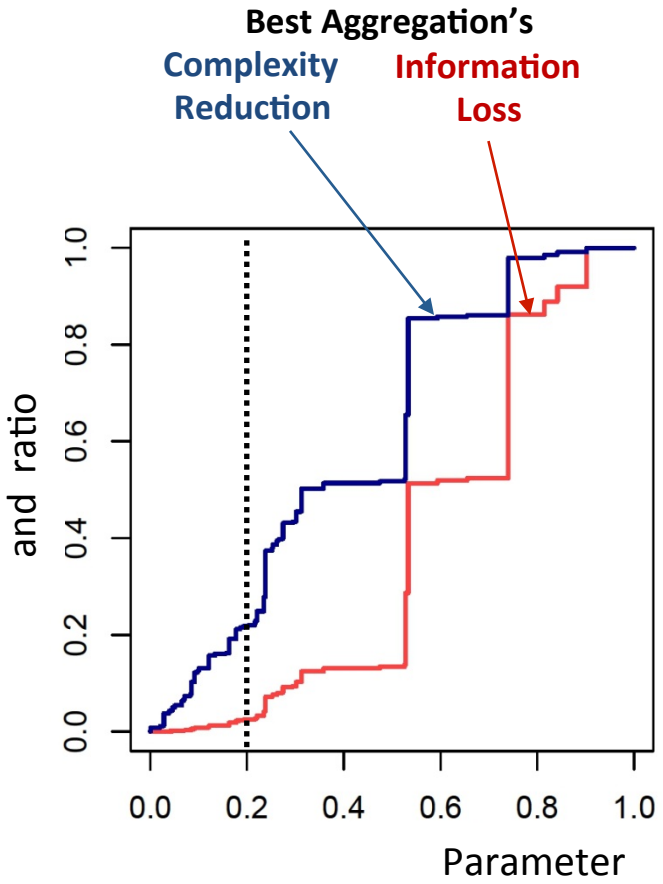
Best Temporal Aggregation

Citations of "Tunisia"

The Guardian (60.000 articles)

from **May 2011** to **Jan. 2013**



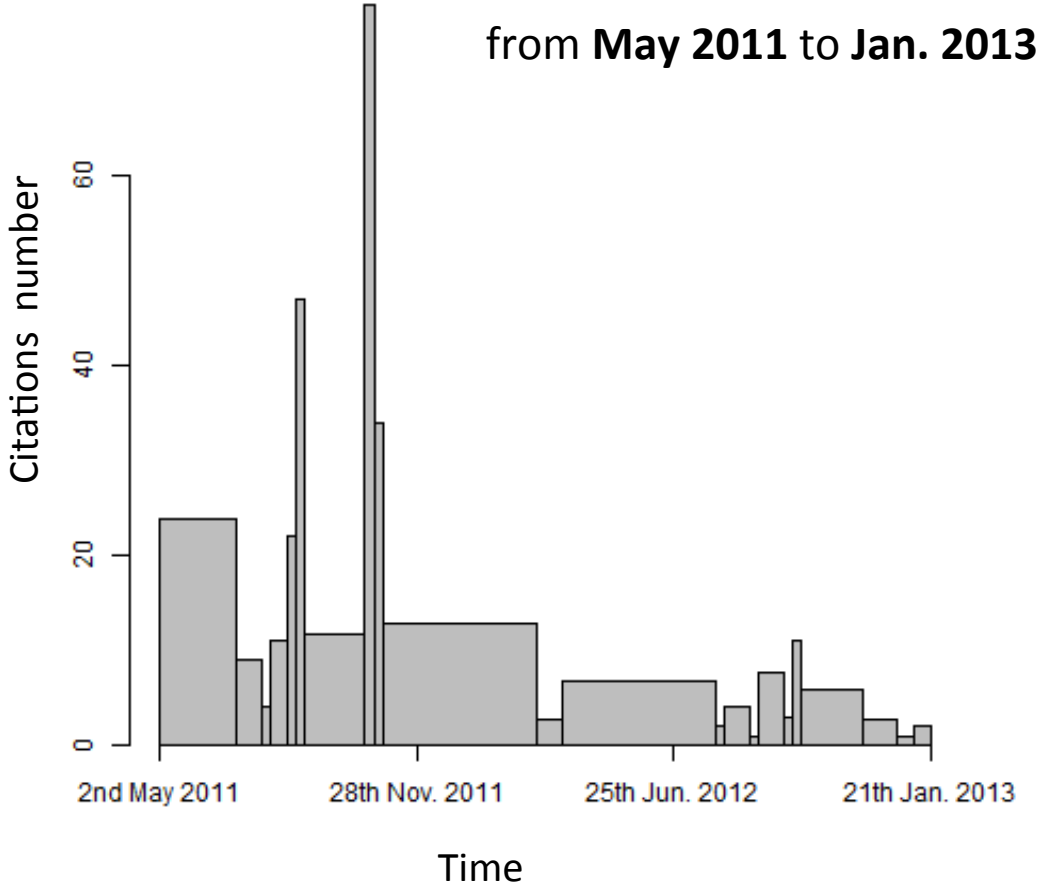


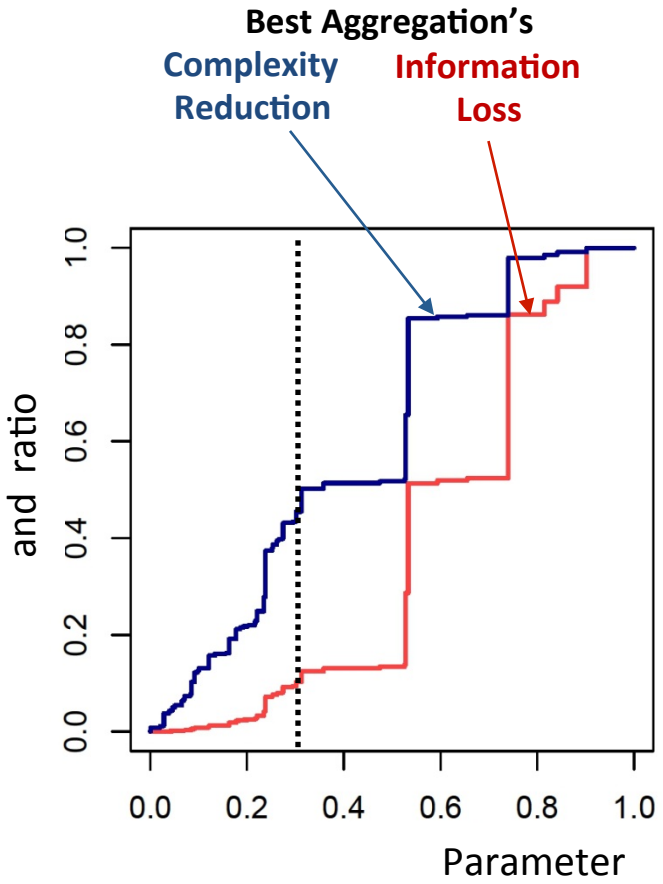
Best Temporal Aggregation

Citations of "Tunisia"

The Guardian (60.000 articles)

from **May 2011** to **Jan. 2013**



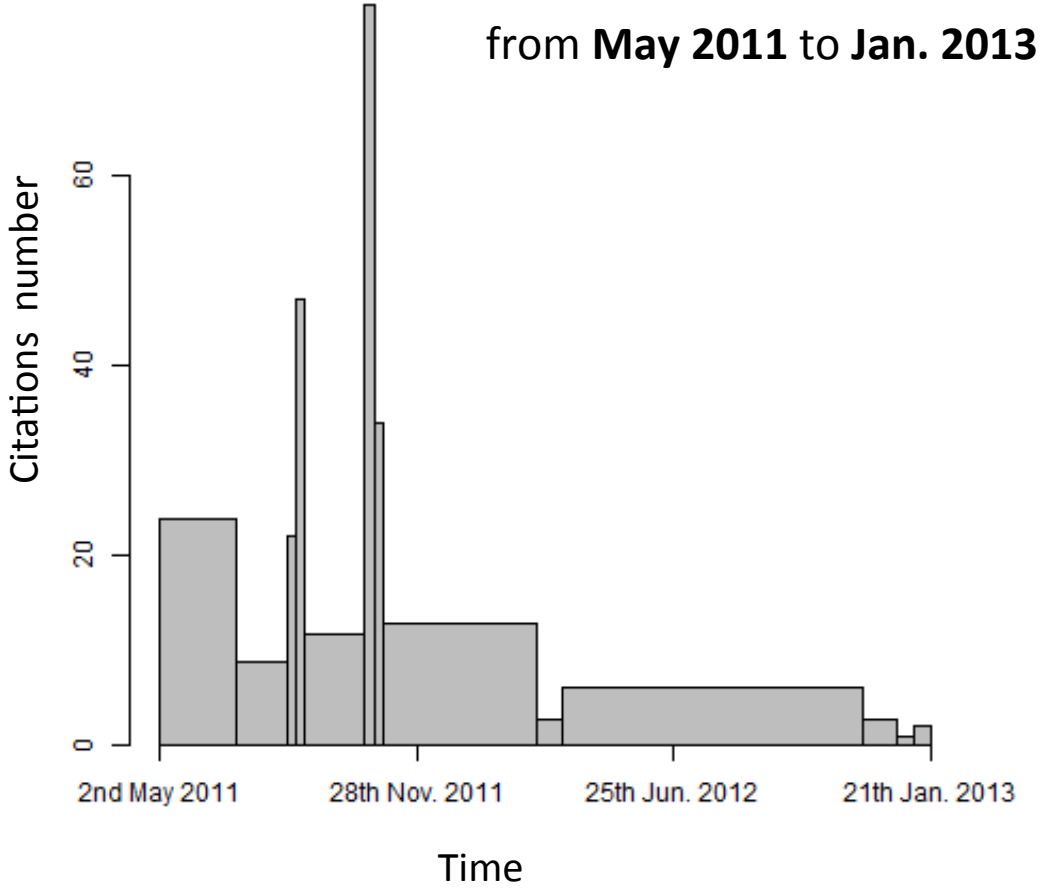


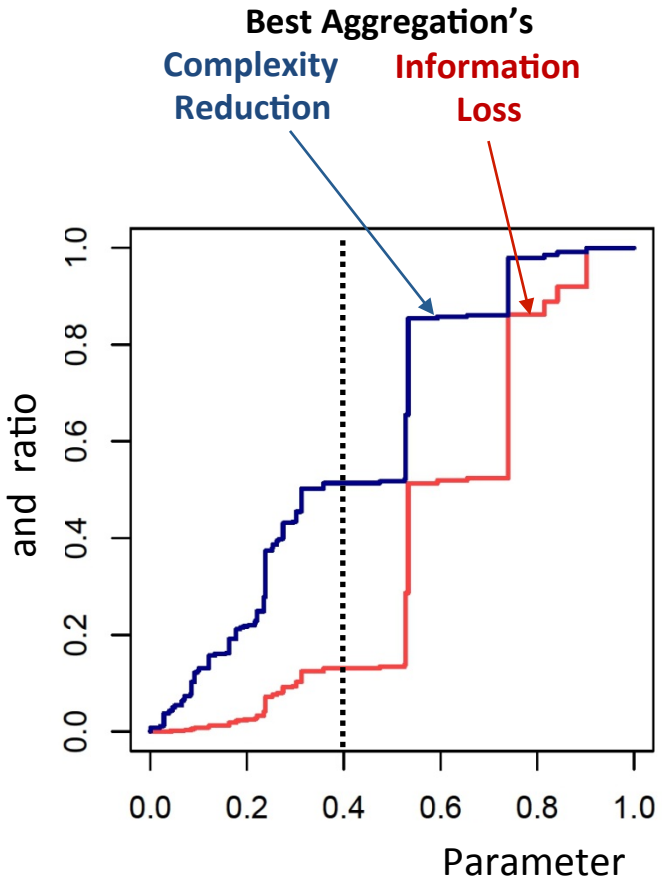
Best Temporal Aggregation

Citations of “Tunisia”

The Guardian (60.000 articles)

from **May 2011** to **Jan. 2013**



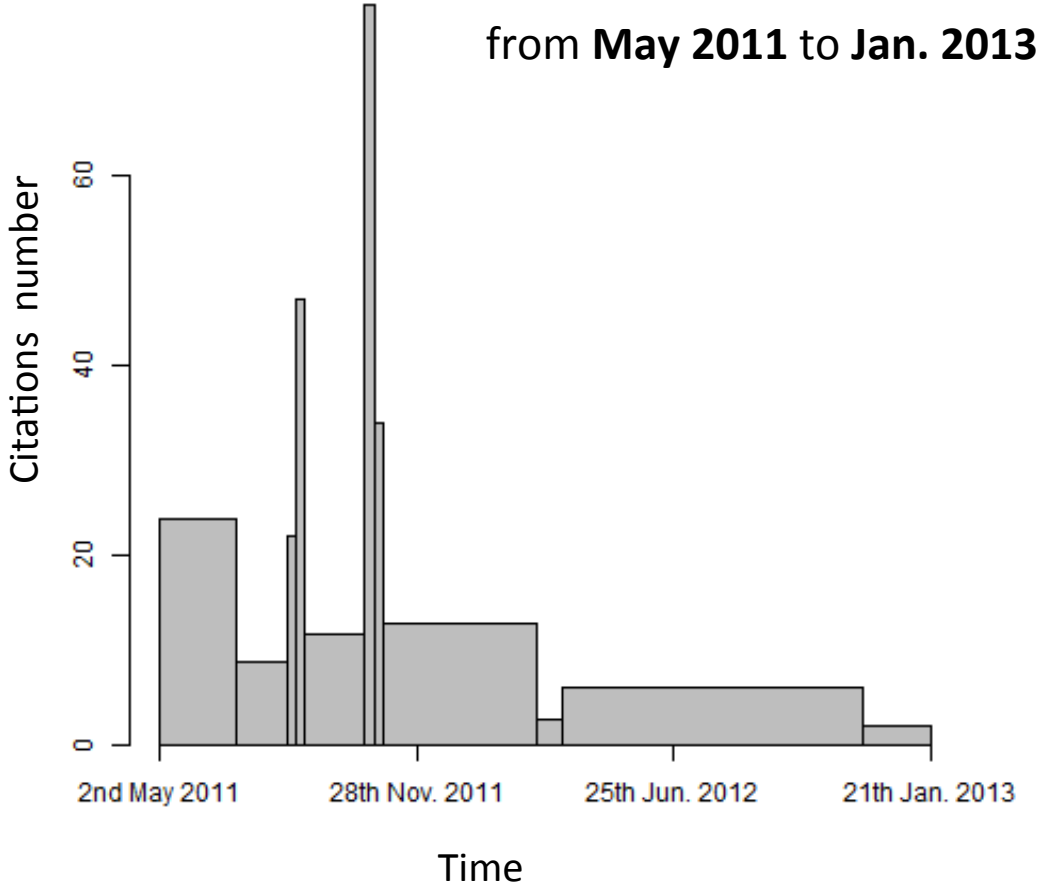


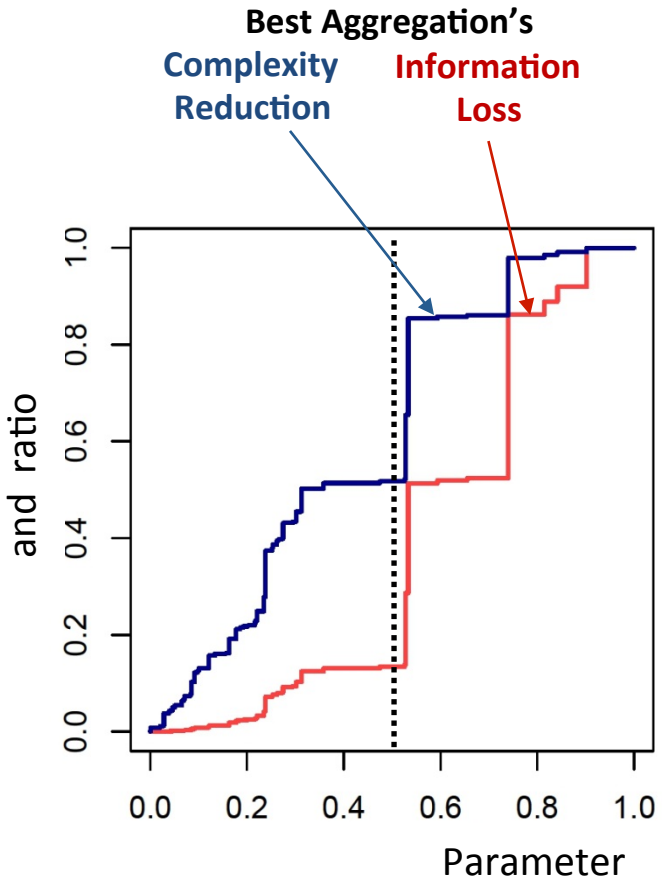
Best Temporal Aggregation

Citations of “Tunisia”

The Guardian (60.000 articles)

from **May 2011** to **Jan. 2013**





Lamarche-Perrin Approach

To characterize the aggregation process

→ The algebra of possible partitions

To preserve the system's semantics

→ A constrained partitioning method

To aggregate according to several dimension

→ Some constraints expressing the system's topology

To evaluate and compare the representations

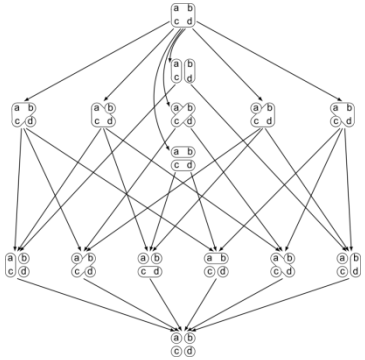
→ Some measures of complexity and information

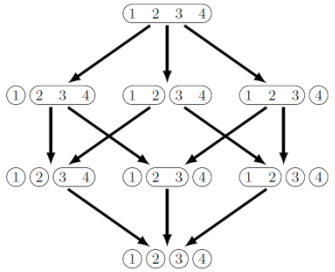
To offer several granularity levels

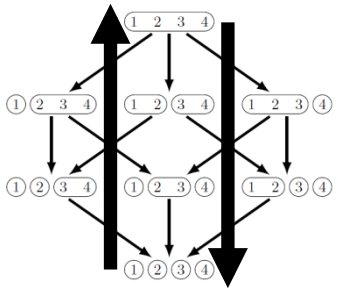
→ The optimization of a compromise

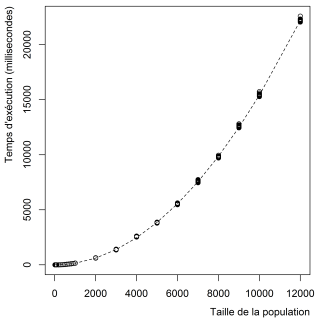
To compute the best representations

→ A generic algorithm of constrained optimization





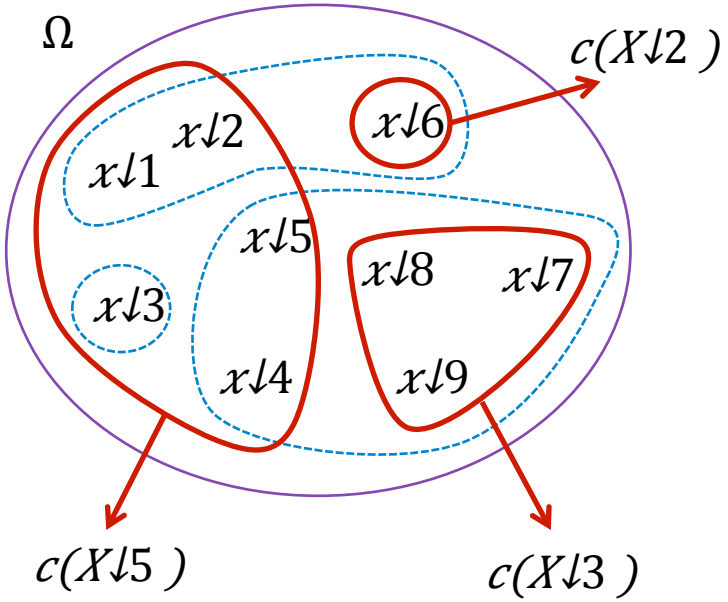




The Set Partitioning Problem

Given:

- a set of individuals $\Omega=\{x_1,\ldots,x_n\}$
- a set of admissible parts $\mathcal{P}=\{X_1,\ldots,X_m\}\subset 2^{\Omega}$
- a cost function $c:\mathcal{P}\to\mathbb{R}$
- the corresponding set of admissible partitions $\mathfrak{P}=\{\mathcal{X}\subset\mathcal{P} \text{ such that } \mathcal{X} \text{ is a partition of } \Omega\}$



Problem: Find an admissible partition that minimizes the cost function:

$$\mathcal{X}^{*}=\arg\min_{\mathcal{X}\in\mathfrak{P}}\left(\sum_{X\in\mathcal{X}}c(X)\right)$$

→ **NP-complete!**

The Generic 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 $\mathfrak{L}$.

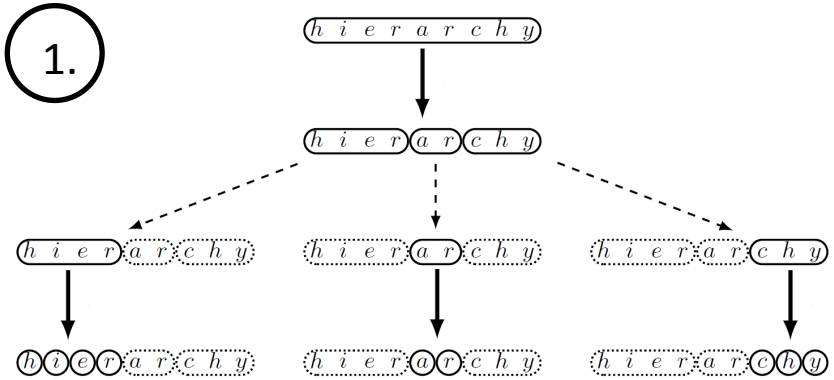
Generic: solve any instance of the SPP
→ but inefficient for special versions

Designing dedicated implementations:

1. Analysing the generic execution
2. Building appropriate data structures
3. Deriving a specialized algorithm

Application to the Hierarchical SPP

1.



2.

Data Structure

- Set of parts: rooted tree
- Optimal partition: cut of the tree
- Algorithm: depth-first search

3.

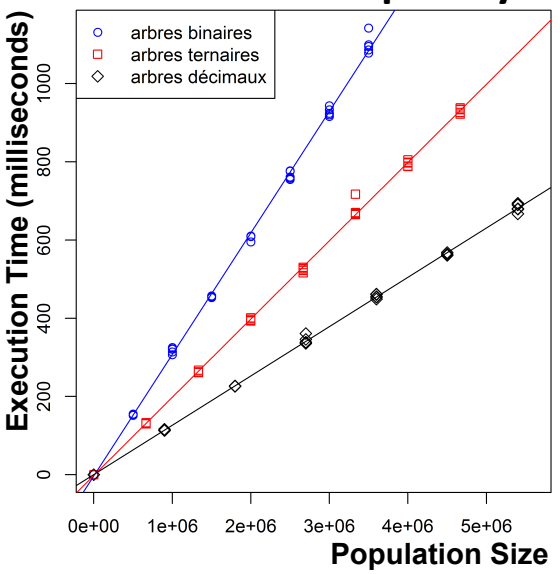
Algorithm 1 for the HSPP

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).

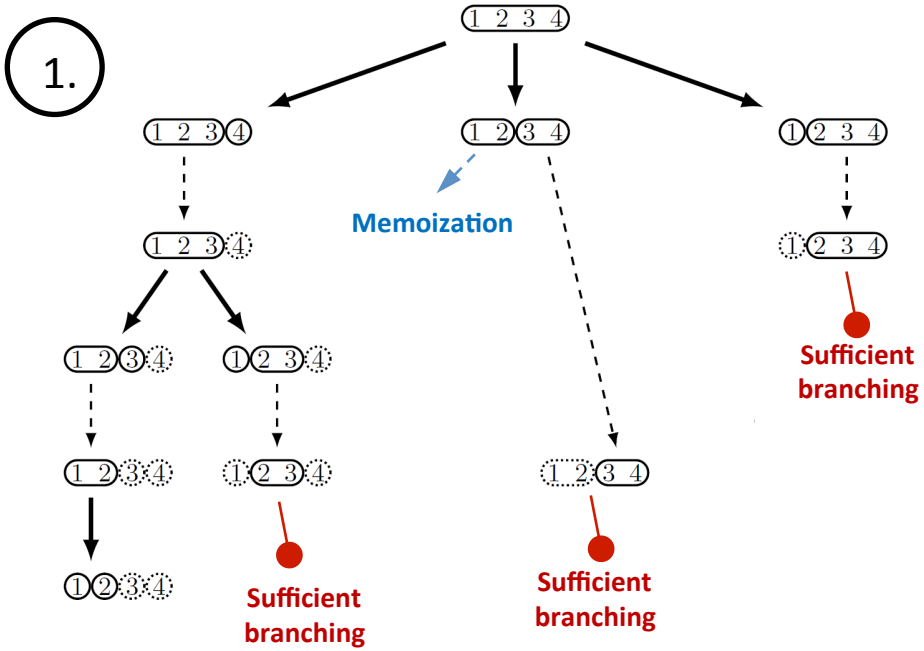
```
procedure SOLVEHSPP(node)
    if node has no child then
        node.optimalCost ← node.cost
        node.optimalCut ← true
    else
        MCost ← node.cost
        μCost ← 0
        for each child of node do
            SOLVEHSPP(child)
            μCost ← μCost + child.optimalCost
        node.optimalCost ← max(μCost, MCost)
        node.optimalCut ← (μCost < MCost)
```

Linear Complexity



Application to the Ordered SPP

1.



2.

Data Structure

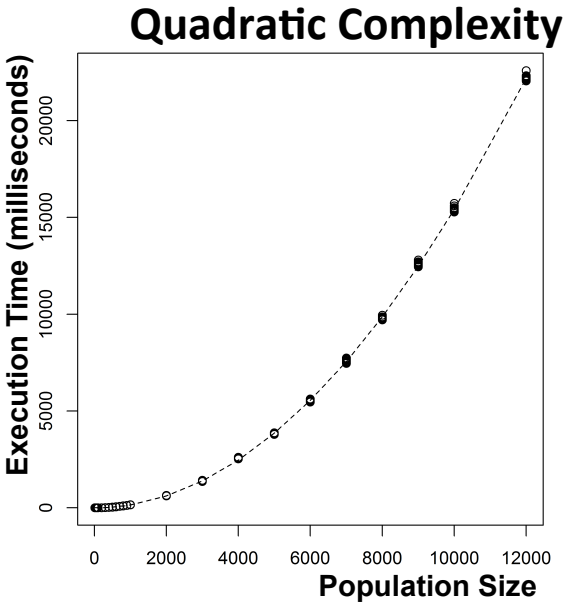
- Set of parts: triangular matrix
- Optimal partition: array of cuts
- Algorithm: dynamic programming

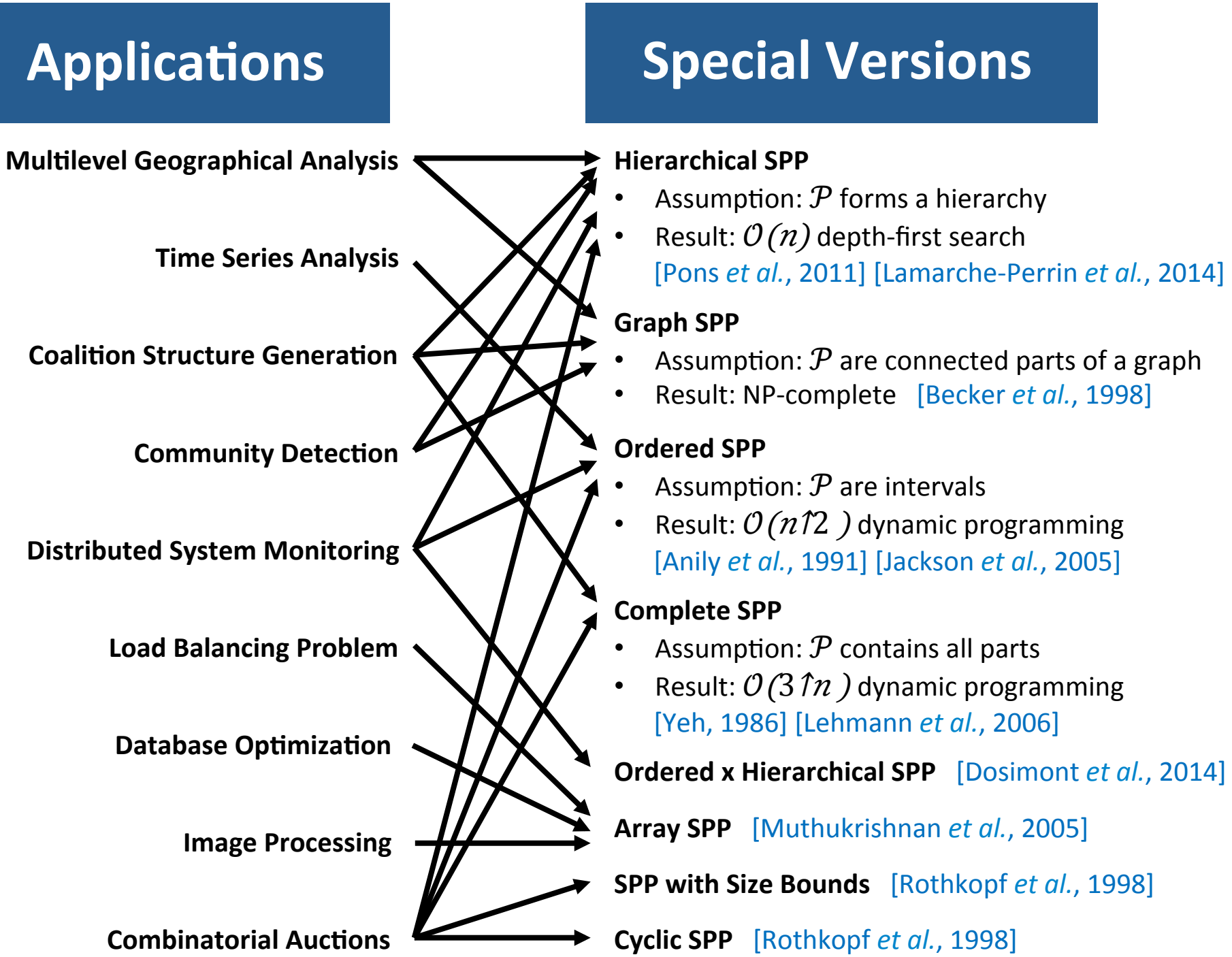
3.

Algorithm 2 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
```




Applications
Special Versions
Multilevel Geographical Analysis
Time Series Analysis
Coalition Structure Generation
Community Detection
Distributed System Monitoring
Load Balancing Problem
Database Optimization
Image Processing
Combinatorial Auctions
Hierarchical SPP
• Assumption: $\mathcal{P}$ forms a hierarchy
• Result: $\mathcal{O}(n)$ depth-first search
[Pons et al., 2011] [Lamarche-Perrin et al., 2014]
Graph SPP
• Assumption: $\mathcal{P}$ are connected parts of a graph
• Result: NP-complete [Becker et al., 1998]
Ordered SPP
• Assumption: $\mathcal{P}$ are intervals
• Result: $\mathcal{O}(n^2)$ dynamic programming
[Anily et al., 1991] [Jackson et al., 2005]
Complete SPP
• Assumption: $\mathcal{P}$ contains all parts
• Result: $\mathcal{O}(3^n)$ dynamic programming
[Yeh, 1986] [Lehmann et al., 2006]
Ordered x Hierarchical SPP [Dosimont et al., 2014]
Array SPP [Muthukrishnan et al., 2005]
SPP with Size Bounds [Rothkopf et al., 1998]
Cyclic SPP [Rothkopf et al., 1998]

Lamarche-Perrin Approach

To characterize the aggregation process

→ The algebra of possible partitions

To preserve the system's semantics

→ A constrained partitioning method

To aggregate according to several dimension

→ Some constraints expressing the system's topology

To evaluate and compare the representations

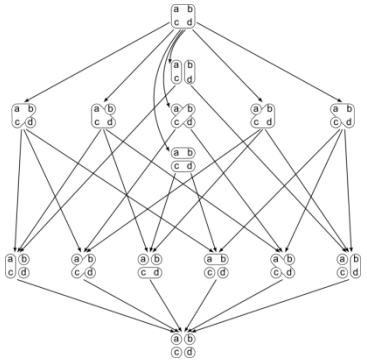
→ Some measures of complexity and information

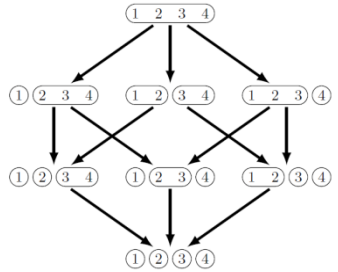
To offer several granularity levels

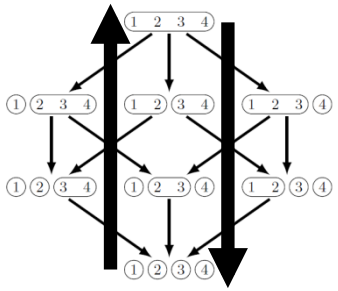
→ The optimization of a compromise

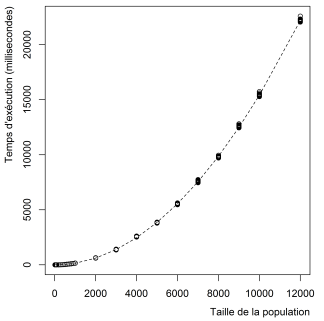
To compute the best representations

→ A generic algorithm of constrained optimization

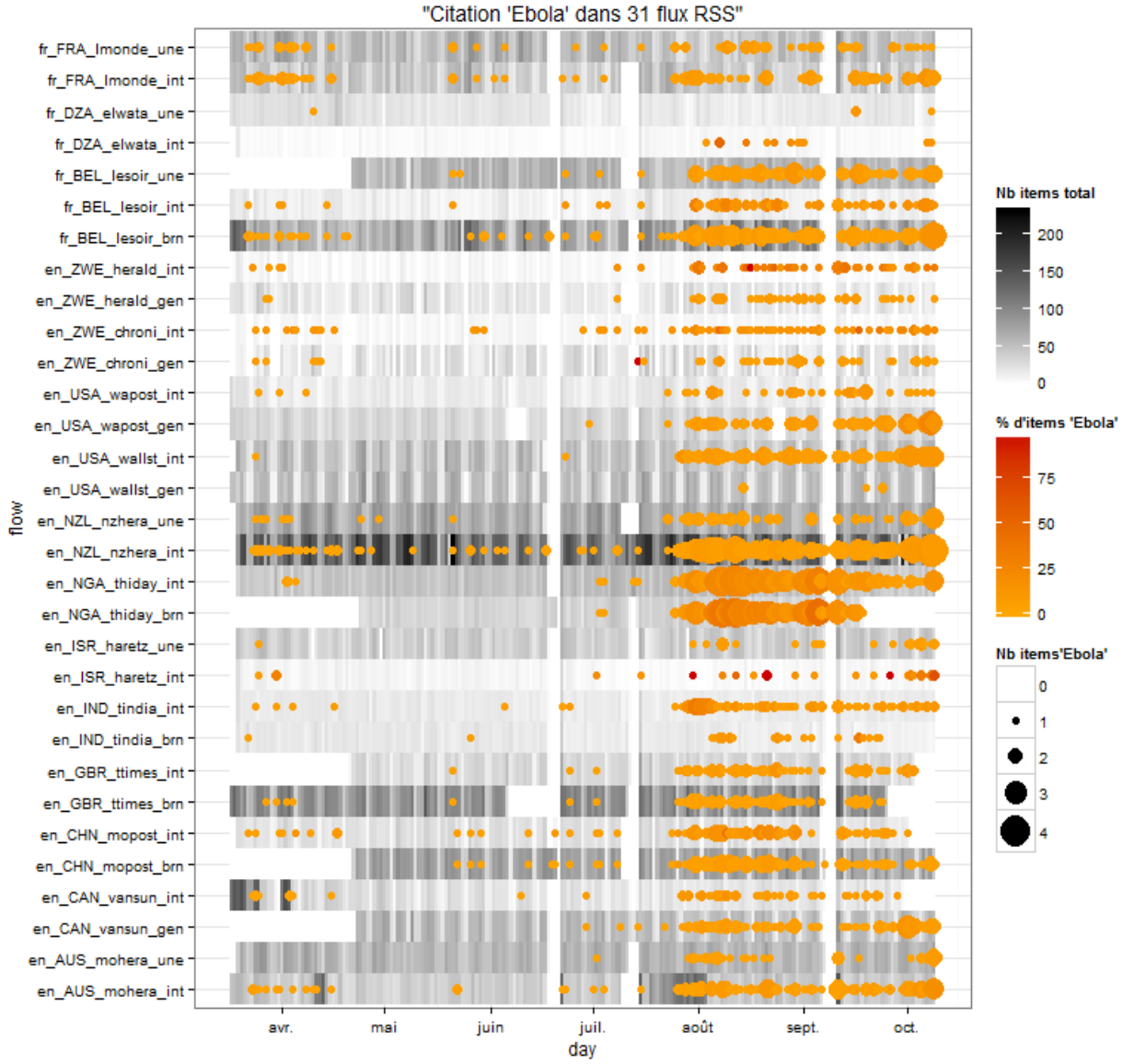




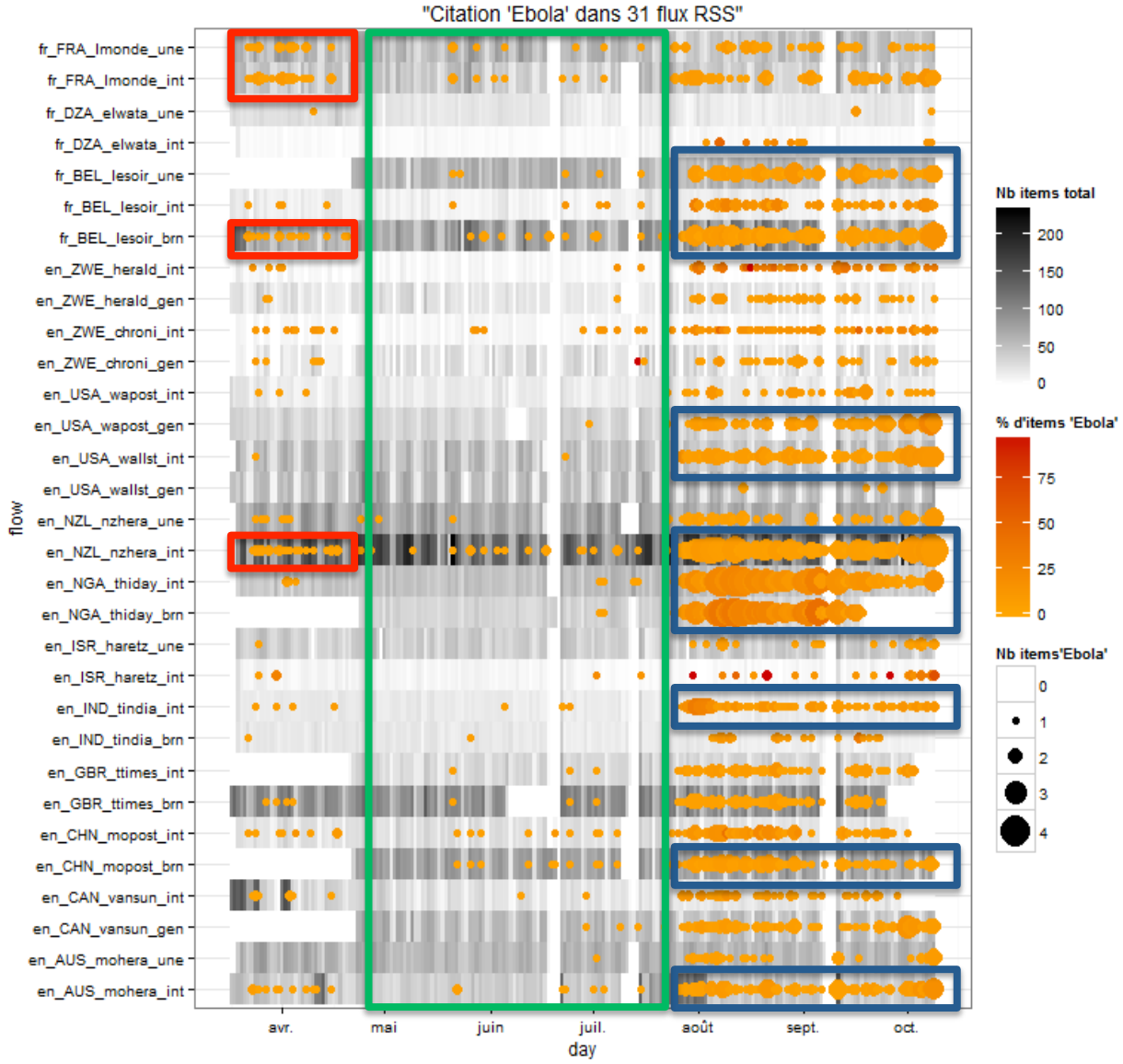




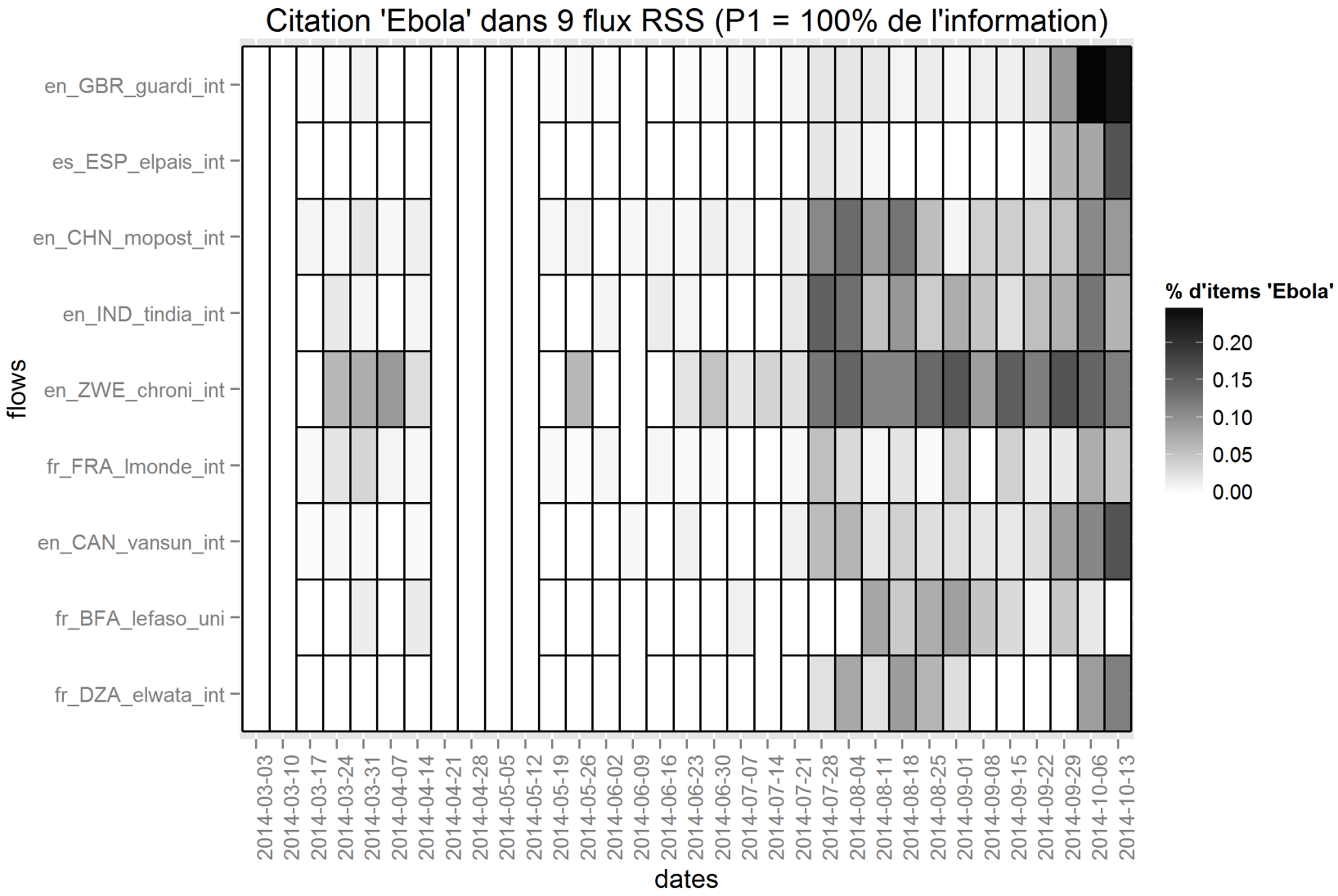
Media Aggregation



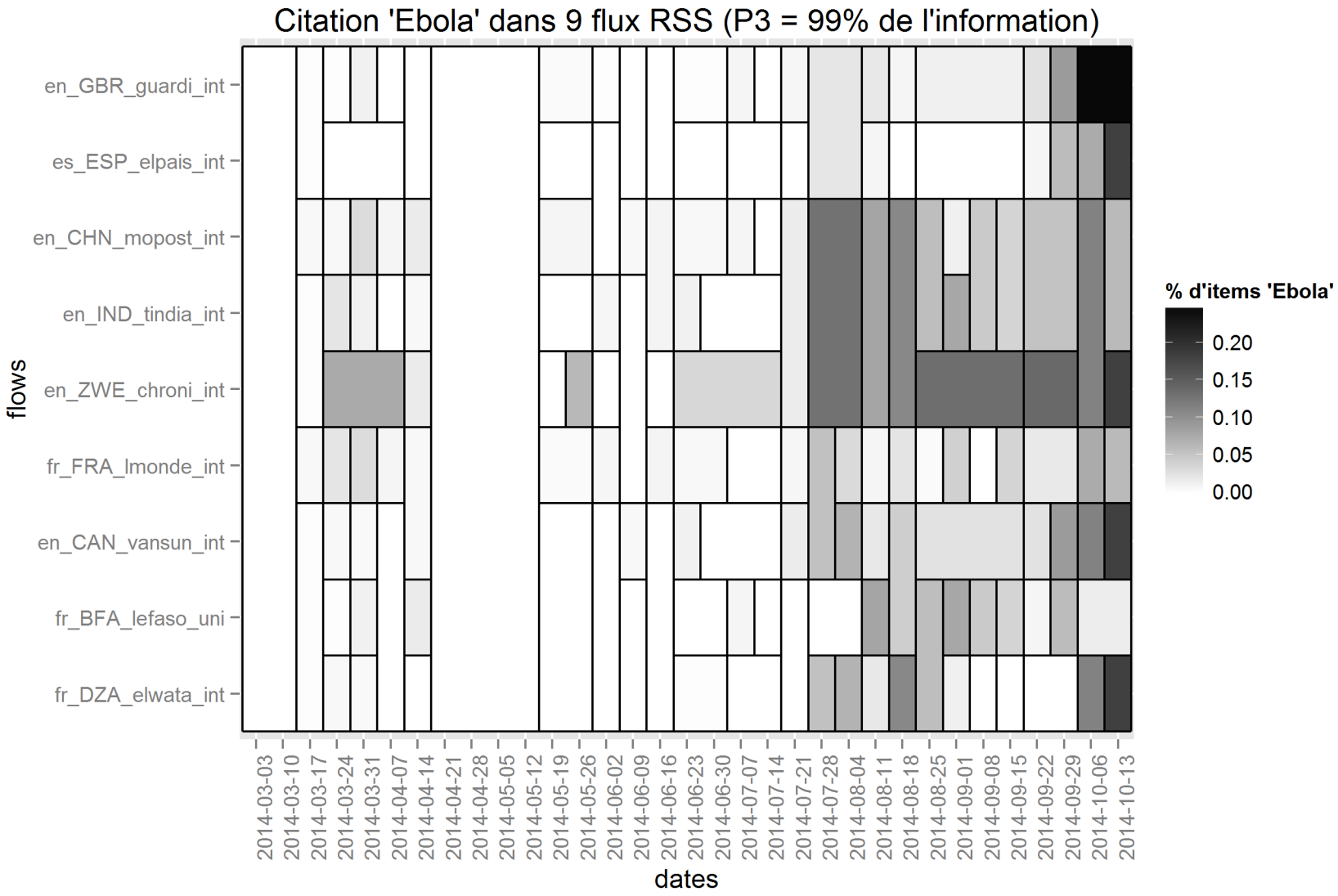
Media Aggregation



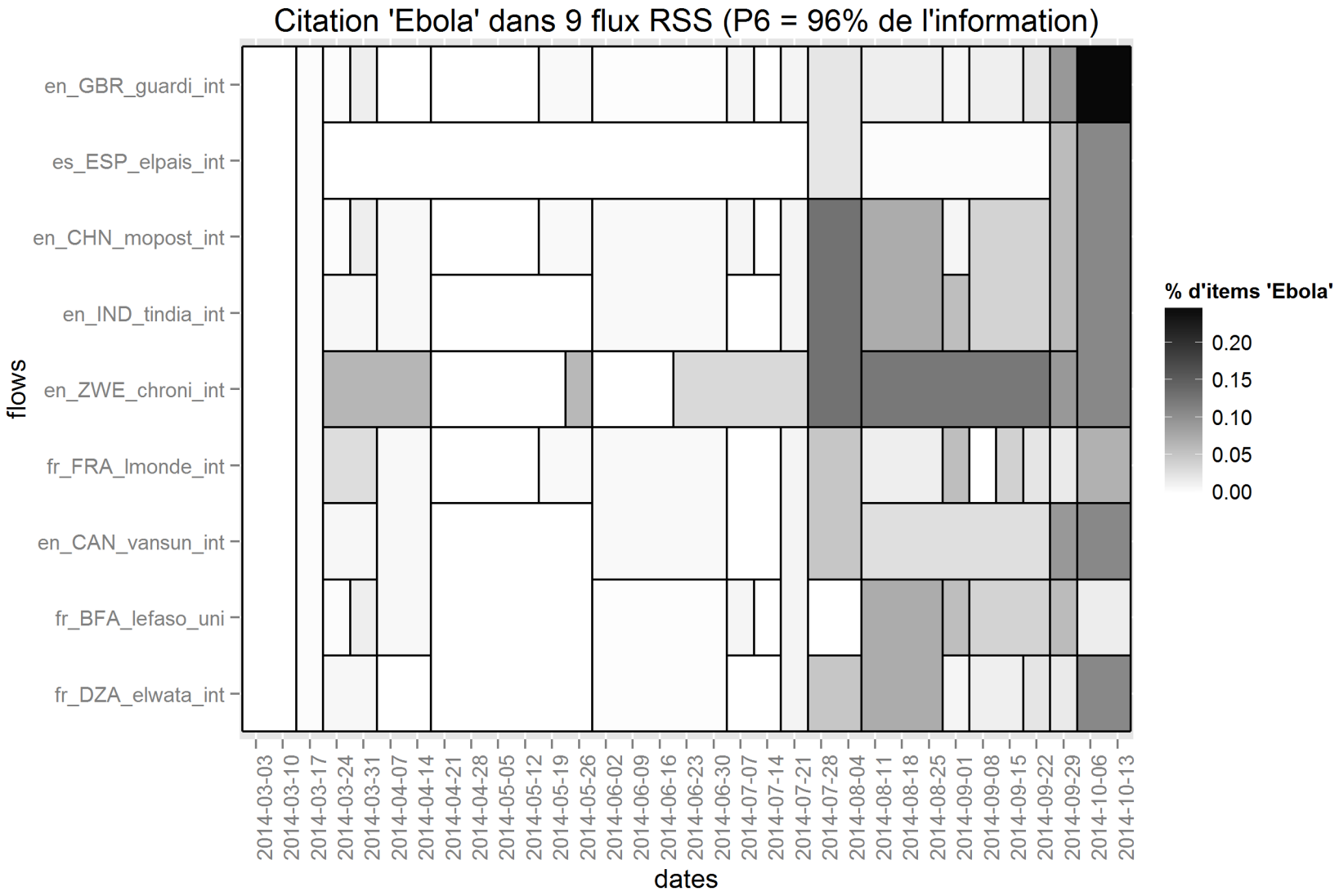
Media Aggregation



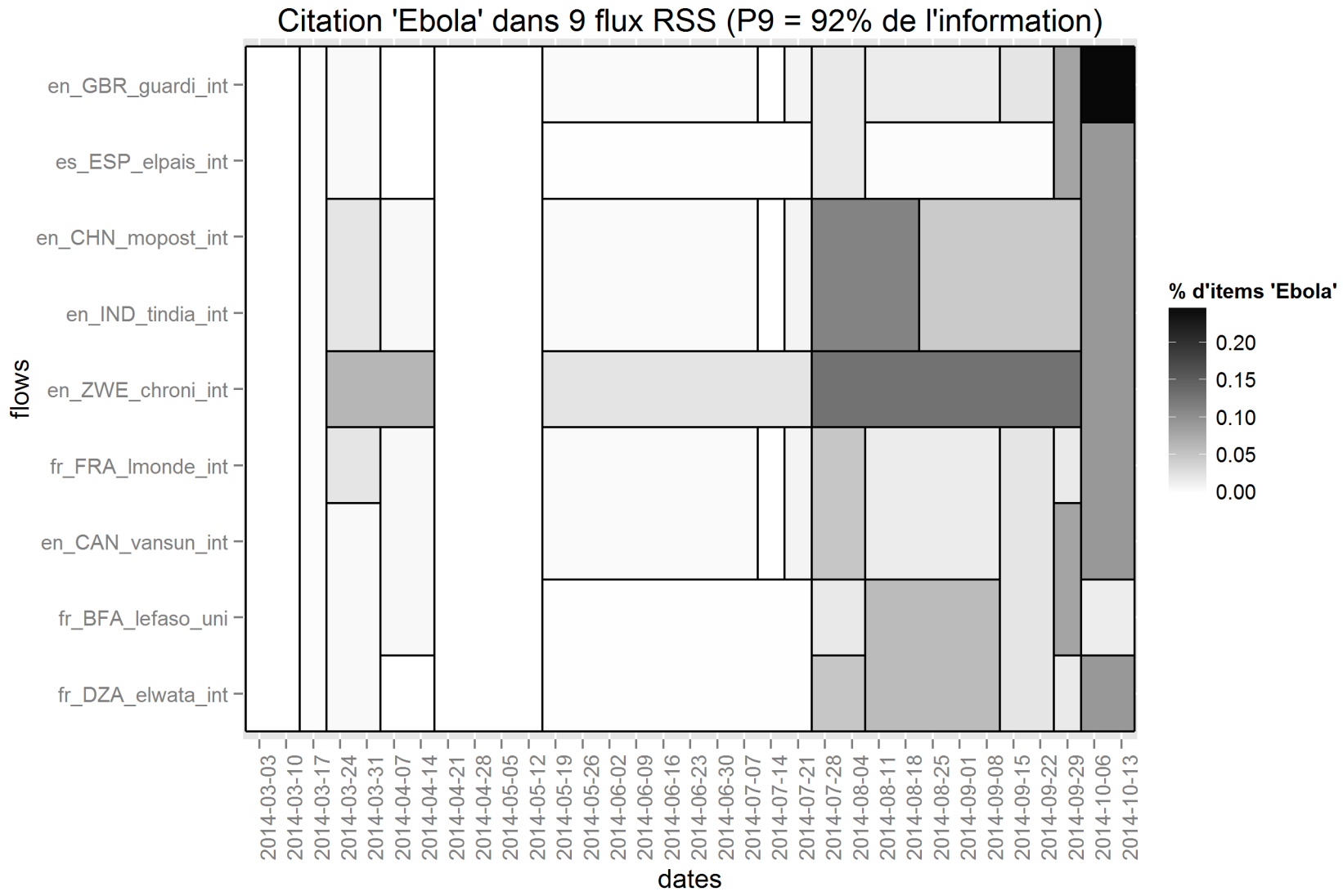
Media Aggregation



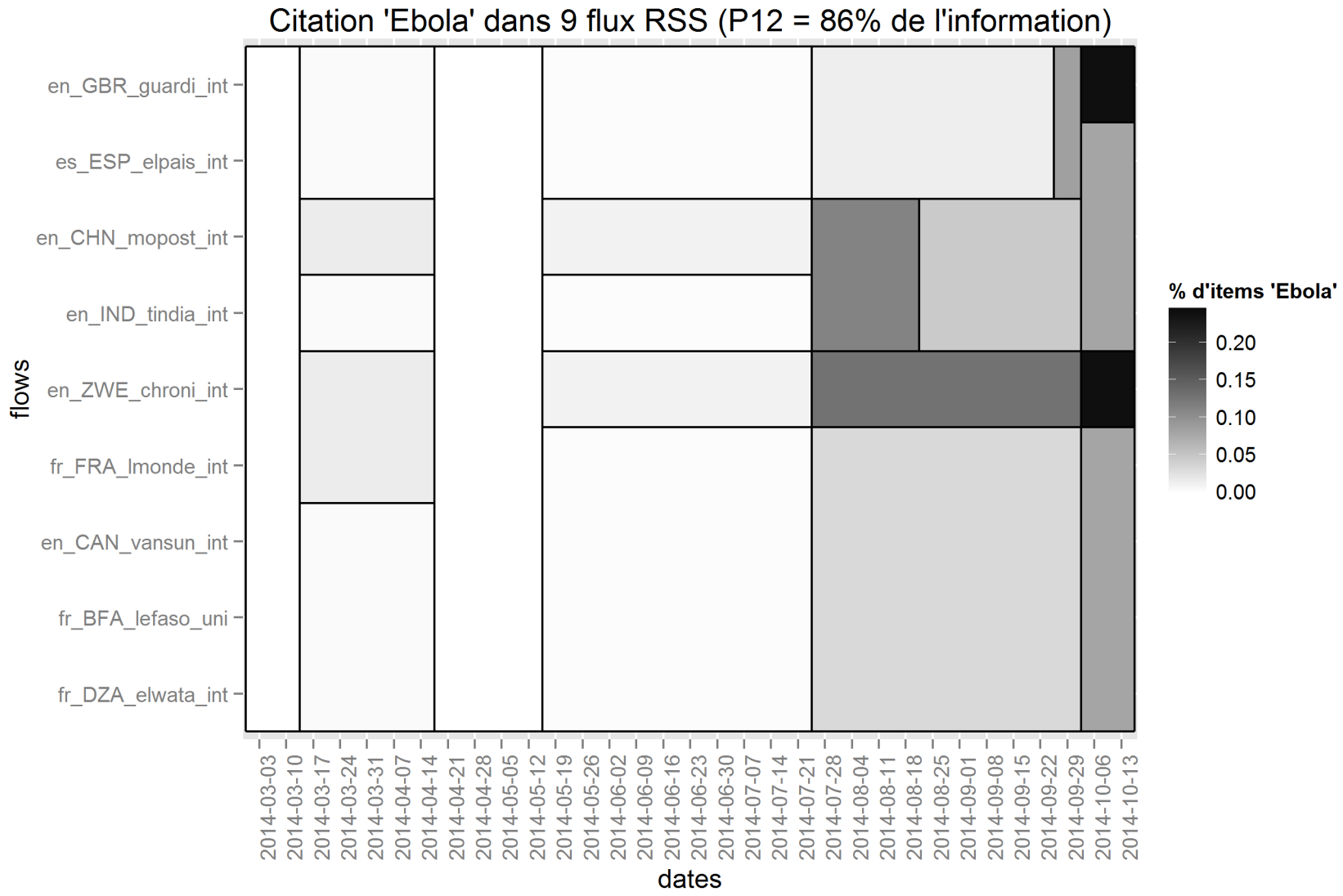
Media Aggregation



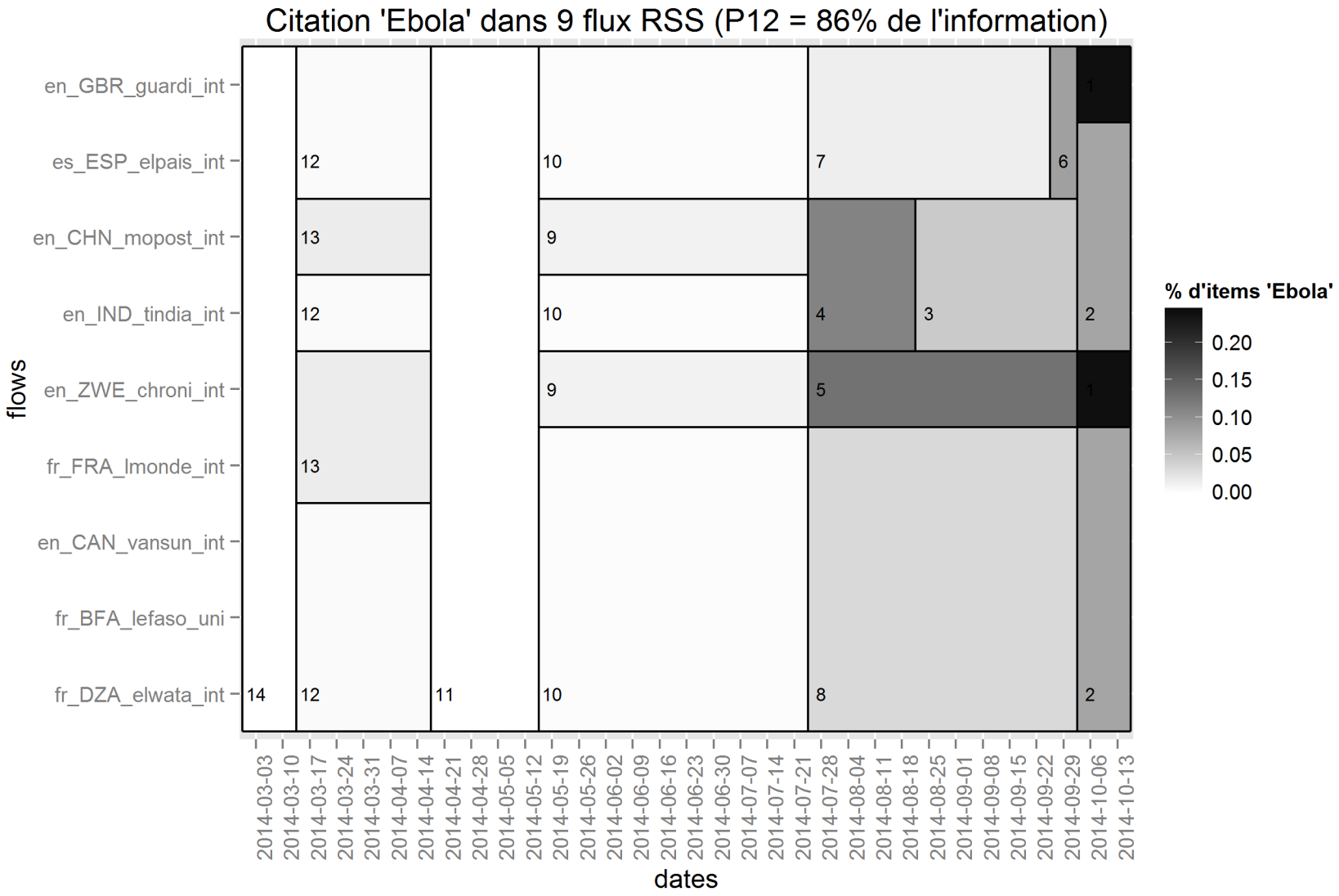
Media Aggregation



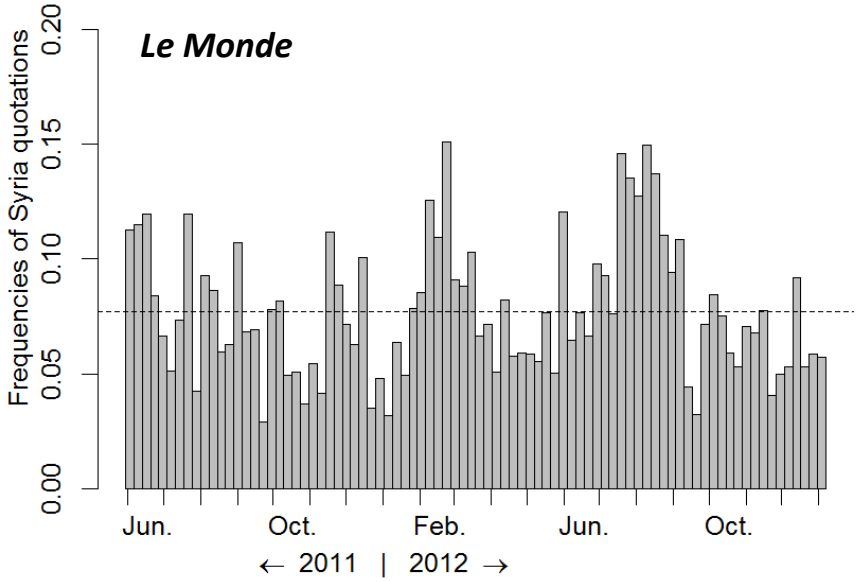
Media Aggregation

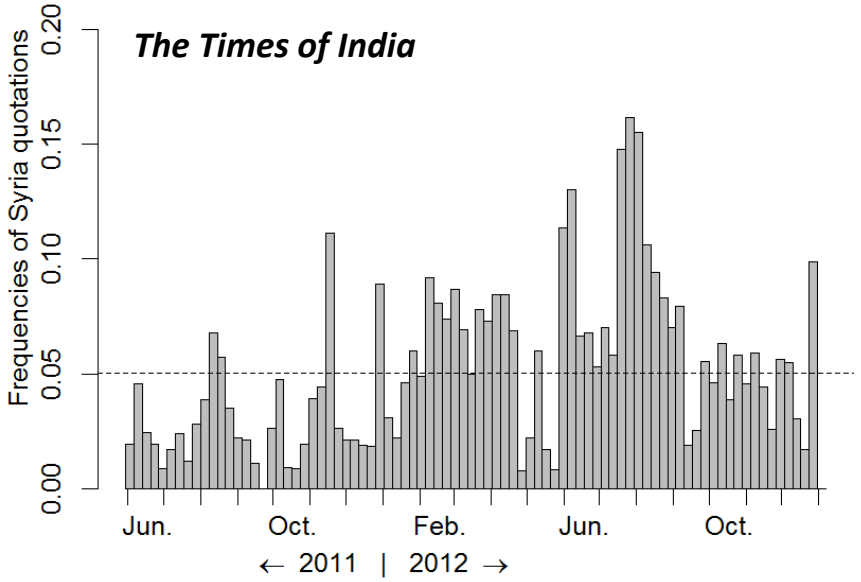


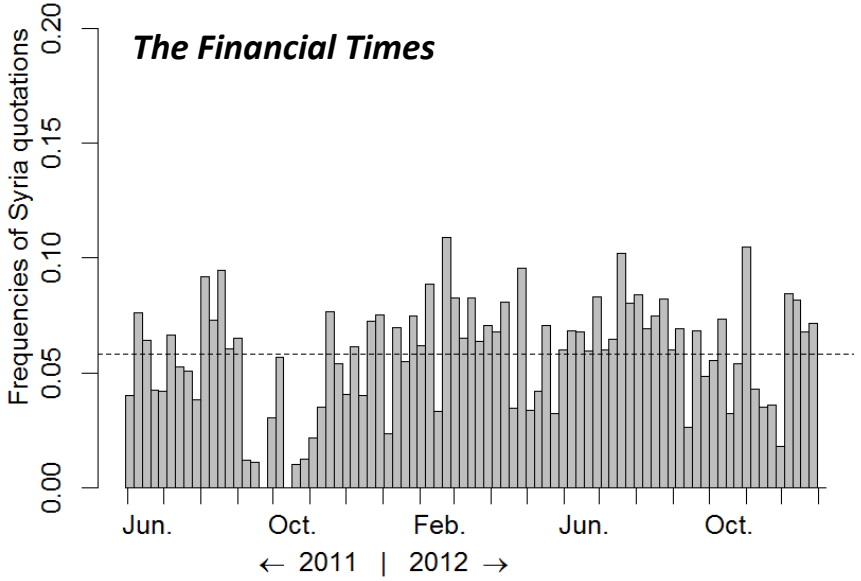
Media Aggregation

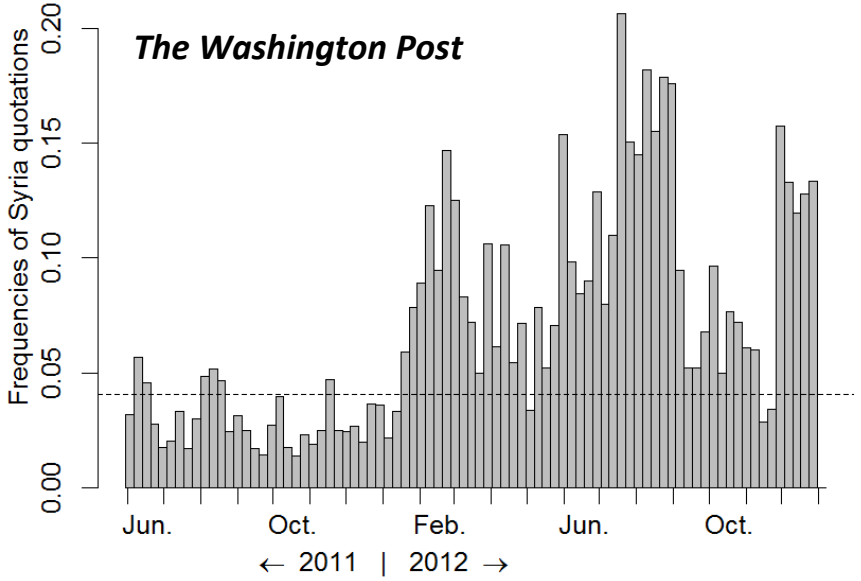


Information Loss = 0%

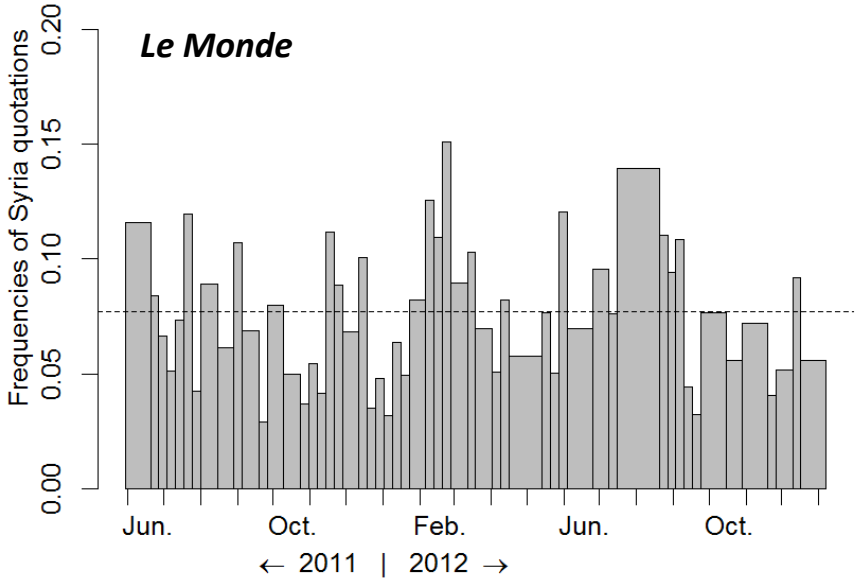


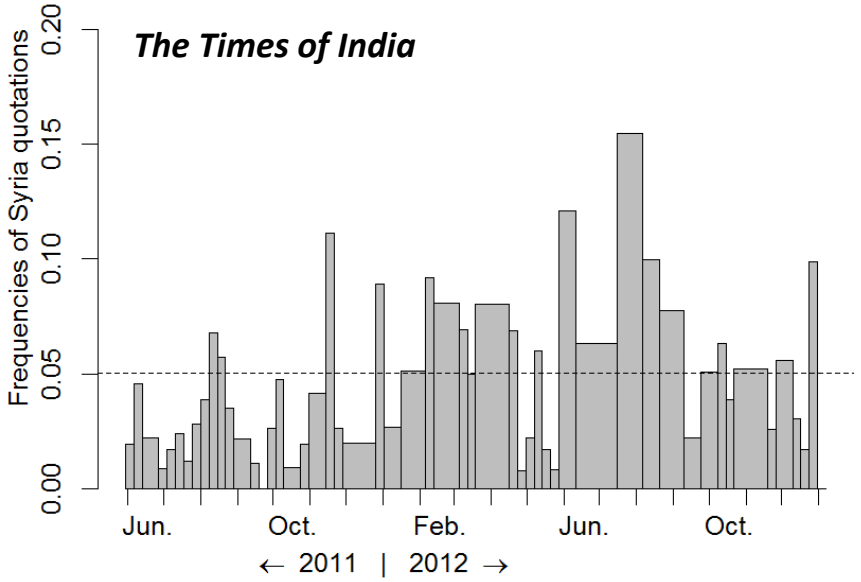


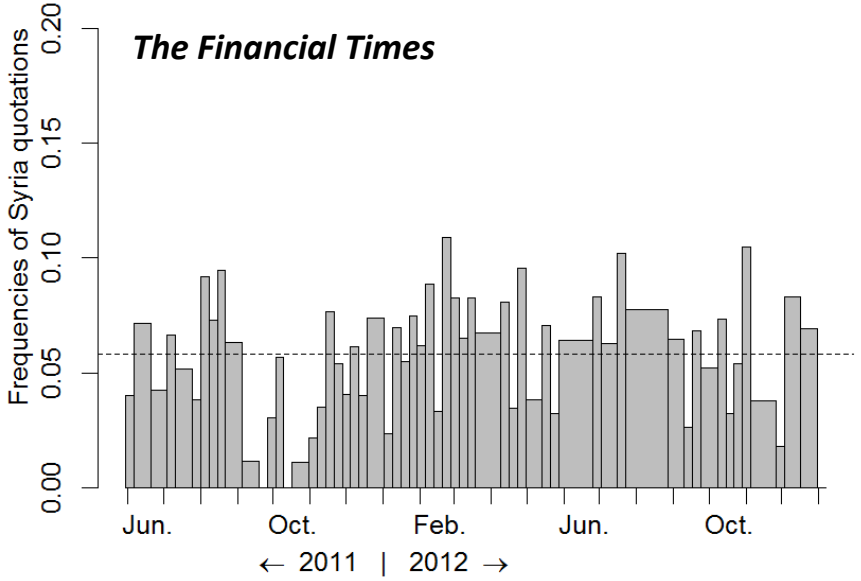


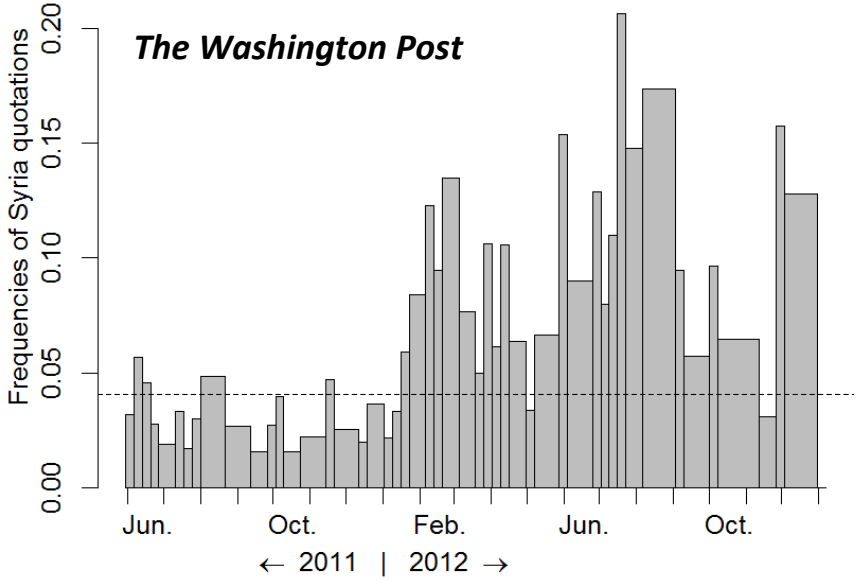


Information Loss = 1%

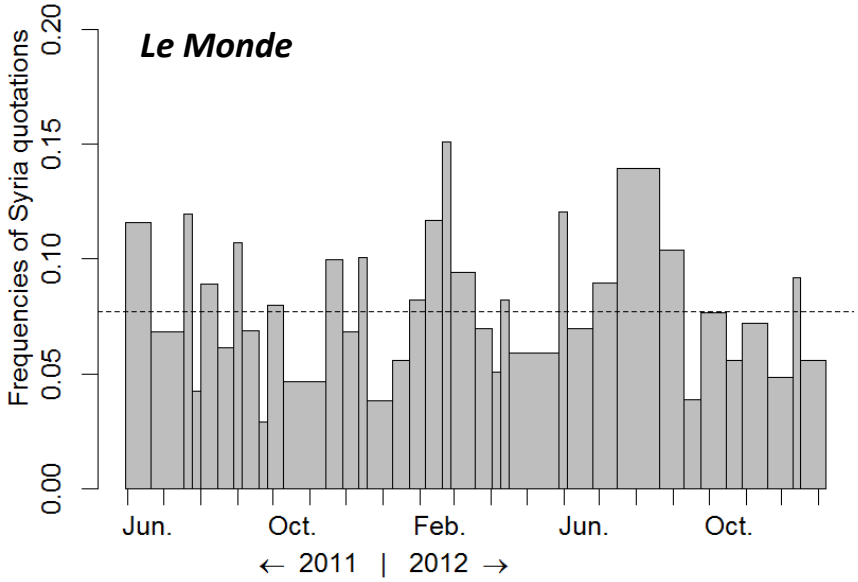


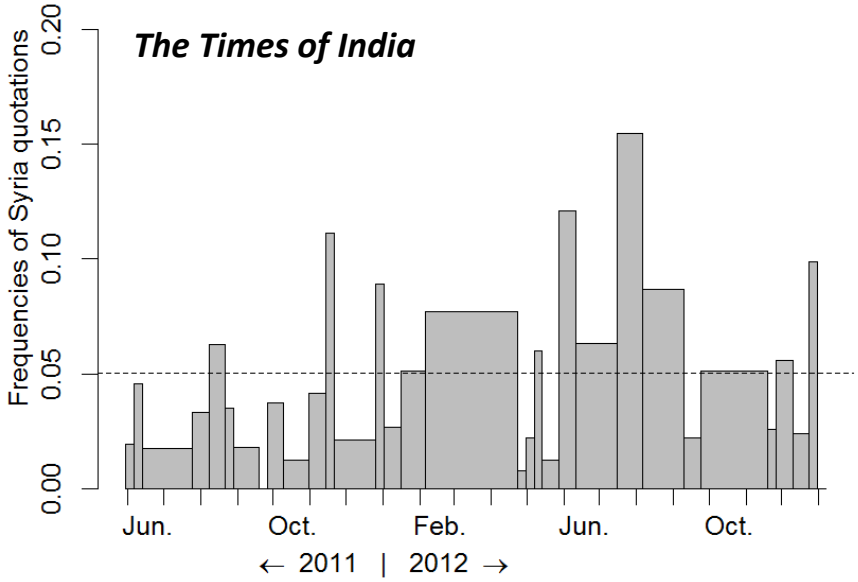


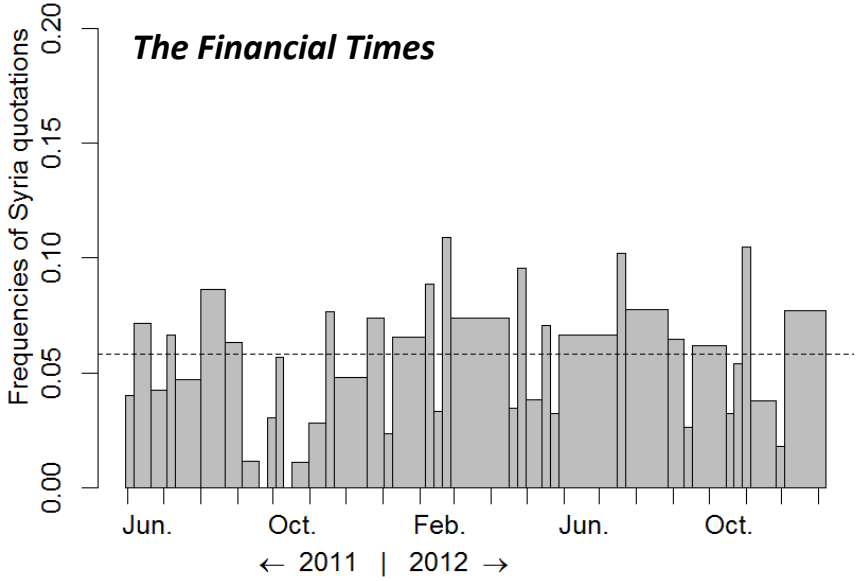


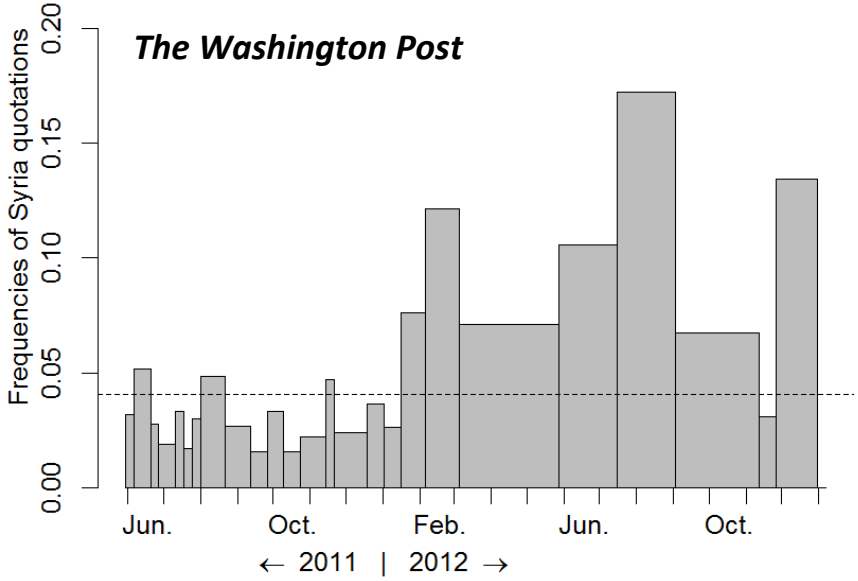


Information Loss = 5%

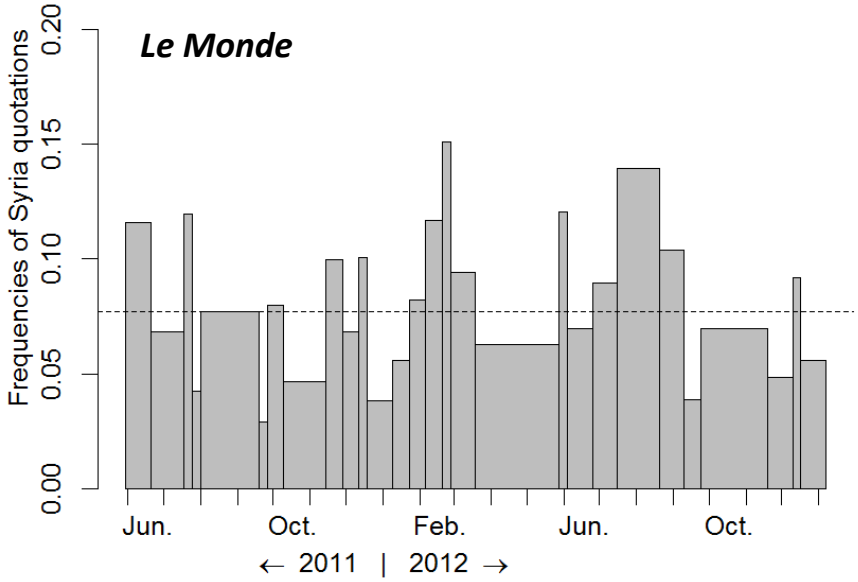


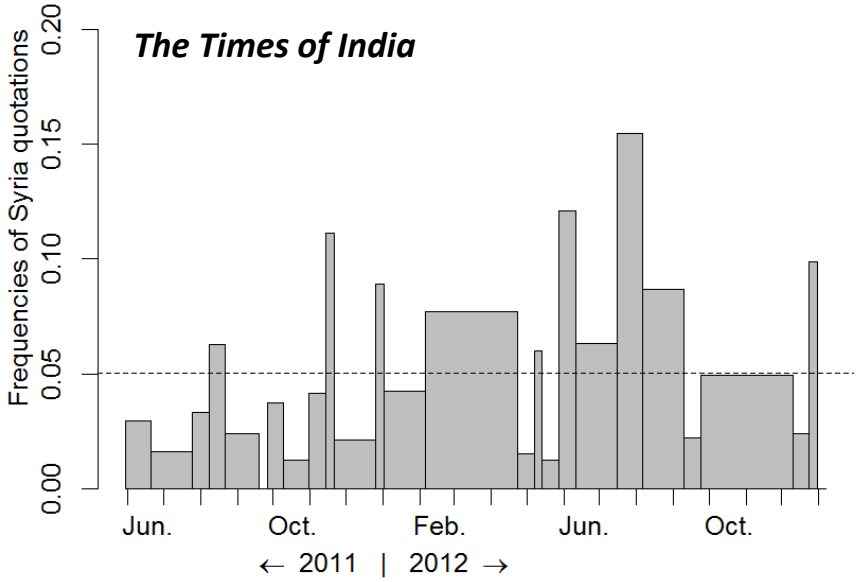


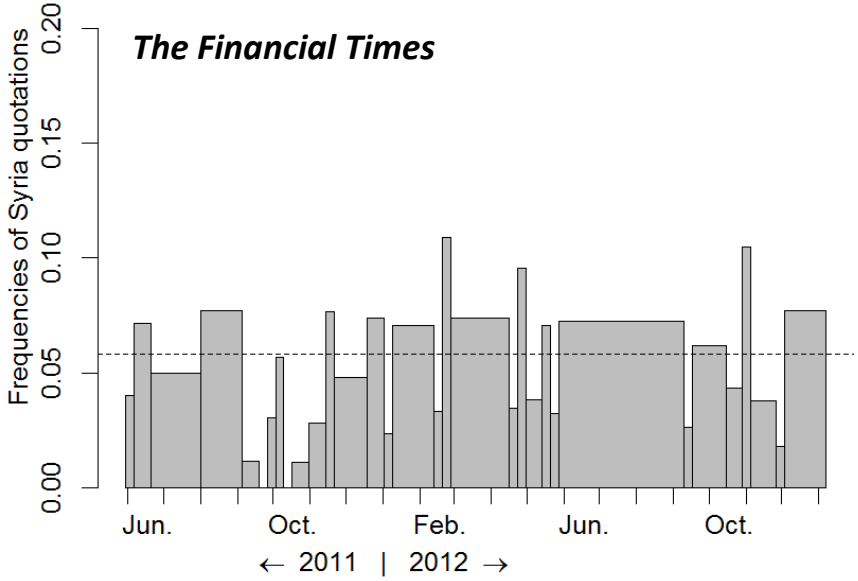


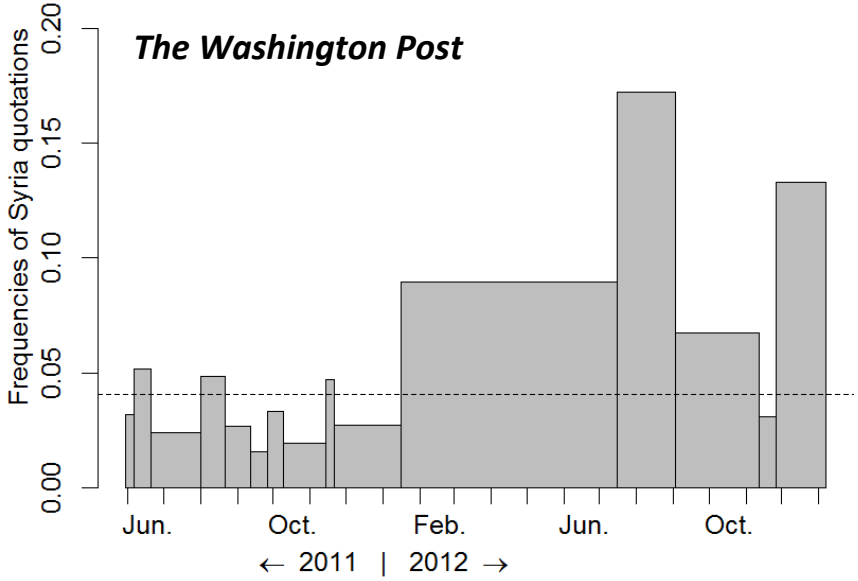


Information Loss = 10%

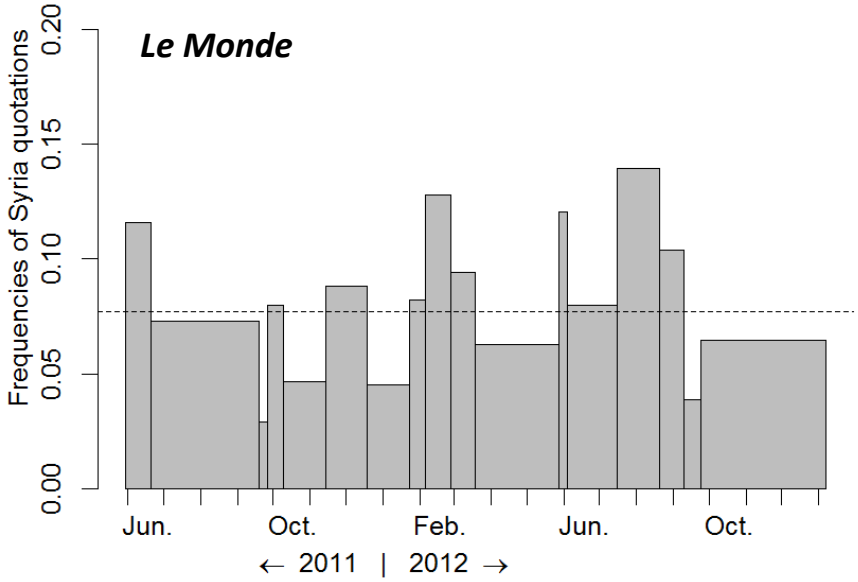


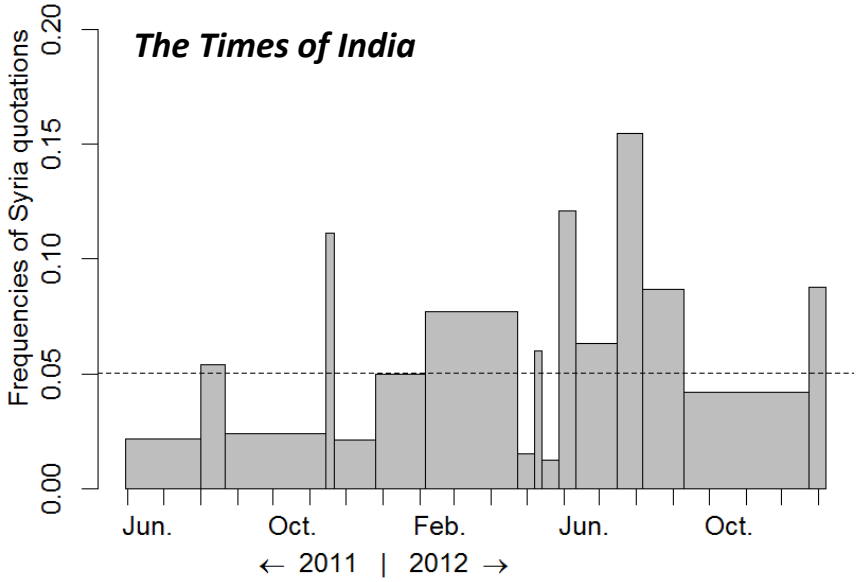


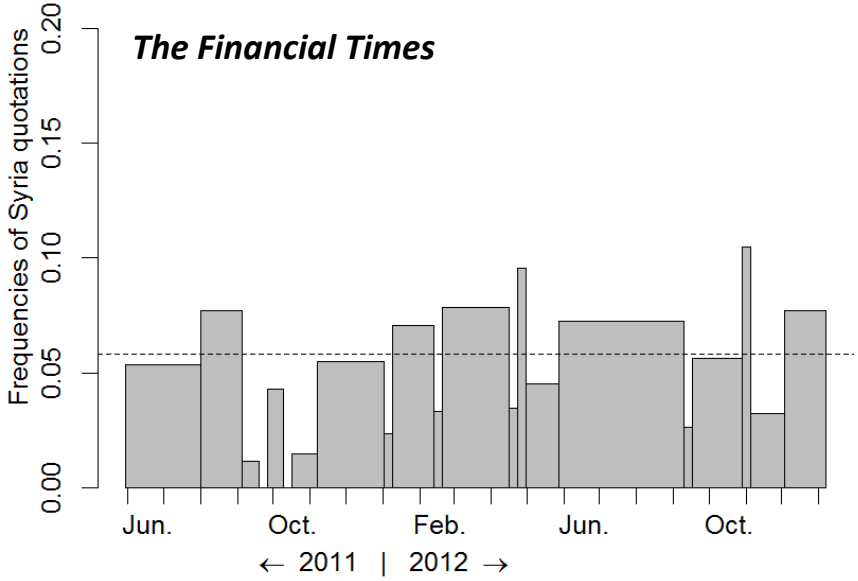


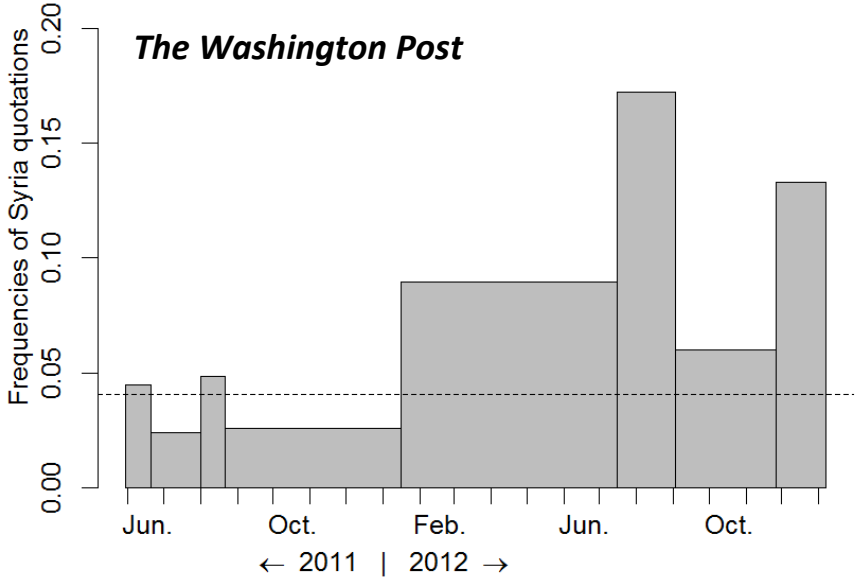


Information Loss = 20%

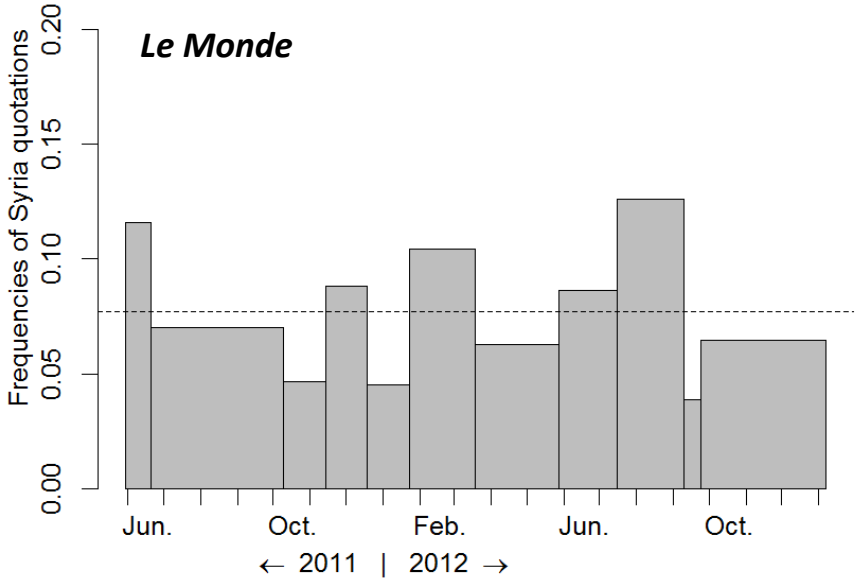


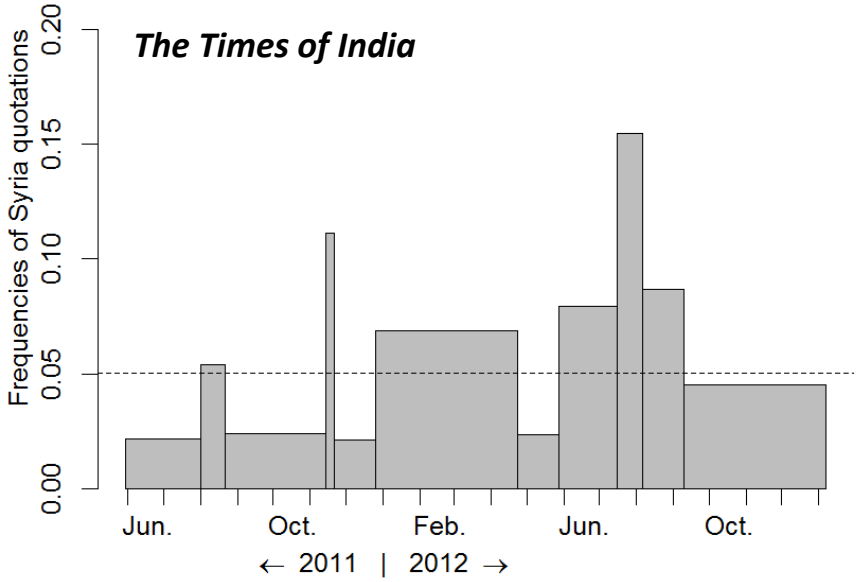


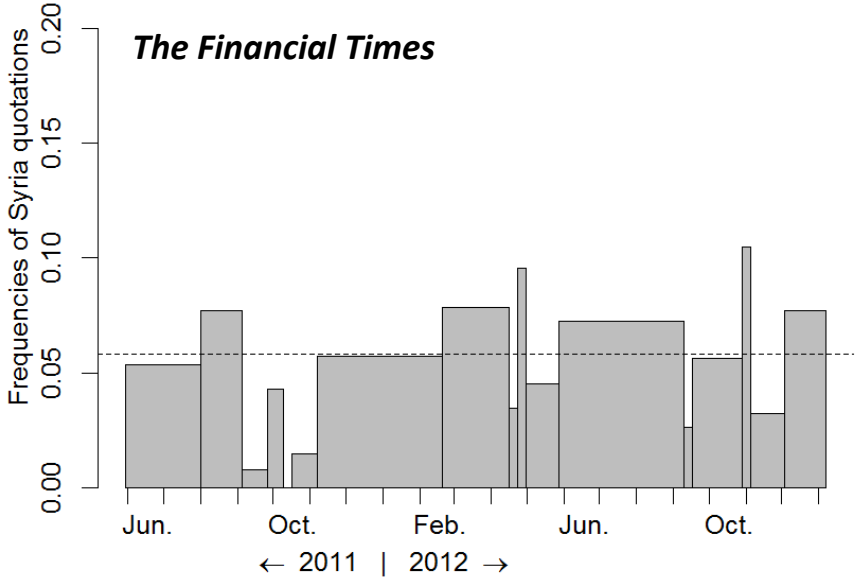


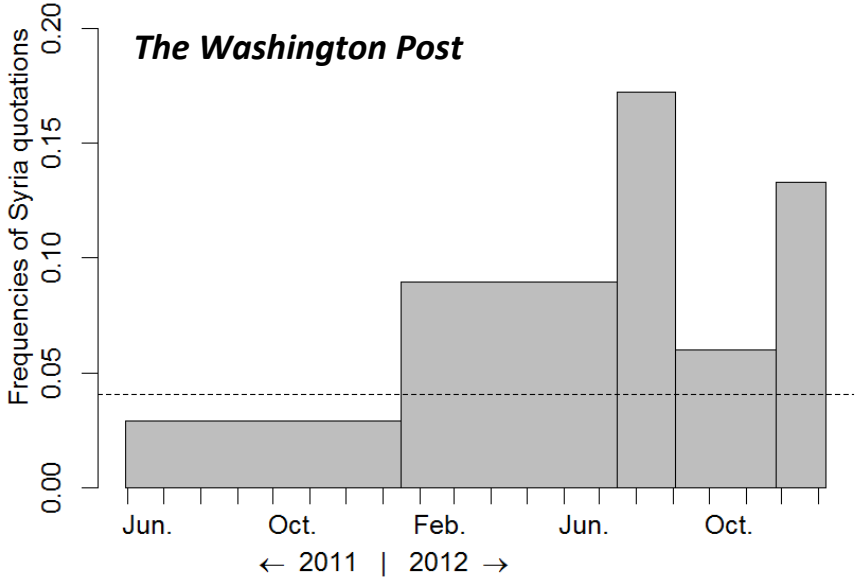


Information Loss = 30%

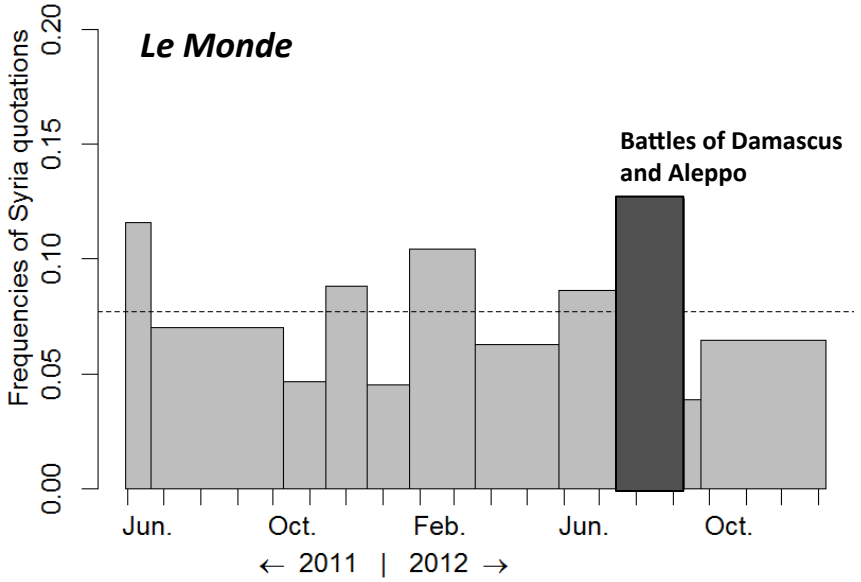


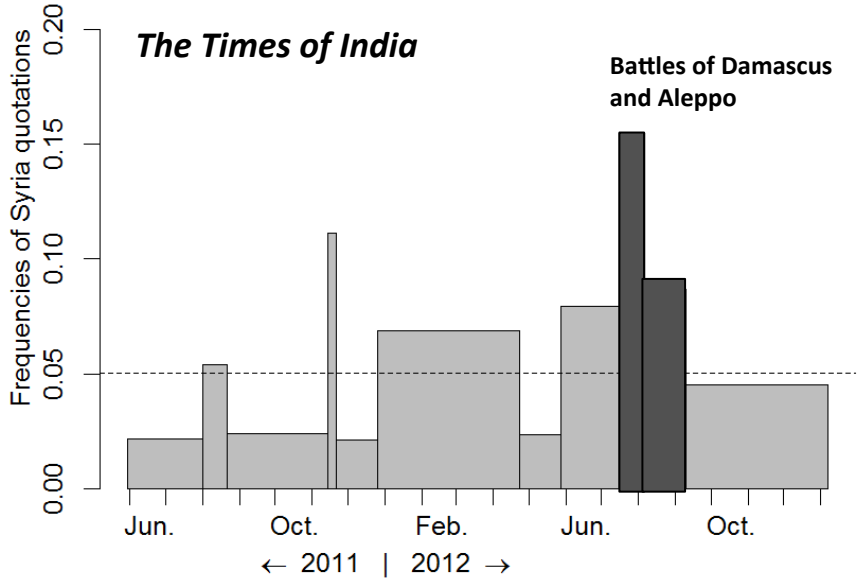


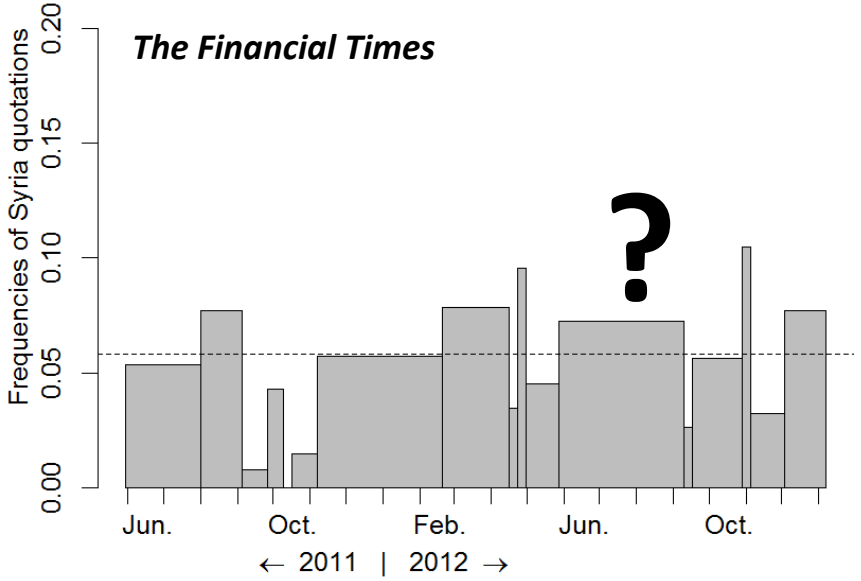


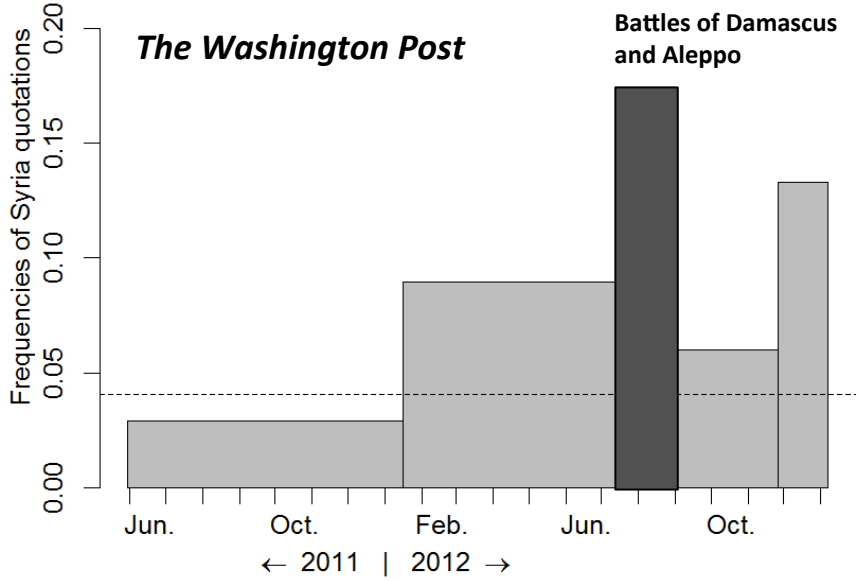


Information Loss = 30%









Information Loss = 30%