Characterizing measures for the assessment of cluster analysis and community detection

Nejat Arinik
Vincent Labatut
Rosa Figueiredo

Laboratoire Informatique d'Avignon, University of Avignon, France.

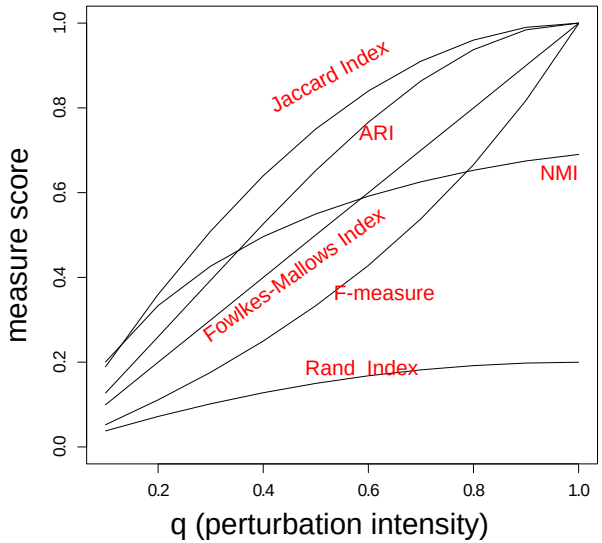
MARAMI
Oct 15, 2020

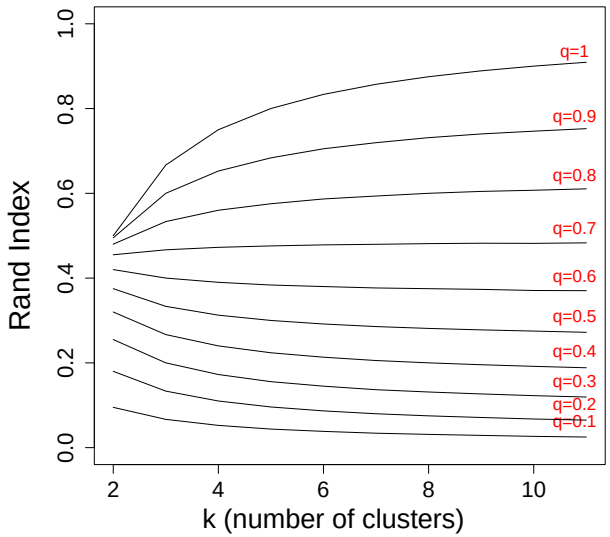
Context

- **Context:** comparing two non-overlapping partitions through an external measure $\longrightarrow$ cluster analysis, graph partitioning
  - ground-truth vs. estimated partition, 2 estimated partitions

Context

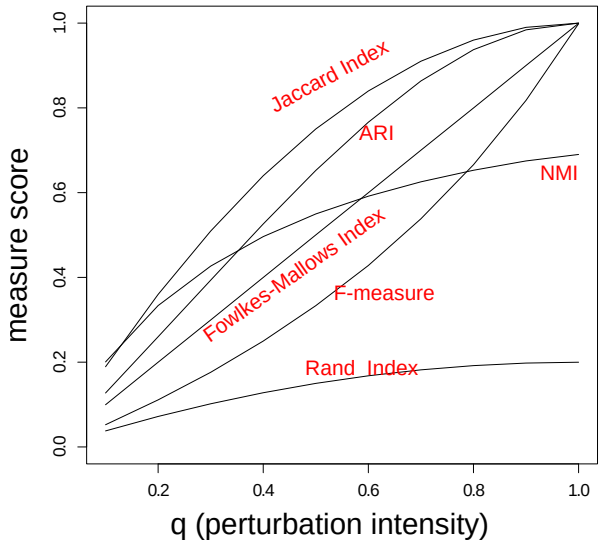
- **Context:** comparing two non-overlapping partitions through an external measure ⟶ cluster analysis, graph partitioning
  - ground-truth vs. estimated partition, 2 estimated partitions
- **Issues:**
  - profusion of available measures ⟶ trend to follow popular measures
  - lack of comprehensive comparison

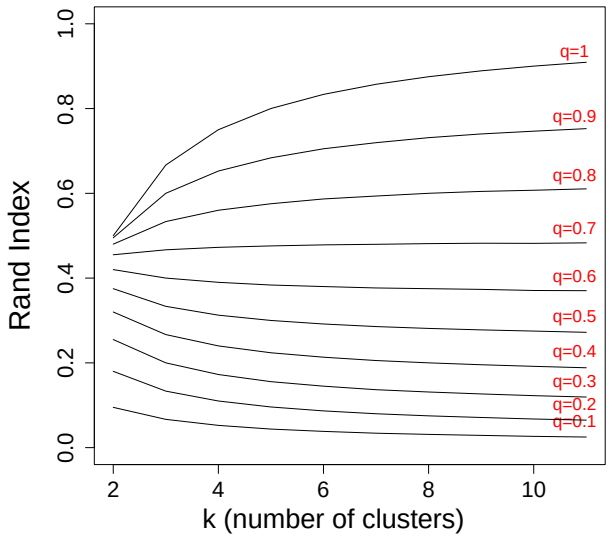




Context

- **Context:** comparing two non-overlapping partitions through an external measure ⟶ cluster analysis, graph partitioning
  - ground-truth vs. estimated partition, 2 estimated partitions
- **Issues:**
  - profusion of available measures ⟶ trend to follow popular measures
  - lack of comprehensive comparison

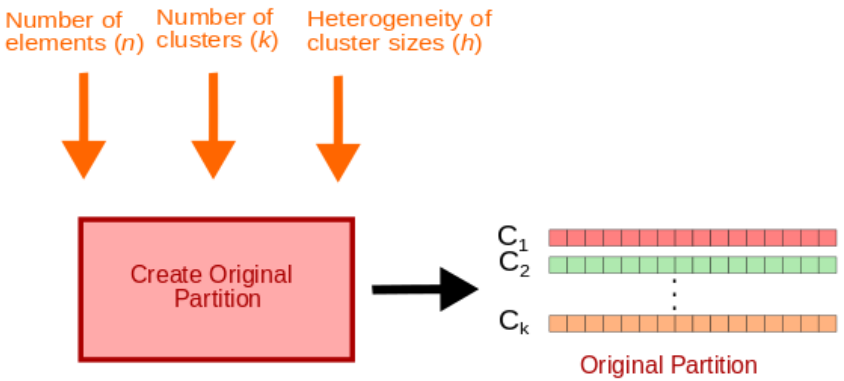




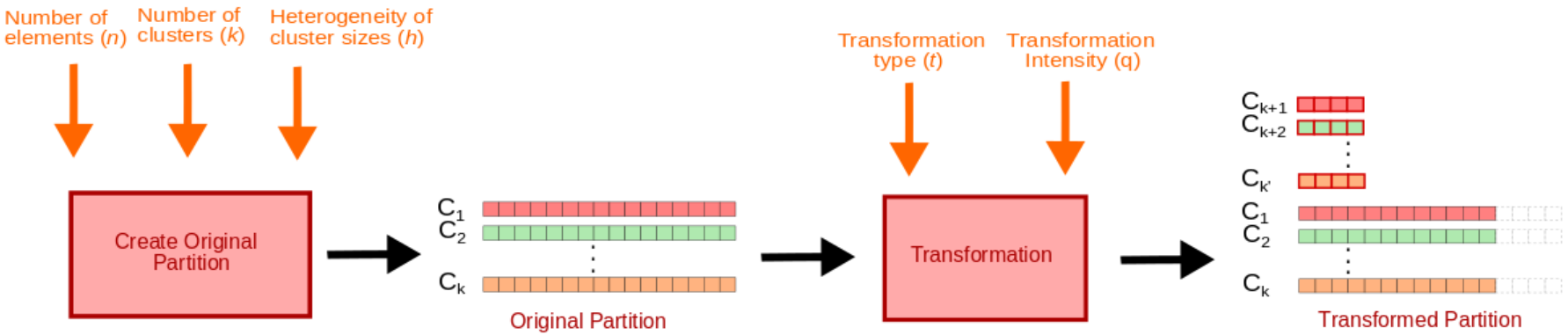
- **Our solution:** a new framework of evaluation

Outline

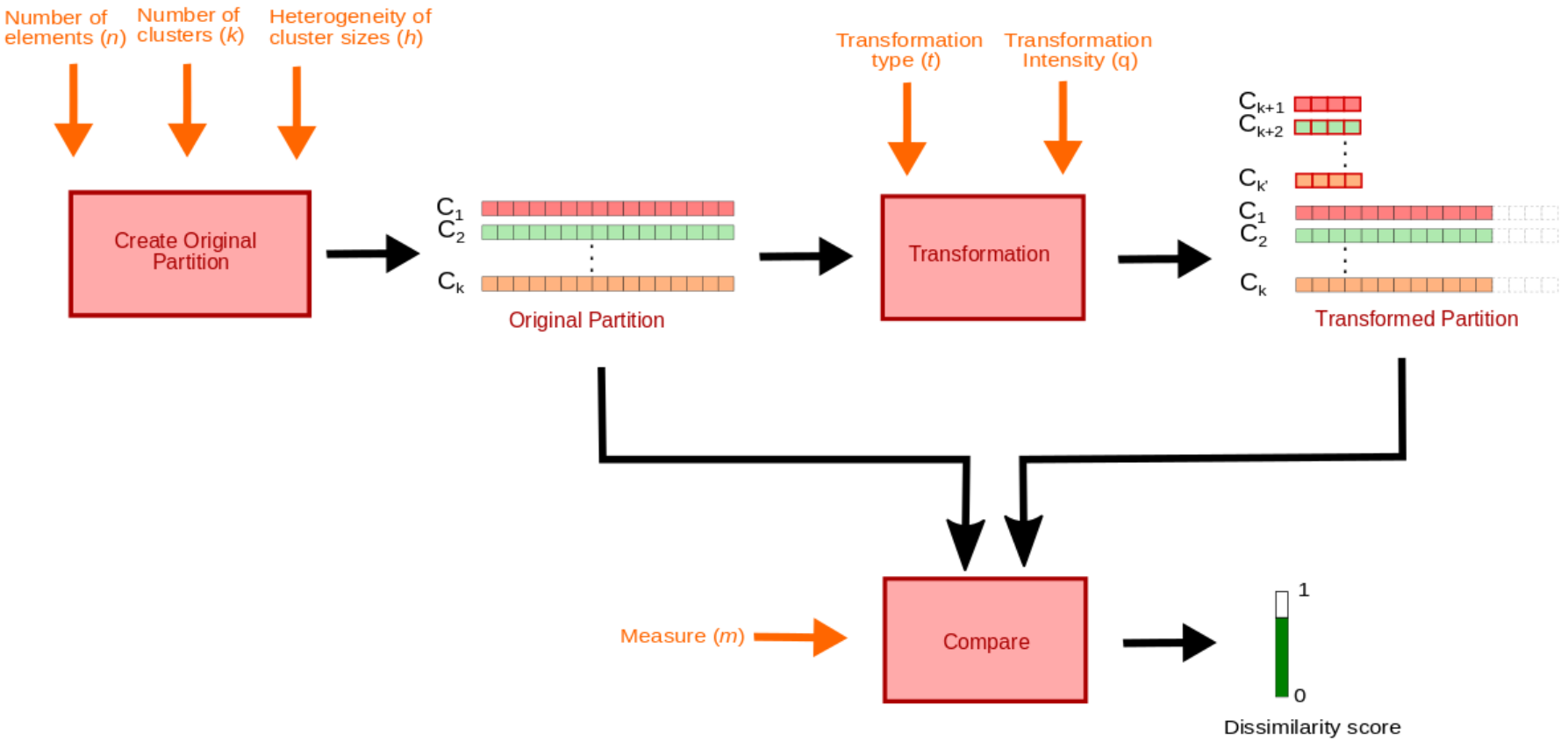
Our framework: Characterizing of the measures



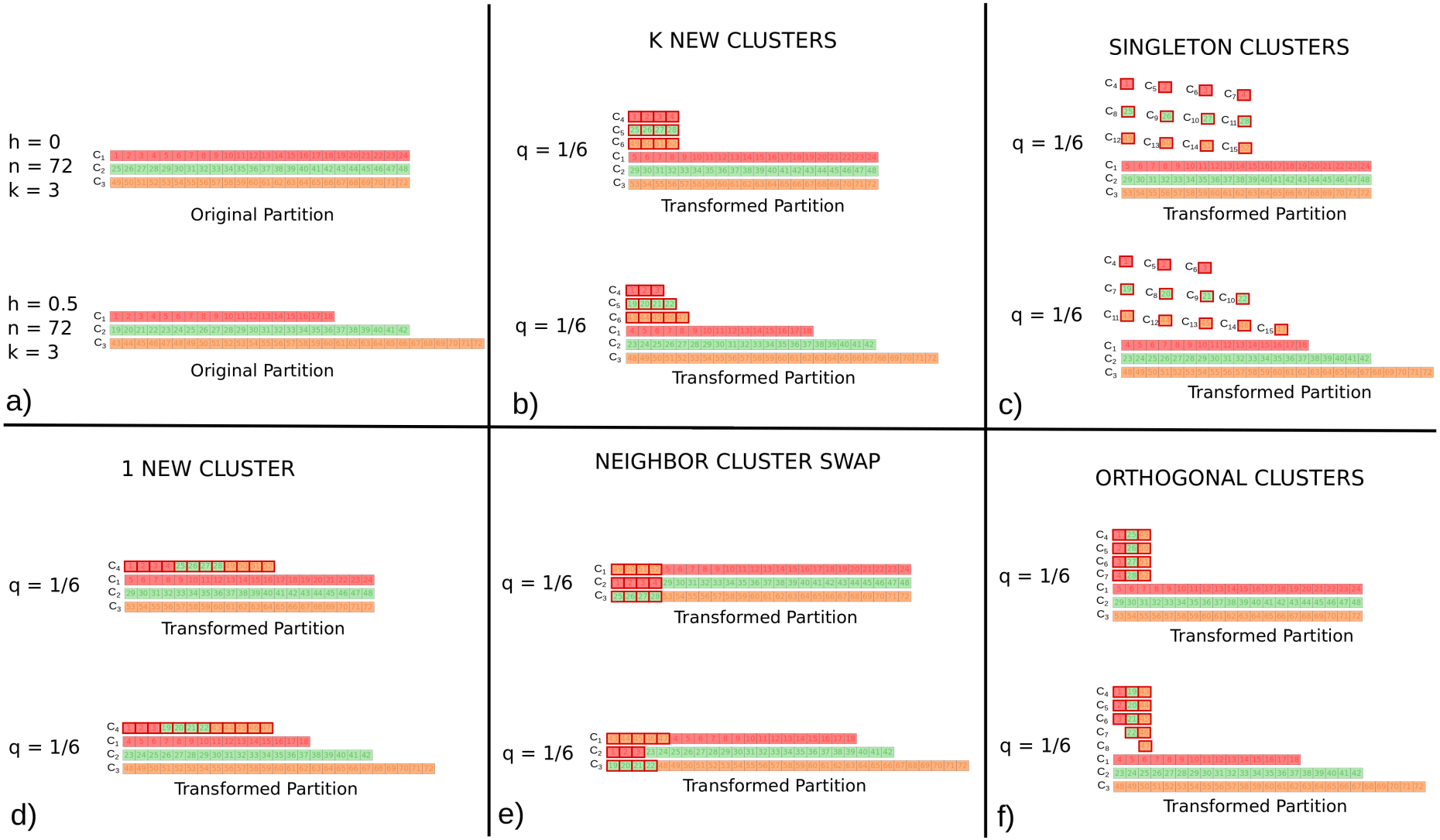
Our framework: Characterizing of the measures



Our framework: Characterizing of the measures



Our framework: proposed deterministic transformations



Our framework: Regression Analysis

Our multiple linear regression model:

$$
\begin{aligned}
y = \sum_i \sum_j \Big(& \beta_{0ij} t_i m_j \\
& + \beta_{1ij} n t_i m_j + \beta_{2ij} k t_i m_j + \beta_{3ij} p t_i m_j + \beta_{4ij} h t_i m_j \\
& + \beta_{5ij} n k t_i m_j + \beta_{6ij} n h t_i m_j + \beta_{7ij} n p t_i m_j + \beta_{8ij} k h t_i m_j + \beta_{9ij} k p t_i m_j + \beta_{10ij} h p t_i m_j \Big) \\
& + \epsilon,
\end{aligned}
\tag{1}
$$

- $\beta_{.ij}$: regression coefficients
- t_i $(1 \leq i \leq T)$ and m_j $(1 \leq j \leq M)$: binary dummy variables, where $T =$ number of transformations and $M =$ number of measures
- ϵ: common error

Our framework: Regression Analysis

Our multiple linear regression model:

$$
\begin{aligned}
y = \sum_i \sum_j \Big(& \beta_{0ij} t_i m_j \\
& + \beta_{1ij} n t_i m_j + \beta_{2ij} k t_i m_j + \beta_{3ij} p t_i m_j + \beta_{4ij} h t_i m_j \\
& + \beta_{5ij} n k t_i m_j + \beta_{6ij} n h t_i m_j + \beta_{7ij} n p t_i m_j + \beta_{8ij} k h t_i m_j + \beta_{9ij} k p t_i m_j + \beta_{10ij} h p t_i m_j \Big) \\
& + \epsilon,
\end{aligned}
\tag{1}
$$

- $\beta_{.ij}$: regression coefficients
- t_i $(1 \leq i \leq T)$ and m_j $(1 \leq j \leq M)$: binary dummy variables, where T = number of transformations and M = number of measures
- ϵ: common error

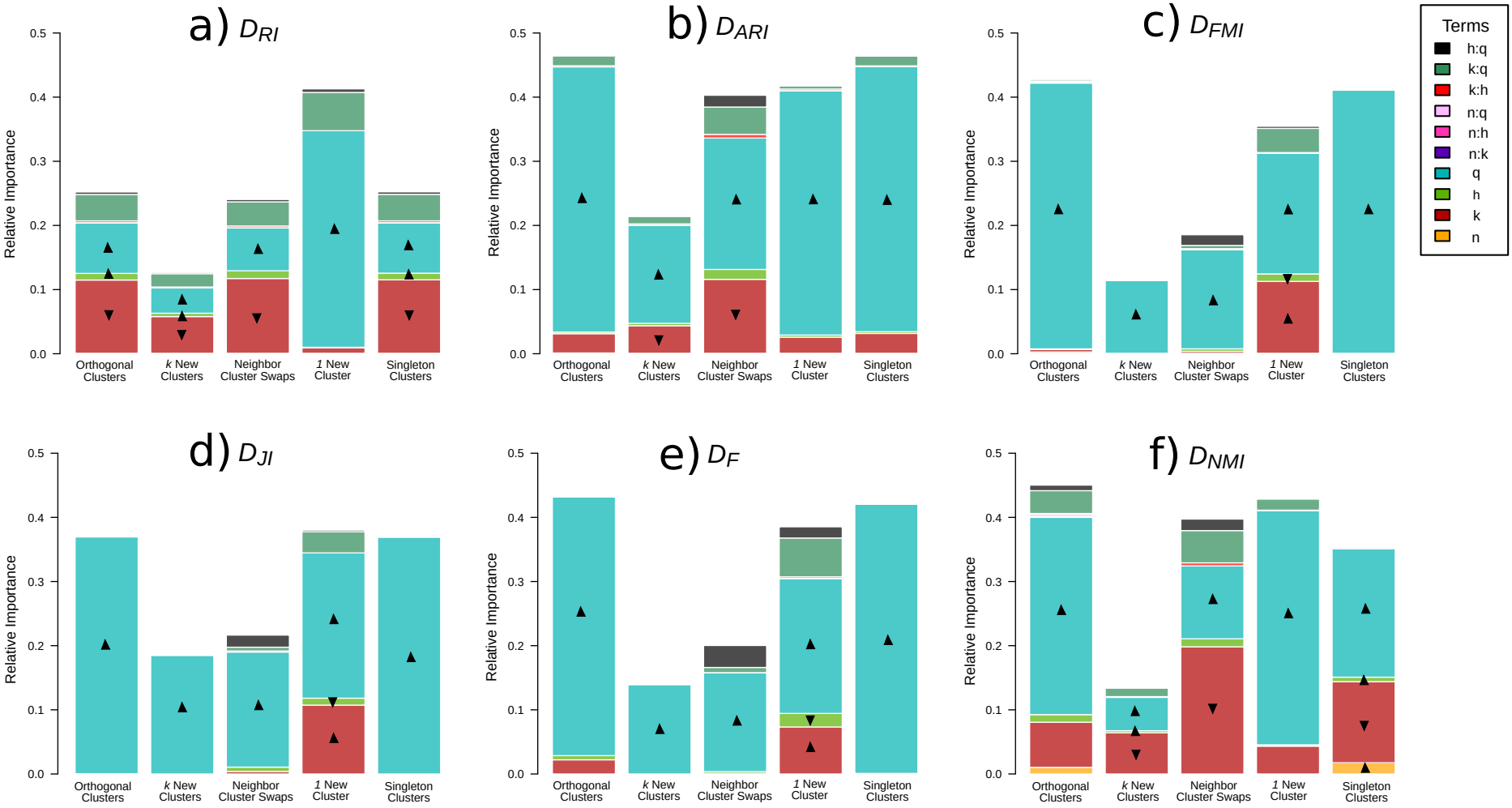
- **Relative importance analysis** $\longrightarrow$ squared beta weights

Experiments

- Our data: $50,000$ pairs of partitions
  - measures *(**m**)* = {Rand Index (D_{RI}), Adjusted Rand Index (D_{ARI}), Fowlkes-Mallows Index (D_{FMI}), Jaccard Index (D_{JI}), F-measure (D_F), Normalized Mutual Information (D_{NMI})}
  - transformations *(**t**)* = {K New Cluster, Singleton Clusters, 1 New Cluster, Neighbor Cluster Swap, Orthogonal Clusters}
  - number of elements *(**n**)* = $3240, 4320, .., 12960$
  - number of clusters *(**k**)* = $2, 3, .., 11$
  - heterogeneity of cluster sizes *(**h**)* = $0, 0.1, .., 0.9$
  - transformation intensity *(**q**)* = $0.1, 0.2, .., 1$

- Regression assumptions
  - by design $\longrightarrow$ no collinearity between the quantitative variables
  - large dataset & central limit theorem $\longrightarrow$ no issue with the residuals
  - heteroscedasticity $\longrightarrow$ increase of the variance in y with parameter q

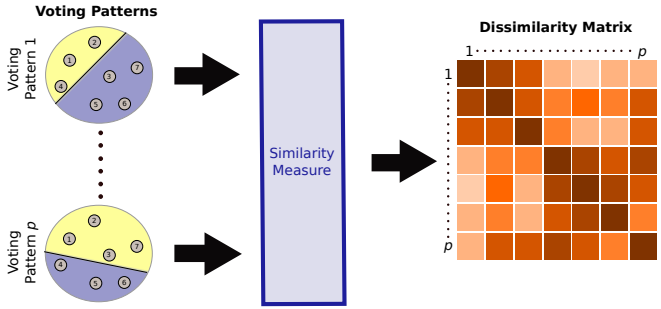
Results


a) D_{RI}
b) D_{ARI}
c) D_{FMI}
d) D_{JI}
e) D_{F}
f) D_{NMI}
Relative Importance
0.0
0.1
0.2
0.3
0.4
0.5
Orthogonal Clusters
k New Clusters
Neighbor Cluster Swaps
1 New Cluster
Singleton Clusters
Terms
h:q
k:q
k:h
n:q
n:h
n:k
q
h
k
n

Practical case: vote application in European Parliament

N. Arinik & R. Figueiredo & V. Labatut. Multiple Partitioning of Multiplex Signed Networks. Social Networks, 2020, 60, 83–102.

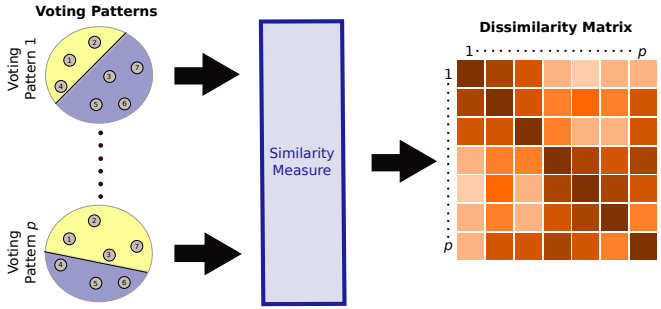
- Requirements:
  - up to 3 clusters
  - n fixed
- Expectations:
  - detecting an extra cluster, or a missing one, is an important change $\longrightarrow$ the difference of k between the original and transformed partitions
  - the effect of k should be stronger than that of h
  - a dissimilarity score should decrease, when h increases

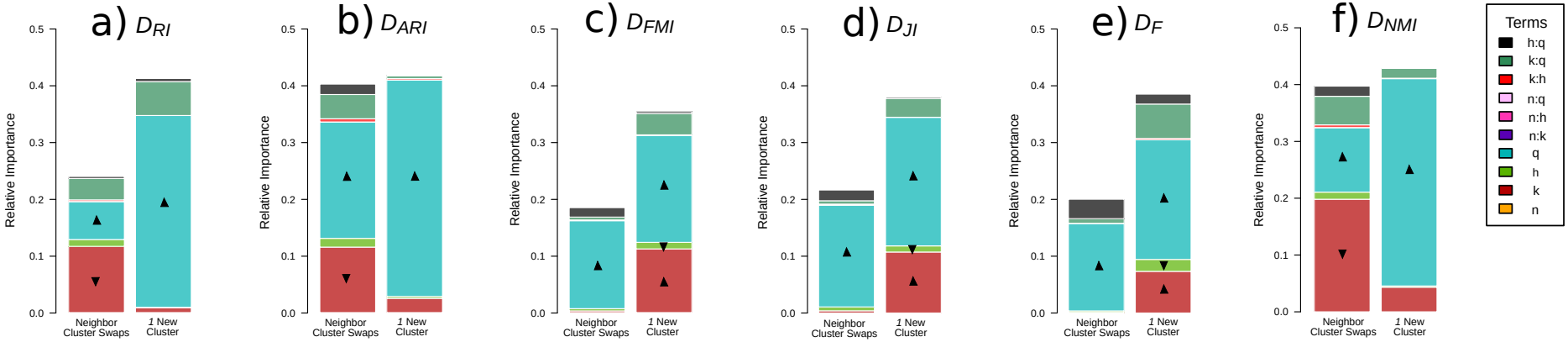


Practical case: vote application in European Parliament

N. Arinik & R. Figueiredo & V. Labatut. Multiple Partitioning of Multiplex Signed Networks. Social Networks, 2020, 60, 83–102.

- Requirements:
  - up to 3 clusters
  - n fixed
- Expectations:
  - detecting an extra cluster, or a missing one, is an important change $\longrightarrow$ the difference of k between the original and transformed partitions
  - the effect of k should be stronger than that of h
  - a dissimilarity score should decrease, when h increases

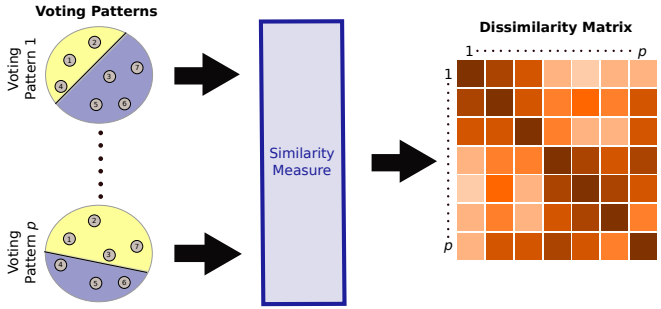


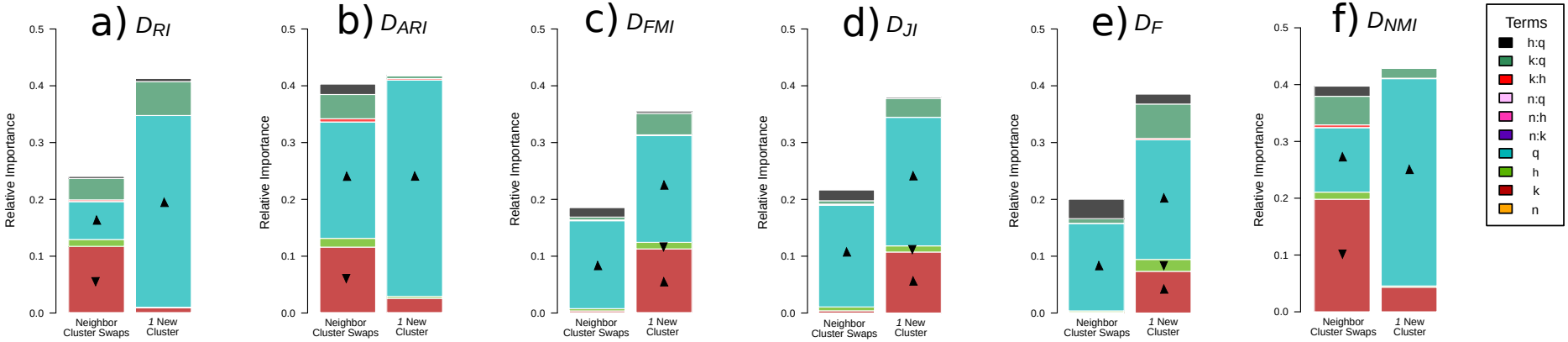


Practical case: vote application in European Parliament

N. Arinik & R. Figueiredo & V. Labatut. Multiple Partitioning of Multiplex Signed Networks. Social Networks, 2020, 60, 83–102.

- Requirements:
  - up to 3 clusters
  - n fixed
- Expectations:
  - detecting an extra cluster, or a missing one, is an important change $\longrightarrow$ the difference of k between the original and transformed partitions
  - the effect of k should be stronger than that of h
  - a dissimilarity score should decrease, when h increases

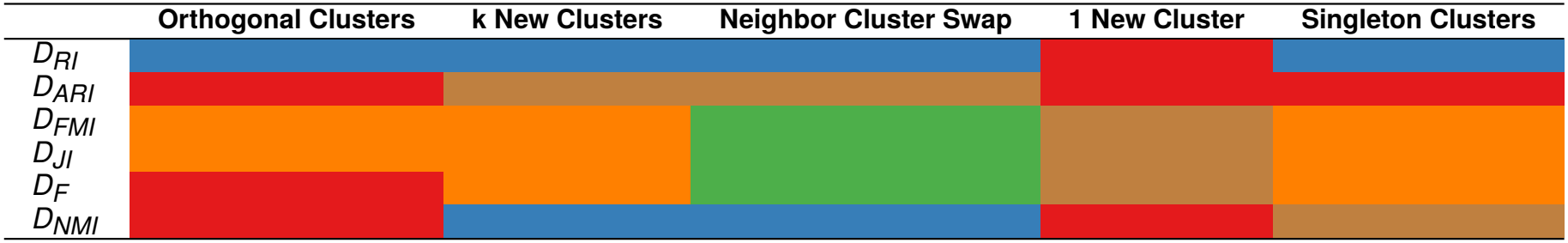




- **selected measure:** D_F (F-measure)

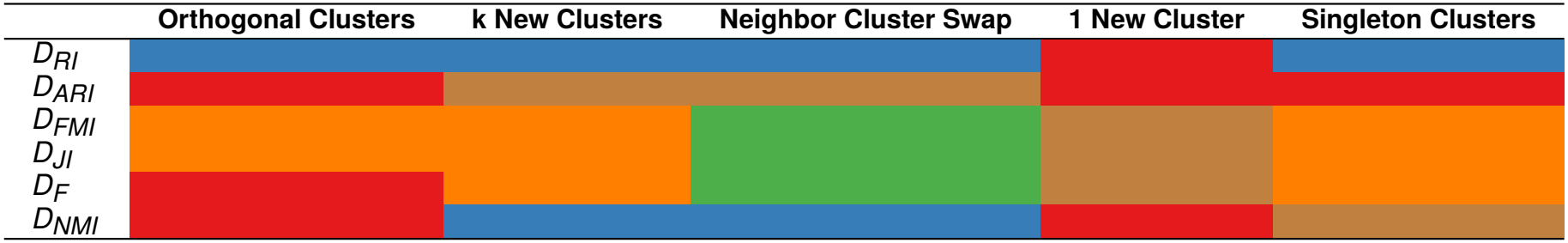
Conclusion & Further research

- a new generic framework of evaluation
- ease of interpretation for the results
- typology of measures based on their performances



Conclusion & Further research

- a new generic framework of evaluation
- ease of interpretation for the results
- typology of measures based on their performances



- evaluation with more measures and transformations
- designing the framework for graph similarity measures
- designing the framework for overlapping partitions

Thank you for your attention!

Contact Information:
Nejat ARINIK
nejat.arinik@univ-avignon.fr