Centrality and Distribution of Partitions according to the Transfer Distance

Lucile Denœud[1], Olivier Hudry[2]

1. Orange Labs, FTRD; 2. Telecom ParisTech

29, 30, 31 October 2008

Outline

Transfer distance between partitions

Definitions and notation

Let $\mathcal{P} = \{P_1, P_2, \dots P_p\}$ and $\mathcal{Q} = \{Q_1, Q_2, \dots Q_q\}$ be two partitions of $X = \{1, \dots, n\}$.
Let Σ be the set of matchings between the classes of $\mathcal{P}$ and the classes of $\mathcal{Q}$ and $\sigma \in \Sigma$.

σ-concordance

$$c_\sigma(\mathcal{P}, \mathcal{Q}) = \sum_{(P_i, Q_j) \in \sigma} \mid P_i \cap Q_j \mid$$

Concordance

$$c(\mathcal{P}, \mathcal{Q}) = \max_{\sigma \in \Sigma} c_\sigma(\mathcal{P}, \mathcal{Q}) = \max_{\sigma \in \Sigma} \sum_{(P_i, Q_j) \in \sigma} \mid P_i \cap Q_j \mid$$

Transfer distance

$$t(\mathcal{P}, \mathcal{Q}) = n - c(\mathcal{P}, \mathcal{Q}) = \min_{\sigma \in \Sigma} (n - \sum_{(P_i, Q_j) \in \sigma} \mid P_i \cap Q_j \mid)$$

Example

$X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,
$\mathcal{P} = (1, 2, 3|4, 5, 6|7, 8, 9)$, $\mathcal{Q} = (1, 3, 5, 6|2, 7, 9|4|8)$.

$\cap$	$Q_1 = \{1, 3, 5, 6\}$	$Q_2 = \{2, 7, 9\}$	$Q_3 = \{4\}$	$Q_4 = \{8\}$
$P_1 = \{1, 2, 3\}$	2	1	0	0
$P_2 = \{4, 5, 6\}$	2	0	1	0
$P_3 = \{7, 8, 9\}$	0	2	0	1

Example

$X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,
$\mathcal{P} = (1, 2, 3|4, 5, 6|7, 8, 9)$, $\mathcal{Q} = (1, 3, 5, 6|2, 7, 9|4|8)$.

$\cap$	$Q_1 = \{1, 3, 5, 6\}$	$Q_2 = \{2, 7, 9\}$	$Q_3 = \{4\}$	$Q_4 = \{8\}$
$P_1 = \{1, 2, 3\}$	**2**	1	0	0
$P_2 = \{4, 5, 6\}$	2	**0**	1	0
$P_3 = \{7, 8, 9\}$	0	2	**0**	1

$$\sigma = \{(P_1, Q_1), (P_2, Q_2), (P_3, Q_3)\}$$

$$c_\sigma(\mathcal{P}, \mathcal{Q}) = |P_1 \cap Q_1| + |P_2 \cap Q_2| + |P_3 \cap Q_3| = 2$$

Example

$X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,
$\mathcal{P} = (1, 2, 3|4, 5, 6|7, 8, 9)$, $\mathcal{Q} = (1, 3, 5, 6|2, 7, 9|4|8)$.

$\cap$	$Q_1 = \{1, 3, 5, 6\}$	$Q_2 = \{2, 7, 9\}$	$Q_3 = \{4\}$	$Q_4 = \{8\}$
$P_1 = \{1, 2, 3\}$	**2**	1	0	0
$P_2 = \{4, 5, 6\}$	2	0	**1**	0
$P_3 = \{7, 8, 9\}$	0	**2**	0	1

$$\hat{\sigma} = \{(P_1, Q_1), (P_2, Q_3), (P_3, Q_2)\}$$

$$c_{\hat{\sigma}}(\mathcal{P}, \mathcal{Q}) = |P_1 \cap Q_1| + |P_2 \cap Q_3| + |P_3 \cap Q_2| = 5 = c(\mathcal{P}, \mathcal{Q})$$

Example

$X = \{1,2,3,4,5,6,7,8,9\}$,
$\mathcal{P} = (1,2,3|4,5,6|7,8,9)$, $\mathcal{Q} = (1,3,5,6|2,7,9|4|8)$.

$\cap$	$Q_1 = \{1,3,5,6\}$	$Q_2 = \{2,7,9\}$	$Q_3 = \{4\}$	$Q_4 = \{8\}$
$P_1 = \{1,2,3\}$	**2**	1	0	0
$P_2 = \{4,5,6\}$	2	0	**1**	0
$P_3 = \{7,8,9\}$	0	**2**	0	1

$$\hat{\sigma} = \{(P_1, Q_1), (P_2, Q_3), (P_3, Q_2)\}$$

$$c_{\hat{\sigma}}(\mathcal{P}, \mathcal{Q}) = |P_1 \cap Q_1| + |P_2 \cap Q_3| + |P_3 \cap Q_2| = 5 = c(\mathcal{P}, \mathcal{Q})$$

$$t(\mathcal{P}, \mathcal{Q}) = n - c(\mathcal{P}, \mathcal{Q}) = 9 - 5 = 4$$

Example

$$X = \{1,2,3,4,5,6,7,8,9\},$$
$$\mathcal{P} = (1,2,3|4,5,6|7,8,9),\ \mathcal{Q} = (1,3,5,6|2,7,9|4|8).$$

$\cap$	$Q_1 = \{1,3,5,6\}$	$Q_2 = \{2,7,9\}$	$Q_3 = \{4\}$	$Q_4 = \{8\}$
$P_1 = \{1,2,3\}$	**2**	1	0	0
$P_2 = \{4,5,6\}$	2	0	**1**	0
$P_3 = \{7,8,9\}$	0	**2**	0	1

$$\hat{\sigma} = \{(P_1, Q_1), (P_2, Q_3), (P_3, Q_2)\}$$

$$c_{\hat{\sigma}}(\mathcal{P}, \mathcal{Q}) = |P_1 \cap Q_1| + |P_2 \cap Q_3| + |P_3 \cap Q_2| = 5 = c(\mathcal{P}, \mathcal{Q})$$

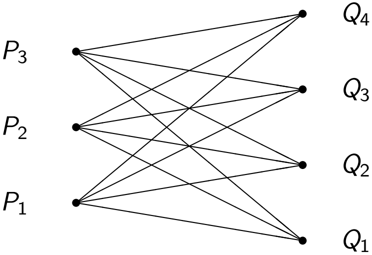
$$t(\mathcal{P}, \mathcal{Q}) = n - c(\mathcal{P}, \mathcal{Q}) = 9 - 5 = 4$$

$$(1,2,3 \mid 4,5,6 \mid 7,8,9) \to (1,3 \mid 4,5,6 \mid 2,7,8,9) \to (1,3,5 \mid 4,6 \mid 2,7,8,9)$$
$$\to (1,3,5,6 \mid 4 \mid 2,7,8,9) \to (1,3,5,6 \mid 4 \mid 2,7,9 \mid 8).$$

Computation of the transfer distance

The compuation of the concordance comes down to determine a maximum matching in a bipartite graph (Day, 1981).

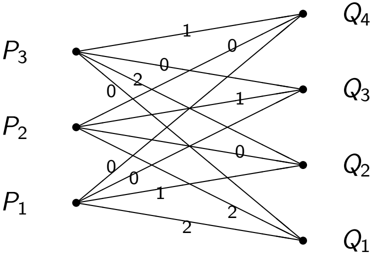
Example : $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,
$\mathcal{P} = (1, 2, 3|4, 5, 6|7, 8, 9)$, $\mathcal{Q} = (1, 3, 5, 6|2, 7, 9|4|8)$.



Computation of the transfer distance

The compuation of the concordance comes down to determine a maximum matching in a bipartite graph (Day, 1981).

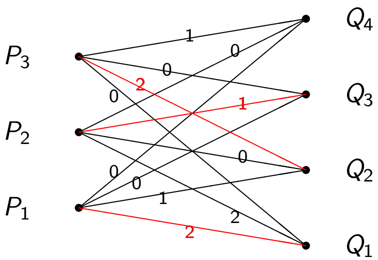
Example : $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,
$\mathcal{P} = (1, 2, 3|4, 5, 6|7, 8, 9)$, $\mathcal{Q} = (1, 3, 5, 6|2, 7, 9|4|8)$.



Computation of the transfer distance

The compuation of the concordance comes down to determine a maximum matching in a bipartite graph (Day, 1981).

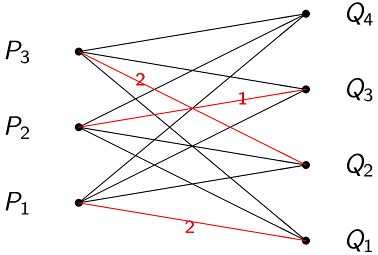
Example : $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,
$\mathcal{P} = (1, 2, 3|4, 5, 6|7, 8, 9)$, $\mathcal{Q} = (1, 3, 5, 6|2, 7, 9|4|8)$.



Computation of the transfer distance

The compuation of the concordance comes down to determine a maximum matching in a bipartite graph (Day, 1981).

Example : $X = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$,
$\mathcal{P} = (1, 2, 3|4, 5, 6|7, 8, 9)$, $\mathcal{Q} = (1, 3, 5, 6|2, 7, 9|4|8)$.



Resolution : Hungarian method (Kuhn, 1955) : $O(\max(p, q)^3)$.

Close partitions in terms of transfers

Experimental study : methodology

Let X be a given set with n elements, and let $\mathbb{P}_n$ denote the set of the partitions on X. Let $\mathcal{P}$ be a given partition on X (*reference partition*). We study the distribution of the partitions of $\mathbb{P}_n$ according to the transfer distance to the partition $\mathcal{P}$.

- $n \leqslant 12$: Enumeration of the set $\mathbb{P}_n$.

  Example : For $n = 12$, $|\mathbb{P}_n| = 4\ 213\ 597$.

- $n > 12$: Sampling of the set $\mathbb{P}_n$.

  Example : For $n = 100$, $|\mathbb{P}_n| \approx 10^{115}$

  We use a drawing with a uniform distribution over $\mathbb{P}_n$ (A. Nijenhuis and H. Wilf (1978) : Random Partition of an n-Set, *Combinatorial algorithms*).

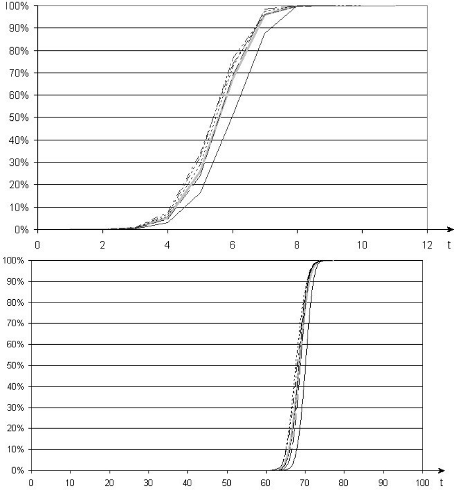
  The drawing of only 10 000 partitions is needed to estimate the studied proportions to 0.01 with confidence level 95%.

Distributions of the partitions of $\mathbb{P}_n$ according to the transfer distance

Distributions according to 5 reference partitions drawn randomly, and average distribution (cumulated percentages).

$n = 12$

$n = 100$



Close partitions in terms of transfers

We consider the reference partition $\mathcal{P}$. Being given a threshold $\alpha \in [0;1]$, we define two critical values $t_\alpha^-(\mathcal{P})$ and $t_\alpha^+(\mathcal{P})$ by :

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})+1\}|}{|\mathbb{P}_n|} > \alpha$$

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})-1\}|}{|\mathbb{P}_n|} > \alpha.$$

Close partitions in terms of transfers

We consider the reference partition $\mathcal{P}$. Being given a threshold $\alpha \in [0;1]$, we define two critical values $t_\alpha^-(\mathcal{P})$ and $t_\alpha^+(\mathcal{P})$ by :

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})+1\}|}{|\mathbb{P}_n|} > \alpha$$

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})-1\}|}{|\mathbb{P}_n|} > \alpha.$$

Example : $\mathcal{P} = (123|45|67|89|10|11|12)$

t	1	2	3	4	5	6	7	8	9
%	0.00	0.06	0.89	7.63	33.94	73.93	95.83	99.84	100

$$\Rightarrow t_{1\%}^- = 3 \text{ and } t_{1\%}^+ = 8$$

Close partitions in terms of transfers

We consider the reference partition $\mathcal{P}$. Being given a threshold $\alpha \in [0;1]$, we define two critical values $t_\alpha^-(\mathcal{P})$ and $t_\alpha^+(\mathcal{P})$ by :

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})+1\}|}{|\mathbb{P}_n|} > \alpha$$

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})-1\}|}{|\mathbb{P}_n|} > \alpha.$$

Definition

Let $\mathcal{Q}$ be a partition of X. We will say that $\mathcal{Q}$ is close to $\mathcal{P}$ at threshold α if $t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-$, and conversely $\mathcal{Q}$ will be considered as far from $\mathcal{P}$ at threshold α if $t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+$.

Close partitions in terms of transfers

We consider the reference partition $\mathcal{P}$. Being given a threshold $\alpha \in [0;1]$, we define two critical values $t_\alpha^-(\mathcal{P})$ and $t_\alpha^+(\mathcal{P})$ by :

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-(\mathcal{P})+1\}|}{|\mathbb{P}_n|} > \alpha$$

$$\frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})\}|}{|\mathbb{P}_n|} \leqslant \alpha, \frac{|\{\mathcal{Q} \in \mathbb{P}_n, t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+(\mathcal{P})-1\}|}{|\mathbb{P}_n|} > \alpha.$$

Definition

Let $\mathcal{Q}$ be a partition of X. We will say that $\mathcal{Q}$ is close to $\mathcal{P}$ at threshold α if $t(\mathcal{P},\mathcal{Q}) \leqslant t_\alpha^-$, and conversely $\mathcal{Q}$ will be considered far from $\mathcal{P}$ at threshold α if $t(\mathcal{P},\mathcal{Q}) \geqslant t_\alpha^+$.

Example :

$$t_{1\%}^- = 3 \text{ and } t_{1\%}^+ = 8$$

$\mathcal{Q}$ is close to $\mathcal{P}$ at threshold 1% if $t(\mathcal{P},\mathcal{Q}) \leqslant 3$, and
$\mathcal{Q}$ is far from $\mathcal{P}$ at threshold 1% if $t(\mathcal{P},\mathcal{Q}) \geqslant 8$.

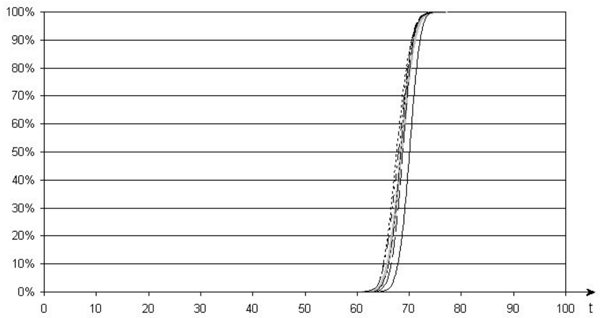
Close partitions in terms of transfers

Conclusion

- The partitions of $\mathbb{P}_n$ are concentrated on a narrow transfer interval.
- The critical values are large and close to each other : there are few transfer values for which we are not able to give an interpretation.
- This study provides non-intuitive information about the transfer distance and may permit to discriminate partitions being at close values of transfer distances.

Centrality of a partition

Centrality of a partition



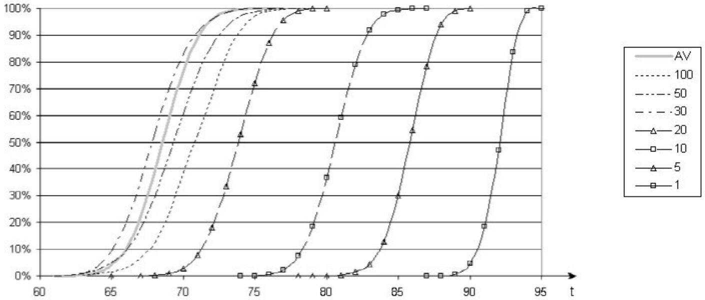
We observe a small shift between the graphs.

Centrality definition

The closer to $\mathcal{P}$ the partitions of $\mathbb{P}_n$ are located (the observed values of transfer are small, or, in other words, the distribution graph is close to the y-axis), the more $\mathcal{P}$ can be considered as central.

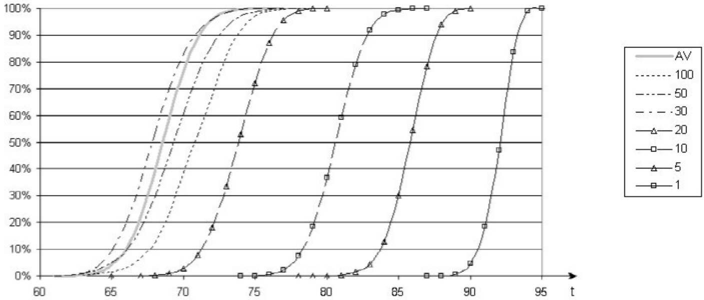
Impact of the number of classes

Distributions of the partitions with respect to reference partitions with p =1, 5, 10, 20, 30, 50 and 100 balanced classes.



Impact of the number of classes

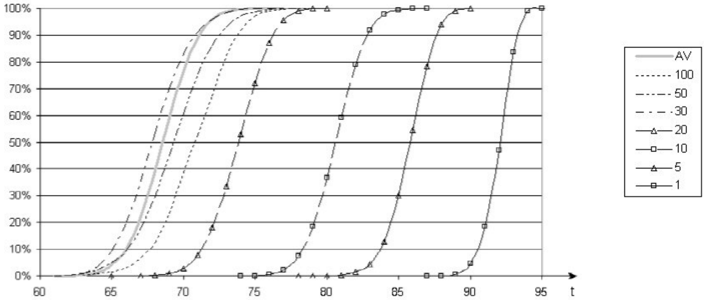
Distributions of the partitions with respect to reference partitions with p =1, 5, 10, 20, 30, 50 and 100 balanced classes.



$$(p = 1) > (p = 5) > (p = 10) > (p = 20) > (p = 30)$$
$$(p = 30) < (p = 50) < (p = 100)$$

Impact of the number of classes

Distributions of the partitions with respect to reference partitions with p =1, 5, 10, 20, 30, 50 and 100 balanced classes.

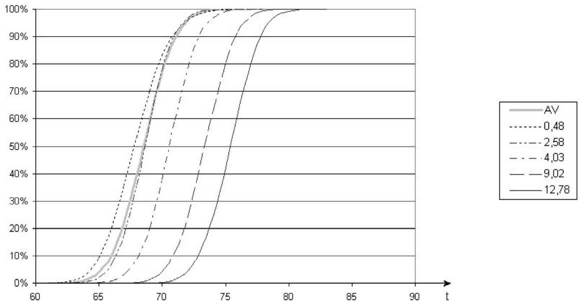


$$(p = 1) > (p = 5) > (p = 10) > (p = 20) > (p = 30)$$
$$(p = 30) < (p = 50) < (p = 100)$$

The centrality seems to be unimodal with the number of classes in the reference partition.

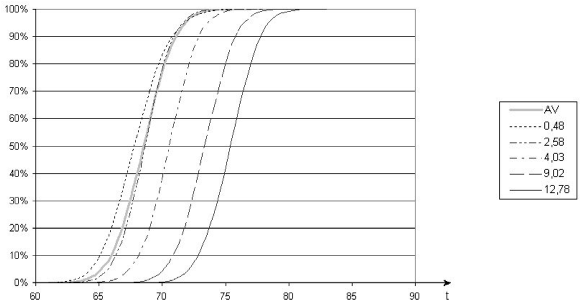
Impact of the balance of classes

Distributions of the partitions with respect to reference partitions with 30 classes more or less balanced (the standard deviation σ of the cardinalities of the classes varies).



Impact of the balance of classes

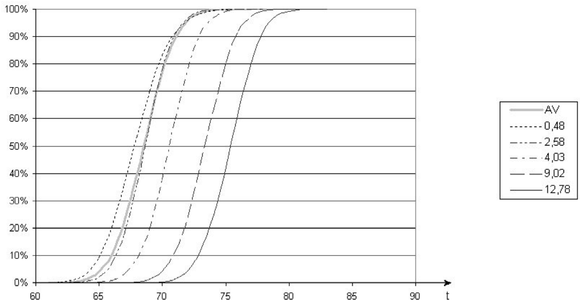
Distributions of the partitions with respect to reference partitions with 30 classes more or less balanced (the standard deviation σ of the cardinalities of the classes varies).



$$(\sigma = 12.78) > (\sigma = 9.02) > (\sigma = 4.03) > (\sigma = 2.58) > (\sigma = 0.48)$$

Impact of the balance of classes

Distributions of the partitions with respect to reference partitions with 30 classes more or less balanced (the standard deviation σ of the cardinalities of the classes varies).



$$(\sigma = 12.78) > (\sigma = 9.02) > (\sigma = 4.03) > (\sigma = 2.58) > (\sigma = 0.48)$$

The centrality seems to be increasing when the balance of classes increases (the standard deviation decreases).

Centrality of a partition

Conjectures

- The centrality is unimodal with the number of classes.
- The centrality is increasing when the classes become more balanced.
- The partition with one class is the most eccentric partition of $\mathbb{P}_n$.

Future work

- Study of the centrality notion.
- Which partition(s) is (are) the most central ? (characteristics near the average ones ?)