LETTRES
SORBONNE
UNIVERSITÉ

Partitionnement des actualités avec MAJORCLUST et distances temporaires

Carlos González

03/12/2020

Table des matières

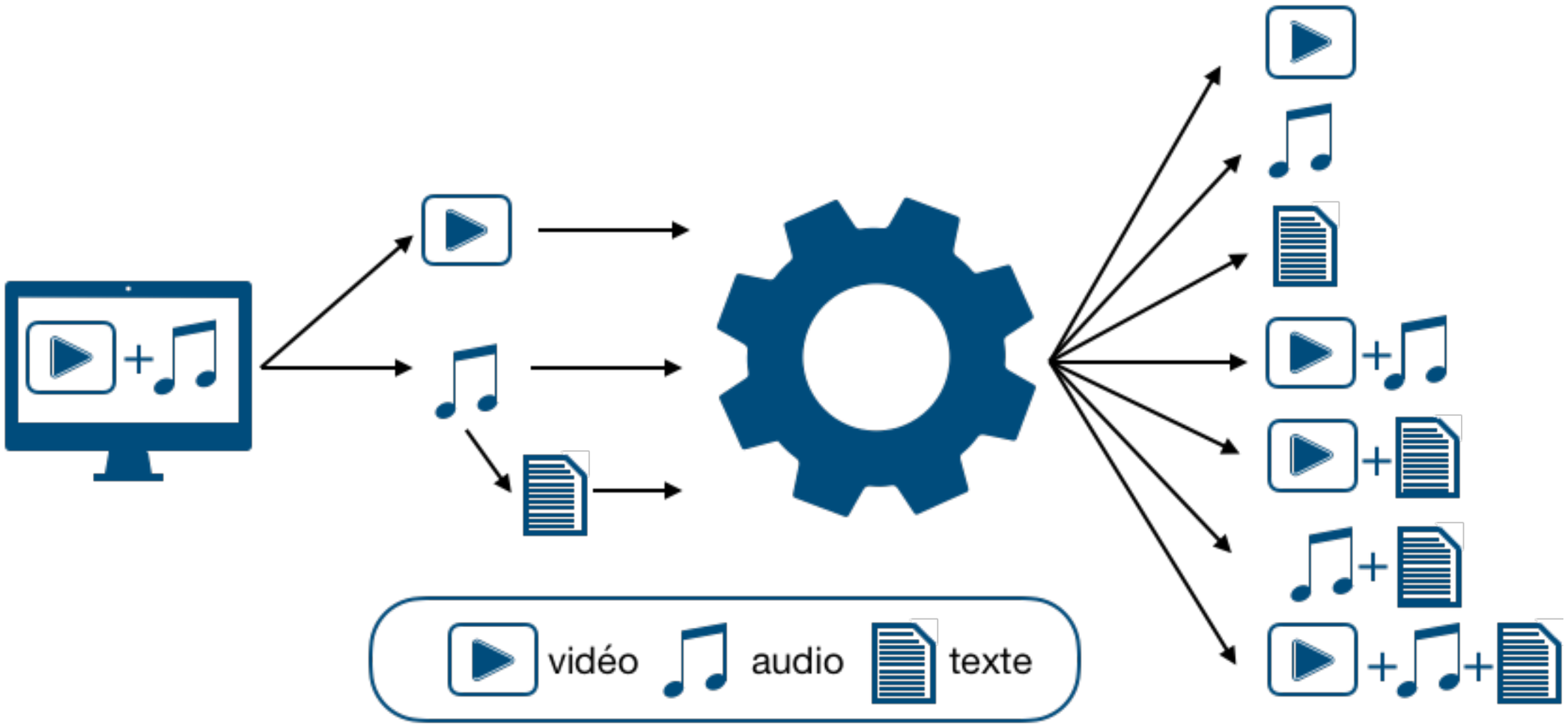
ACTUALITÉS / TÉLÉREPORTAGES

?

→ résumé + RI

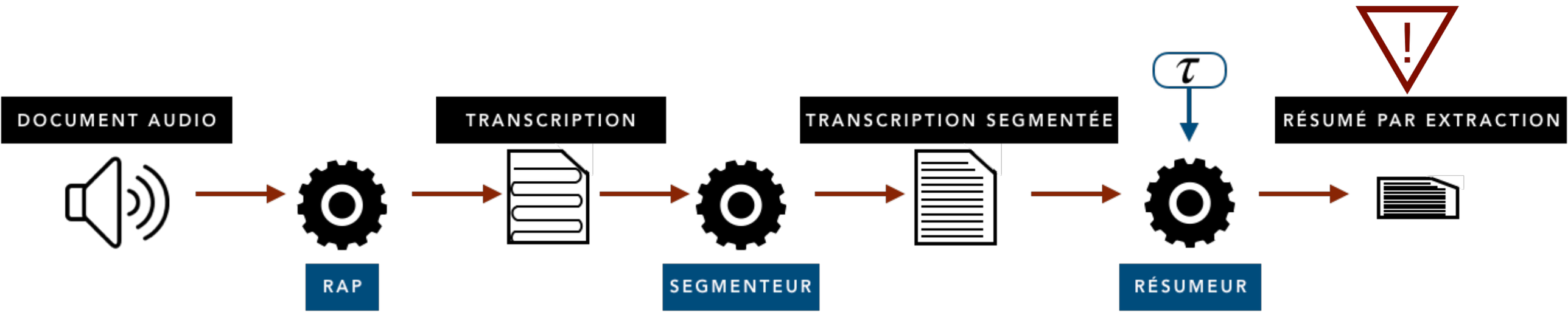
multimédia & multilingue

Introduction



Introduction

Problème



Actualités / Téléreportages
>1 sujet

Partitionnement des sujets
avec MAJORCLUST

documents dépendants
de la temporalité

Partitionnement avec MAJORCLUST

«Structure defines the organization of parts as dominated by the general character of the whole »

On the Nature of Structure and Its Identification

Benno Stein and Oliver Niggemann
Dept. of Mathematics and Computer Science—Knowledge-based Systems,
University of Paderborn, D–33095 Paderborn, Germany
{stein,murray}@uni-paderborn.de

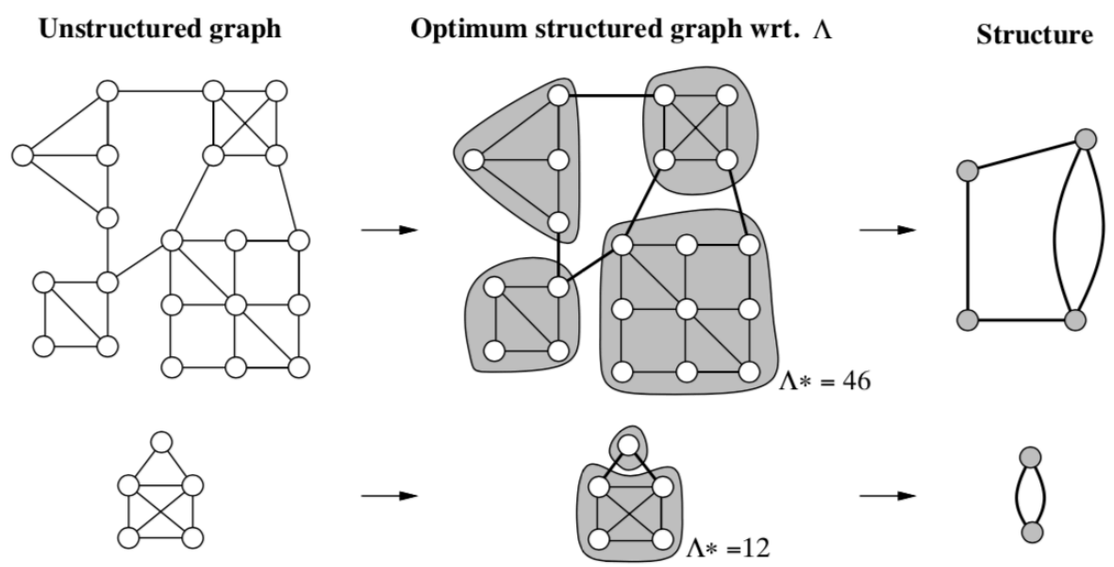


Fig. 3. Examples for decomposing a graph according to our structure definition.

- Stein, Benno, and Oliver Niggemann. **"On the nature of structure and its identification."** *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, Berlin, Heidelberg, 1999.

Partitionnement avec MAJORCLUST

MAJORCLUST.
Input. A graph $G = \langle V, E \rangle$.
Output. A function $c : V \mapsto \mathbf{N}$, which assigns a cluster number to each node.

(1) $n = 0$, $t = \mathit{false}$
(2) $\forall v \in V$ **do** $n = n + 1$, $c(v) = n$ **end**
(3) **while** $t = \mathit{false}$ **do**
(4) $\quad t = \mathit{true}$
(5) $\quad \forall v \in V$ **do**
(6) $\quad\quad c^* = i$ **if** $\left|\{u : \{u, v\} \in E \wedge c(u) = i\}\right|$ is max.
(7) $\quad\quad$ **if** $c(v) \neq c^*$ **then** $c(v) = c^*, t = \mathit{false}$
(8) $\quad$ **end**
(9) **end**

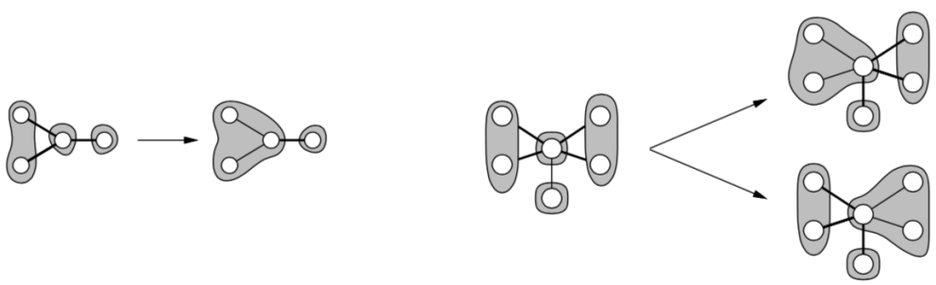


Fig. 4. A definite majority clustering situation (left) and an undecided majority clustering situation (right).

1. Chaque nœud du graphe est attribué à son propre cluster

2. Un nœud adopte le même cluster auquel appartient la majorité de ses voisins
   - Si plusieurs options ⟹ choix aléatoire
   - Si stable ⟹ FIN

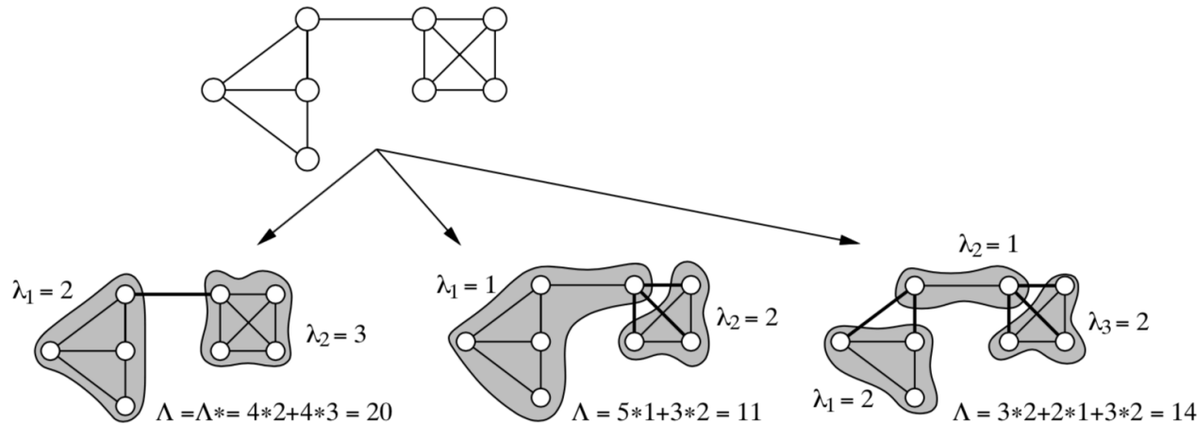


Fig. 2. Graph decompositions and related Λ values.

Λ : Connectivité partielle pondérée

- Stein, Benno, and Oliver Niggemann. **"On the nature of structure and its identification."** *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, Berlin, Heidelberg, 1999.
- Stein, Benno, and S. Meyerzu Eißen. **"Document categorization with MajorClust."** *Proc. 12th Workshop on Information Technology and Systems*. Citeseer, 2002.

Partitionnement avec MAJORCLUST

MAJORCLUST.
Input. A graph $G = \langle V, E \rangle$.
Output. A function $c : V \mapsto \mathbf{N}$, which assigns a cluster number to each node.

(1) $n = 0$, $t = \mathit{false}$
(2) $\forall v \in V$ **do** $n = n + 1$, $c(v) = n$ **end**
(3) **while** $t = \mathit{false}$ **do**
(4) $t = \mathit{true}$
(5) $\forall v \in V$ **do**
(6) $c^* = i$ **if** $\left|\{u : \{u, v\} \in E \wedge c(u) = i\}\right|$ is max.
(7) **if** $c(v) \neq c^*$ **then** $c(v) = c^*, t = \mathit{false}$
(8) **end**
(9) **end**

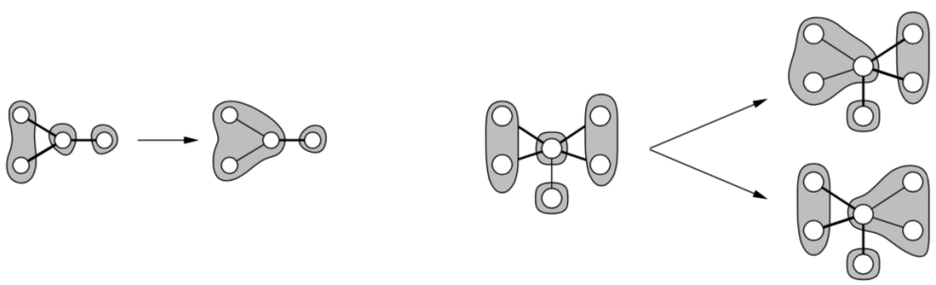


Fig. 4. A definite majority clustering situation (left) and an undecided majority clustering situation (right).

1. Chaque nœud du graphe est attribué à son propre cluster

2. Un nœud adopte le même cluster auquel appartient la majorité de ses voisins
   - Si plusieurs options ⟹ choix aléatoire
   - Si stable ⟹ FIN

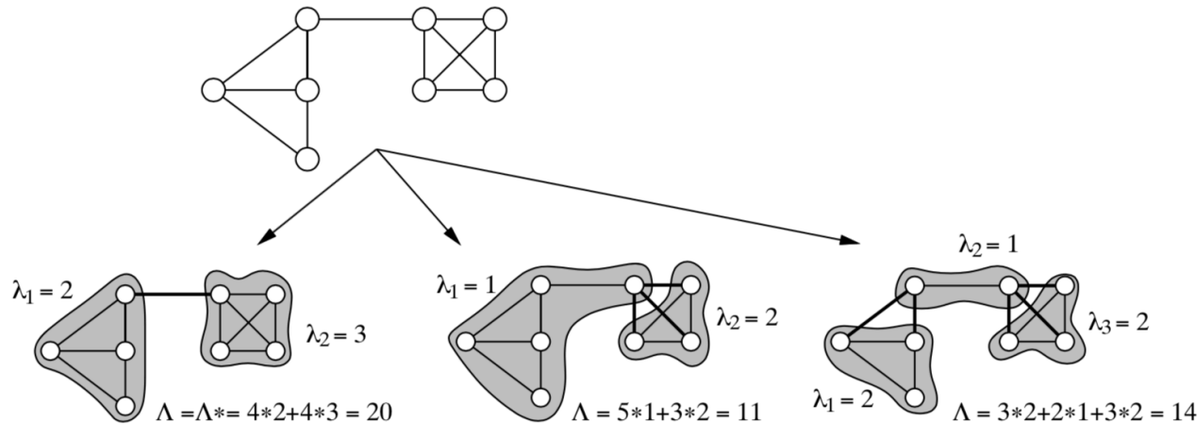


Fig. 2. Graph decompositions and related Λ values.

Λ : Connectivité partielle pondérée

- Stein, Benno, and Oliver Niggemann. **"On the nature of structure and its identification."** *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, Berlin, Heidelberg, 1999.
- Stein, Benno, and S. Meyerzu Eißen. **"Document categorization with MajorClust."** *Proc. 12th Workshop on Information Technology and Systems*. Citeseer, 2002.

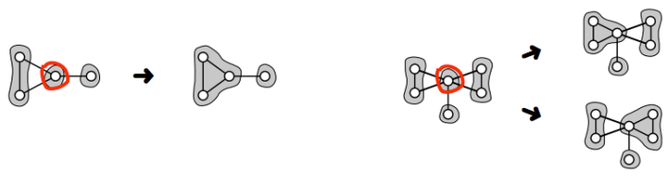
Partitionnement avec MAJORCLUST

MAJORCLUST.
Input. A graph $G = \langle V, E, \varphi \rangle$.
Output. A function $c : V \to \mathbf{N}$, which assigns a cluster number to each node.

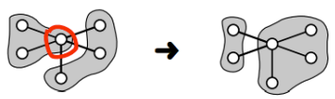
(1) $n = 0, t = \mathit{false}$
(2) $\forall v \in V$ **do** $n = n + 1, c(v) = n$ **end**
(3) **while** $t = \mathit{false}$ **do**
(4) $t = \mathit{true}$
(5) $\forall v \in V$ **do**
(6) $c^* = i$ **if** $\left(\sum_{\substack{c(u)=i, \\ \{u,v\} \in E}} \varphi(u, v) \right)$ is max.
(7) **if** $c(v) \neq c^*$ **then** $c(v) = c^*, t = \mathit{false}$
(8) **end**
(9) **end**

Illustration of Steps (6)+(7) in MAJORCLUST.

A definite majority decision (left), and an indefinite decision (right), when assigning a node to a cluster:



Re-clustering—a node changes its cluster:



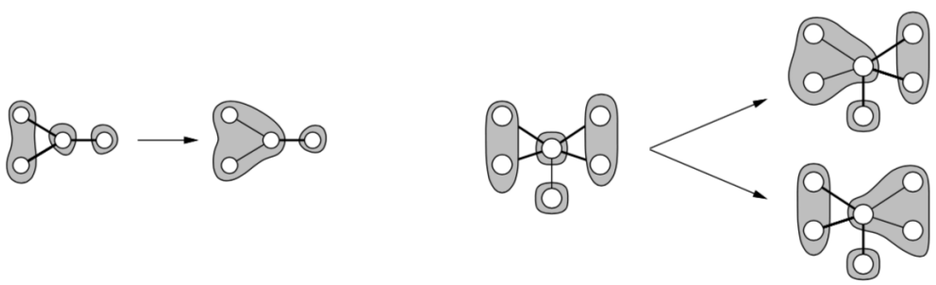


Fig. 4. A definite majority clustering situation (left) and an undecided majority clustering situation (right).

1. Chaque nœud du graphe est attribué à son propre cluster

2. Un nœud adopte le même cluster auquel appartient la majorité de ses voisins
   - Si plusieurs options ⟹ choix aléatoire
   - Si stable ⟹ FIN

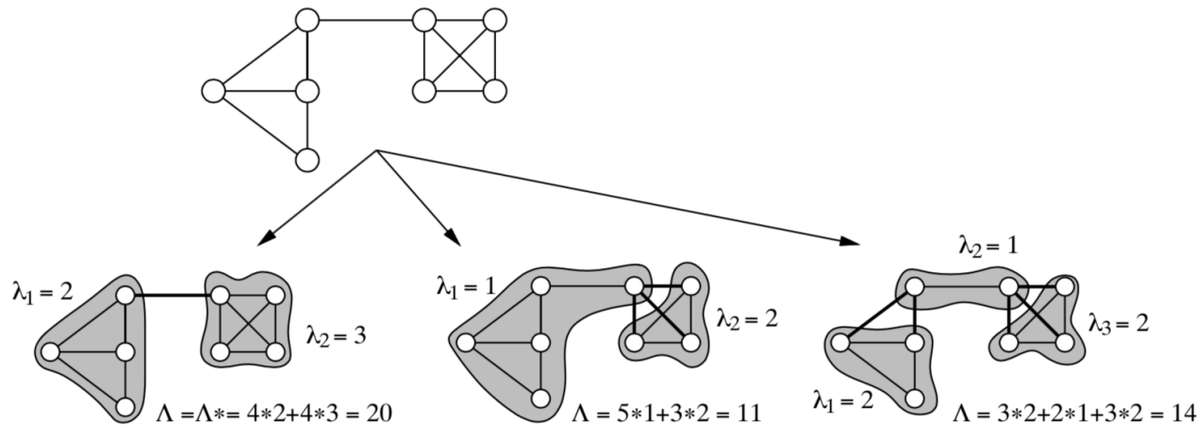


Fig. 2. Graph decompositions and related Λ values.

Λ : Connectivité partielle pondérée

- Stein, Benno, and Oliver Niggemann. **"On the nature of structure and its identification."** *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, Berlin, Heidelberg, 1999.
- Stein, Benno, and S. Meyerzu Eißen. **"Document categorization with MajorClust."** *Proc. 12th Workshop on Information Technology and Systems*. Citeseer, 2002.

Partitionnement avec MAJORCLUST

MAJORCLUST.
Input. A graph $G = \langle V, E, \varphi \rangle$.
Output. A function $c : V \to \mathbf{N}$, which assigns a cluster number to each node.

(1) $n = 0, t = \mathit{false}$
(2) $\forall v \in V$ **do** $n = n + 1, c(v) = n$ **end**
(3) **while** $t = \mathit{false}$ **do**
(4) $t = \mathit{true}$
(5) $\forall v \in V$ **do**
(6) $c^* = i$ **if** $\left(\sum_{\substack{c(u)=i \\ \{u,v\} \in E}} \varphi(u, v) \right)$ is max.
(7) **if** $c(v) \neq c^*$ **then** $c(v) = c^*, t = \mathit{false}$
(8) **end**
(9) **end**

$$\varphi(d_i, d_j) = \cos(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i||\mathbf{d}_j|} \; ; \; \varphi(d_i, d_j) \geq \tau$$

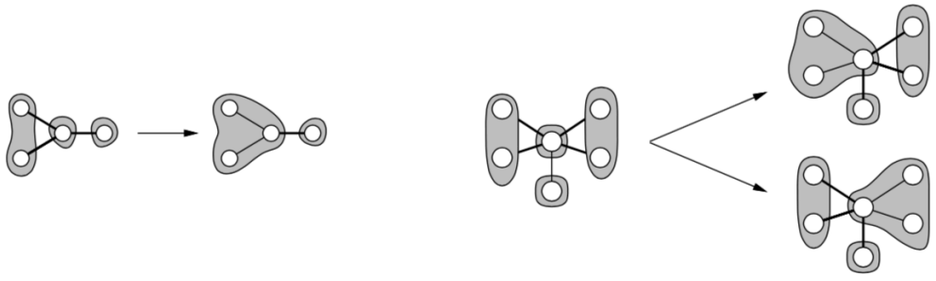
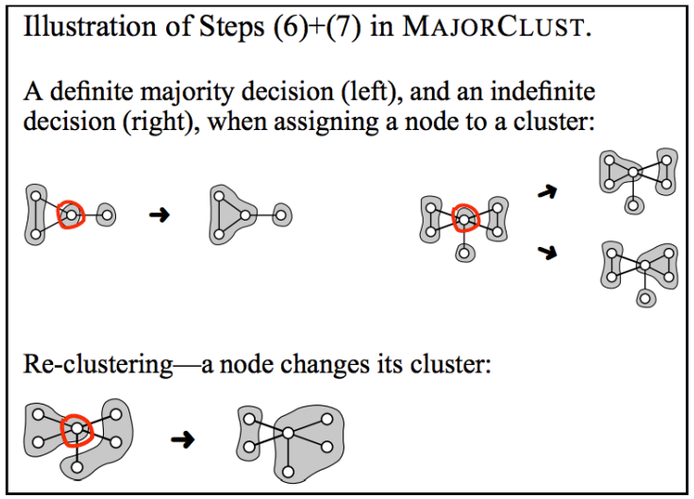


Fig. 4. A definite majority clustering situation (left) and an undecided majority clustering situation (right).



1. Chaque nœud du graphe est attribué à son propre cluster

2. Un nœud adopte le même cluster auquel appartient la majorité de ses voisins
   - Si plusieurs options ⟹ choix aléatoire
   - Si stable ⟹ FIN

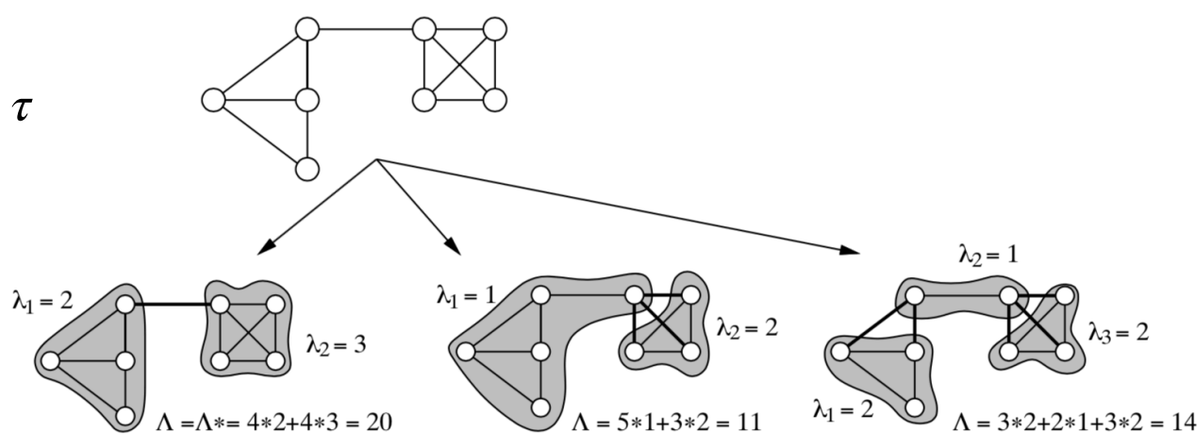


Fig. 2. Graph decompositions and related Λ values.

Λ : Connectivité partielle pondérée

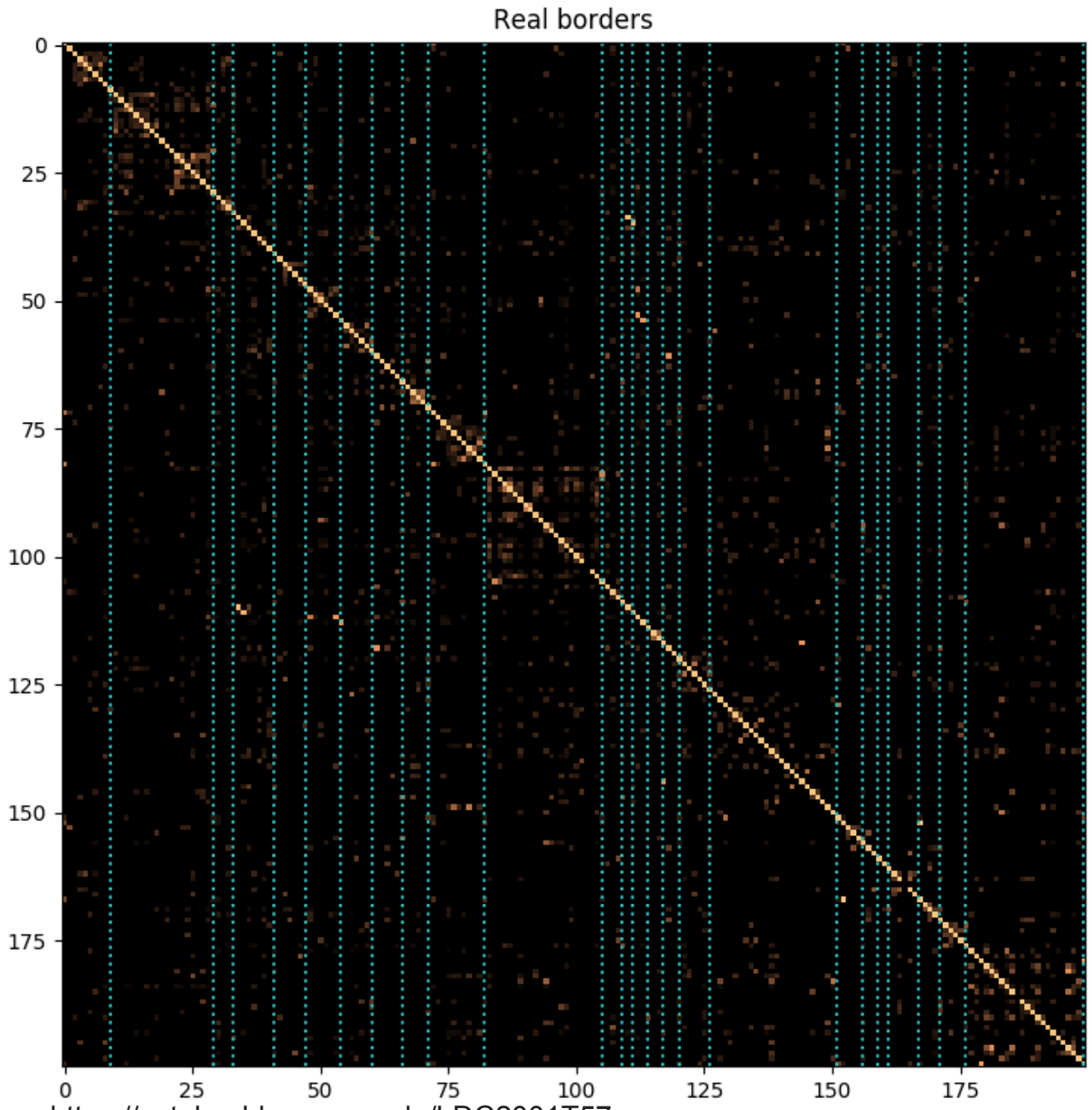
- Stein, Benno, and Oliver Niggemann. **"On the nature of structure and its identification."** *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, Berlin, Heidelberg, 1999.
- Stein, Benno, and S. Meyerzu Eißen. **"Document categorization with MajorClust."** *Proc. 12th Workshop on Information Technology and Systems*. Citeseer, 2002.

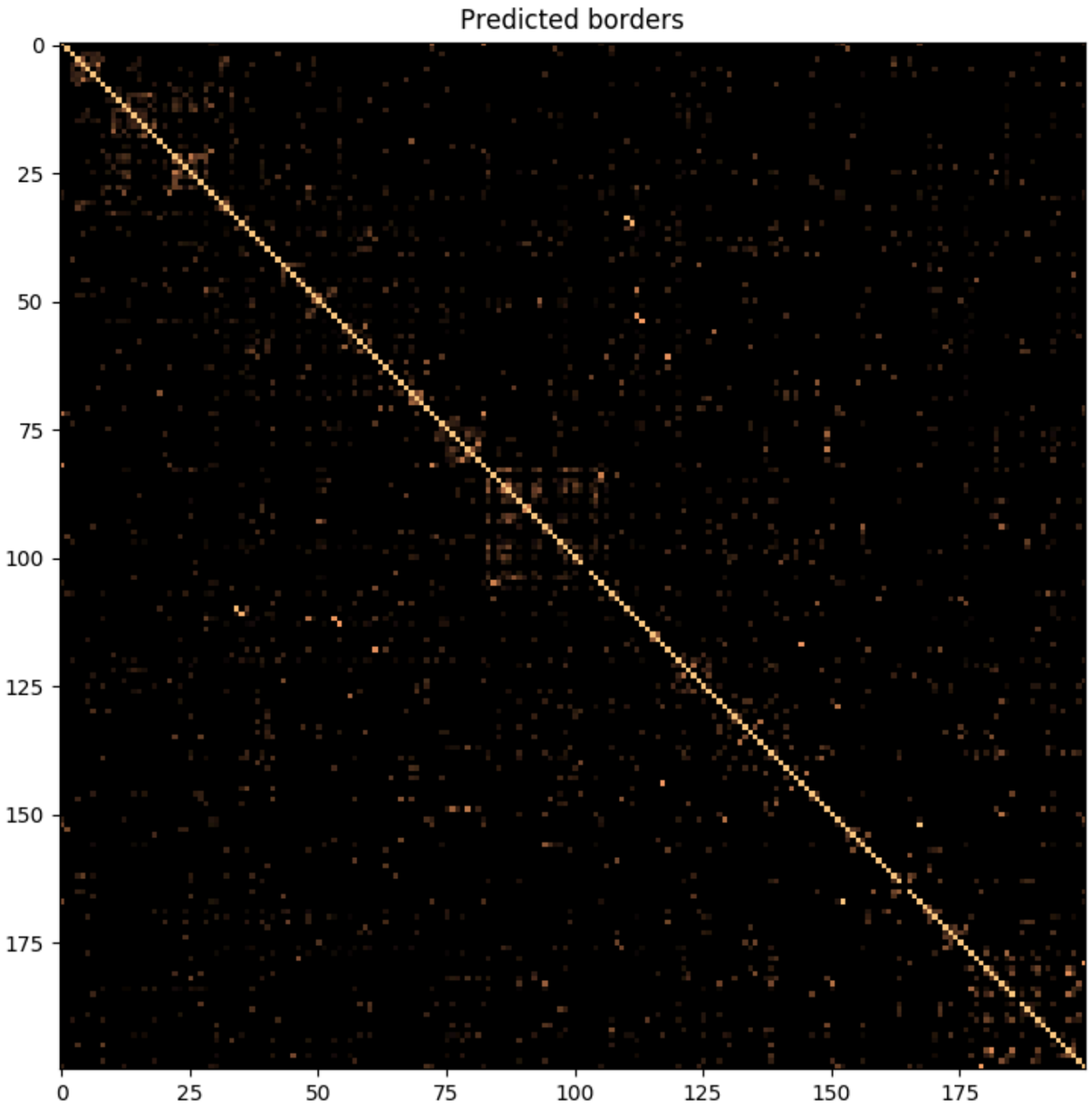
Partitionnement avec MAJORCLUST

Documents dépendants de la temporalité

- Corpus : TDT2 English broadcast news corpus
- Transcriptions : manuelles
- Éléments grammaticaux :
  - verbes
  - noms
  - adjectifs
  - adverbes

- Similarité : $$\varphi(d_i, d_j) = \cos(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i||\mathbf{d}_j|}$$





- https://catalog.ldc.upenn.edu/LDC2001T57

Partitionnement avec MAJORCLUST

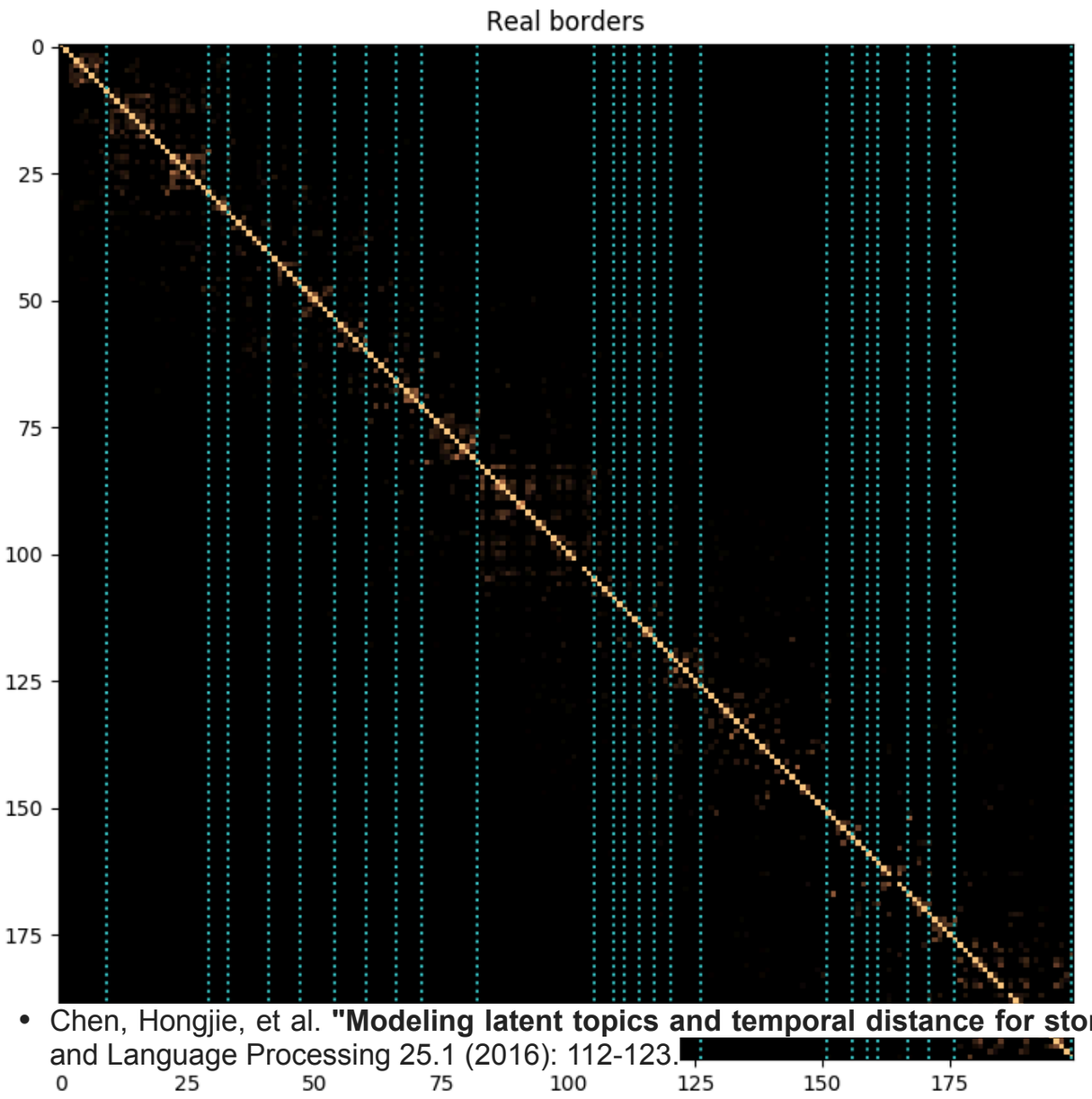
Documents dépendants de la temporalité

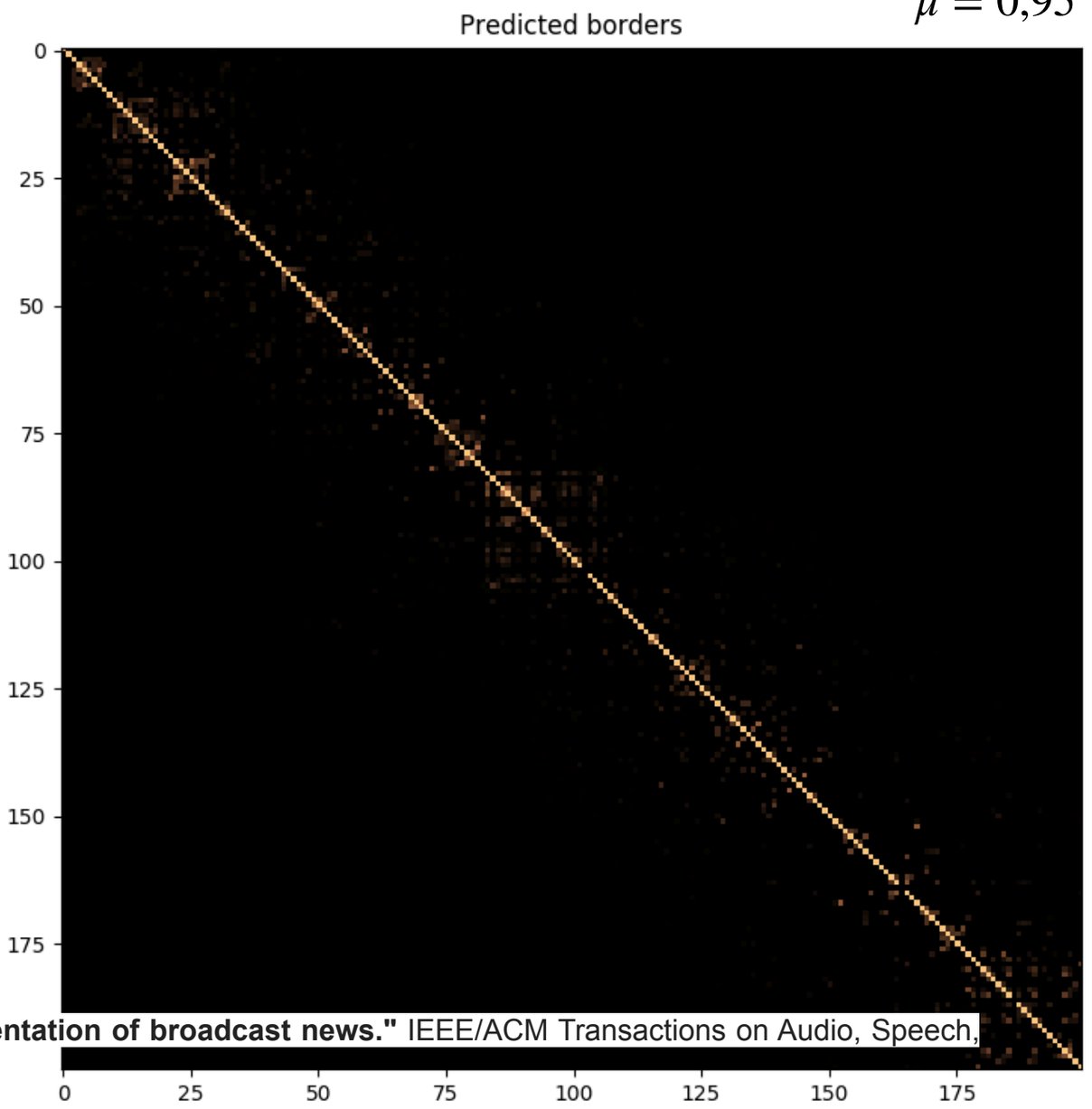
Distance temporelle :

- Réduire le bruit éventuel
- Pénaliser la similitude des phrases hors du voisinage

- Similarité : $\varphi(d_i, d_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i||\mathbf{d}_j|} \cdot \mu^{|i-j|}$

$\mu = 0{,}95$





- Chen, Hongjie, et al. **"Modeling latent topics and temporal distance for story segmentation of broadcast news."** IEEE/ACM Transactions on Audio, Speech, and Language Processing 25.1 (2016): 112-123.

Partitionnement avec MAJORCLUST

Documents dépendants de la temporalité

Laplacian Eigenmaps :

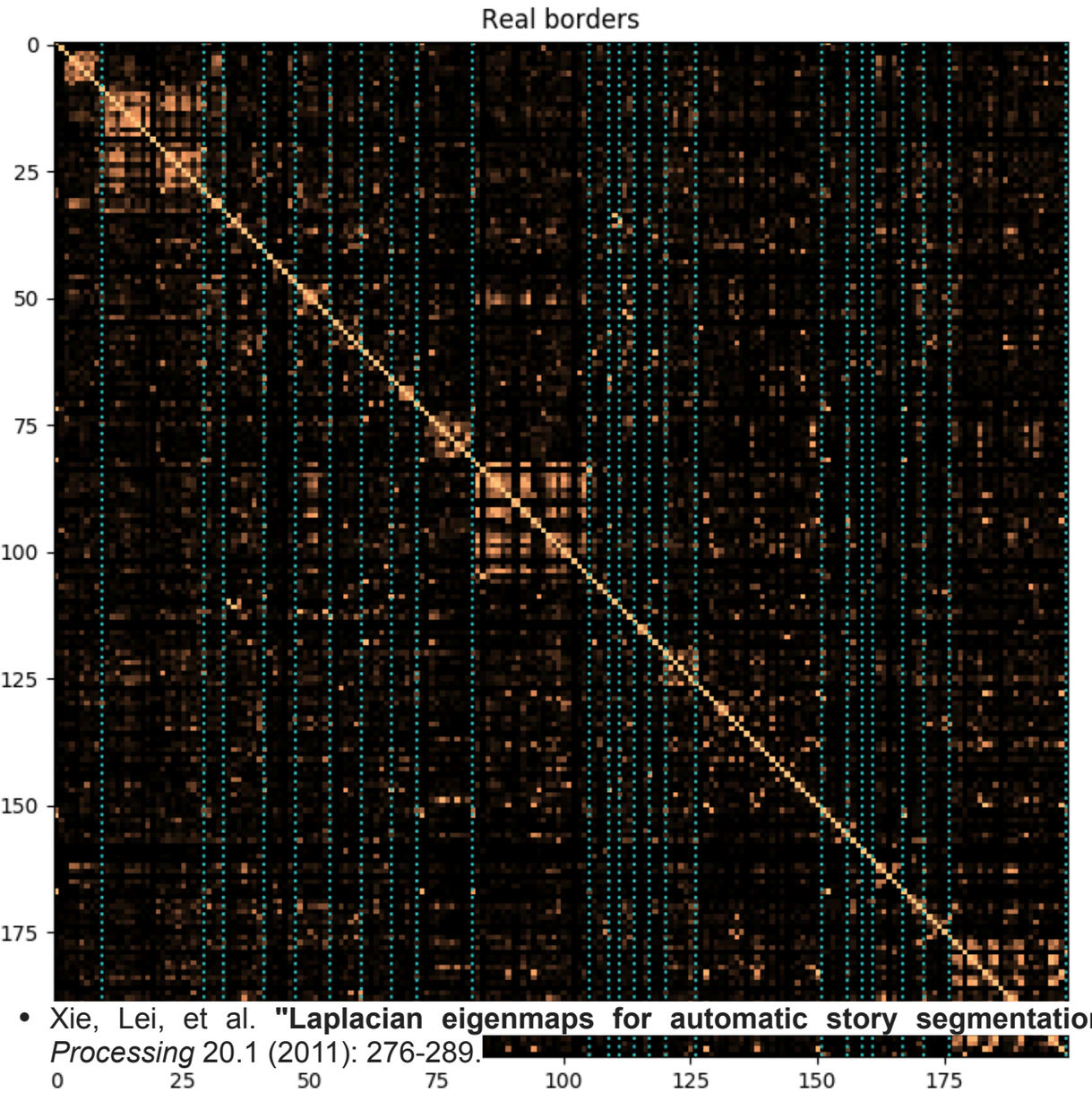
- réduction de la dimensionnalité de la matrice de similarité
- eigen-decomposition
- préserve la structure géométrique des données
- Graph laplacian : **L = C - S**

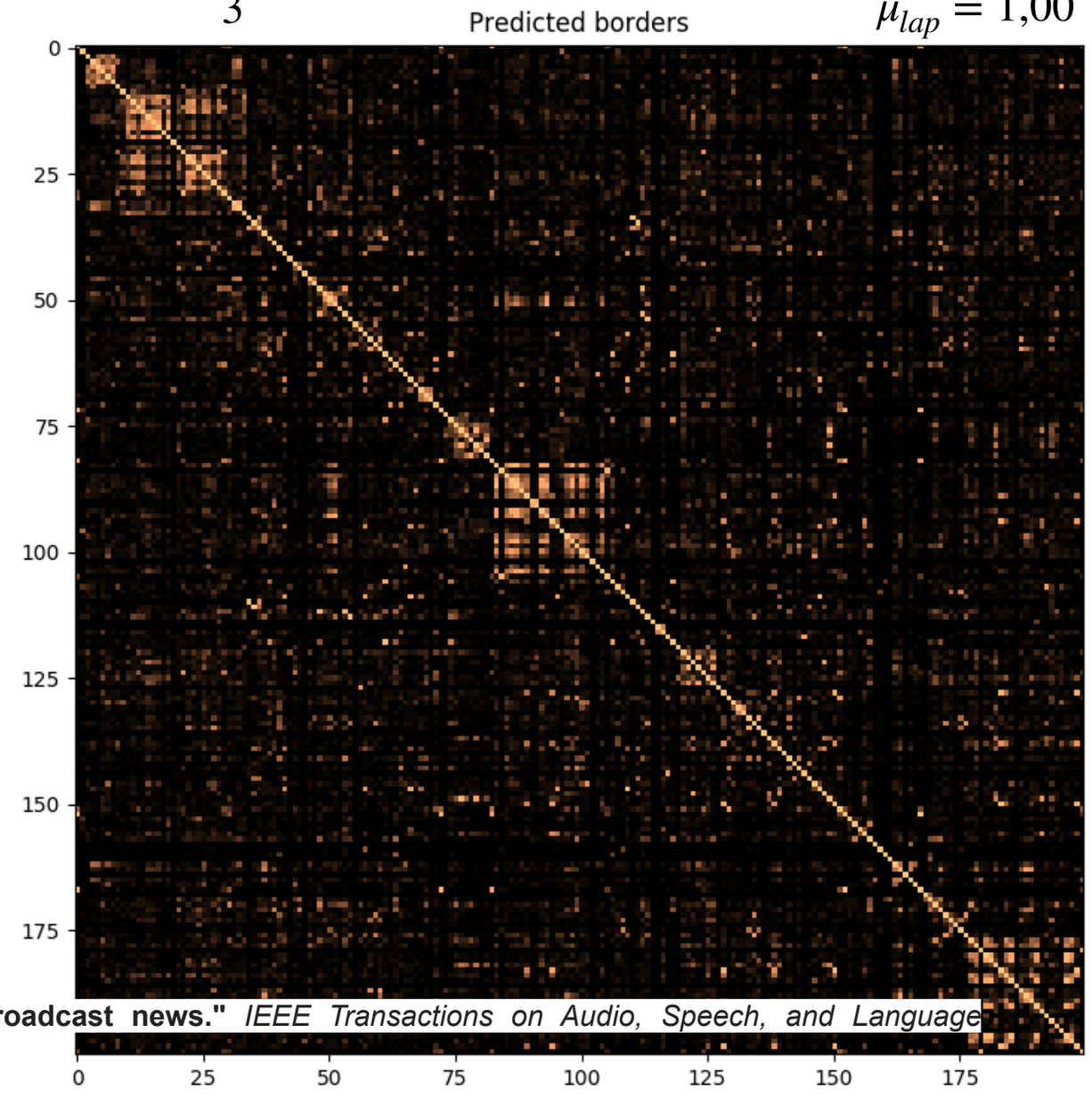
$$c_{i,i} = \sum_{j=1}^{N} s_{i,j}$$

$$\mathfrak{Q} = \frac{1}{2} \sum_{i,j=1}^{N} ||\mathbf{y}_i - \mathbf{y}_j||^2 s_{i,j} = \sum_{q=1}^{Q} \mathbf{f}_q^T \mathbf{L} \mathbf{f}_q$$

$$Q = \frac{|\mathbf{S}|}{3}$$

$\mu = 1{,}00$
$\mu_{lap} = 1{,}00$





- Xie, Lei, et al. **"Laplacian eigenmaps for automatic story segmentation of broadcast news."** *IEEE Transactions on Audio, Speech, and Language Processing* 20.1 (2011): 276-289.

Partitionnement avec MAJORCLUST

Documents dépendants de la temporalité

Laplacian Eigenmaps :

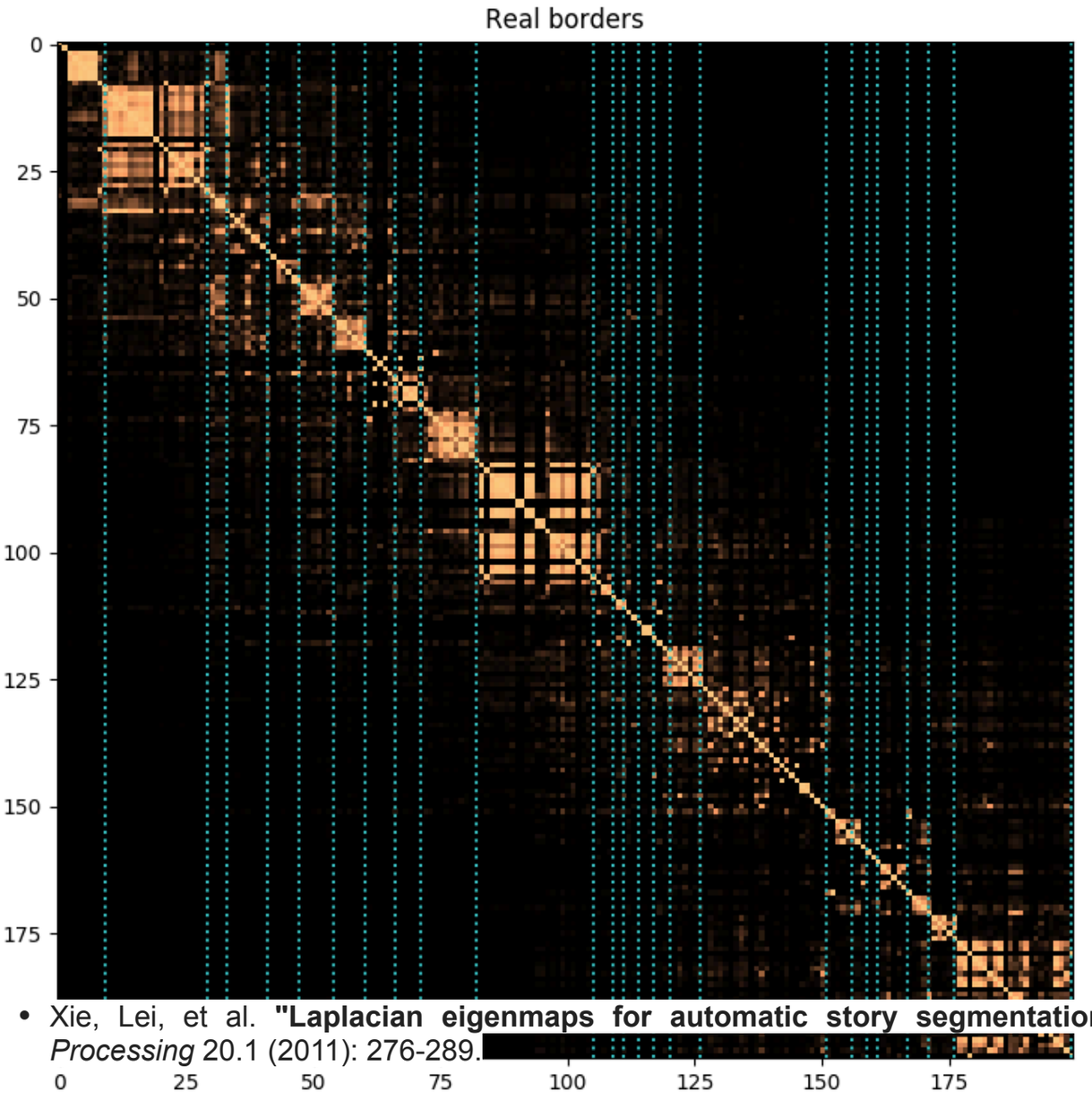
- réduction de la dimensionnalité de la matrice de similarité
- eigen-decomposition
- préserve la structure géométrique des données
- Graph laplacian : **L = C - S**

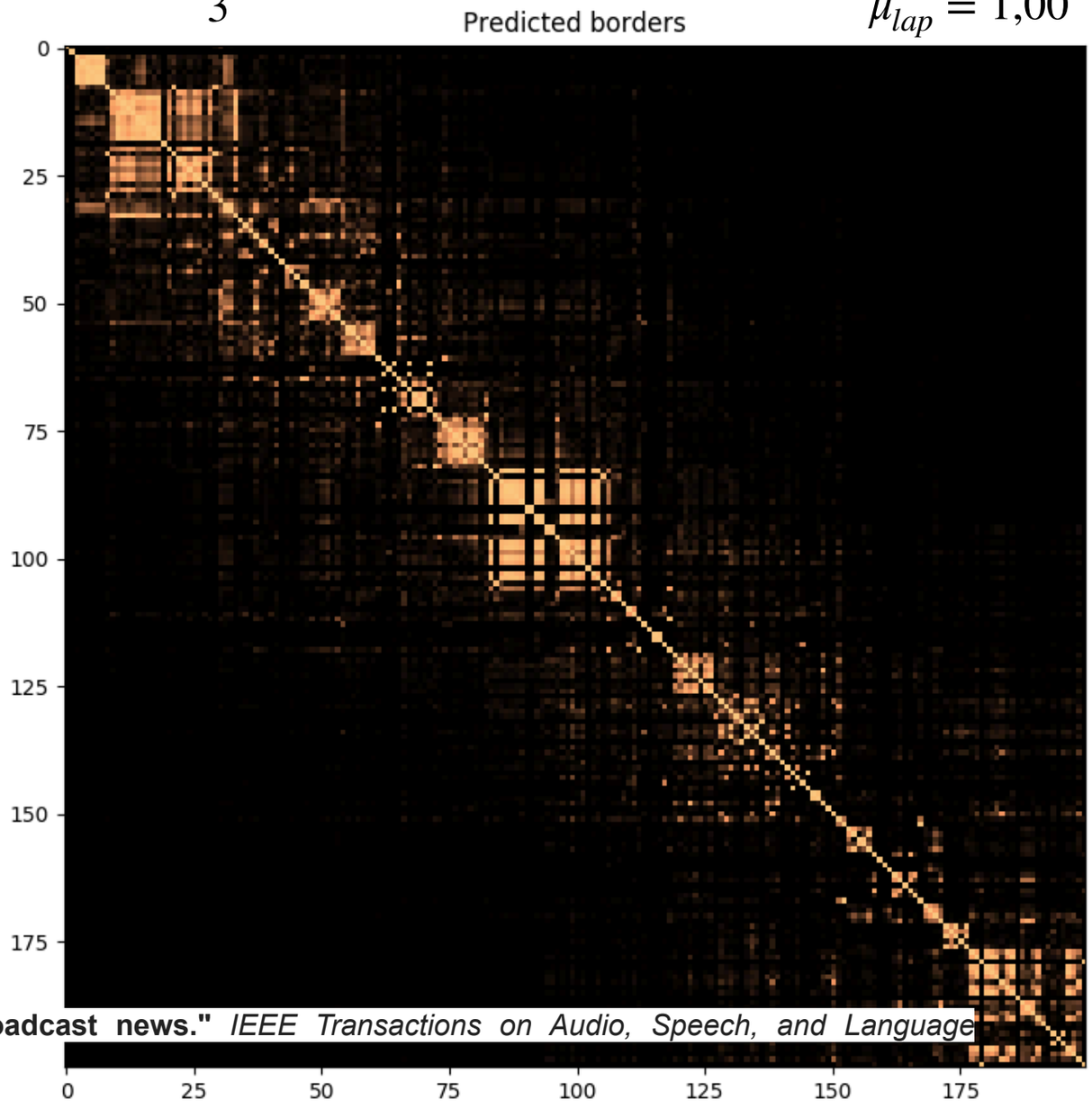
$$c_{i,i} = \sum_{j=1}^{N} s_{i,j}$$

$$\mathfrak{Q} = \frac{1}{2} \sum_{i,j=1}^{N} ||\mathbf{y}_i - \mathbf{y}_j||^2 s_{i,j} = \sum_{q=1}^{Q} \mathbf{f}_q^T \mathbf{L} \mathbf{f}_q$$

$$Q = \frac{|\mathbf{S}|}{3}$$

$\mu = 0{,}95$
$\mu_{lap} = 1{,}00$





- Xie, Lei, et al. **"Laplacian eigenmaps for automatic story segmentation of broadcast news."** *IEEE Transactions on Audio, Speech, and Language Processing* 20.1 (2011): 276-289.

Partitionnement avec MAJORCLUST

Documents dépendants de la temporalité

Laplacian Eigenmaps :

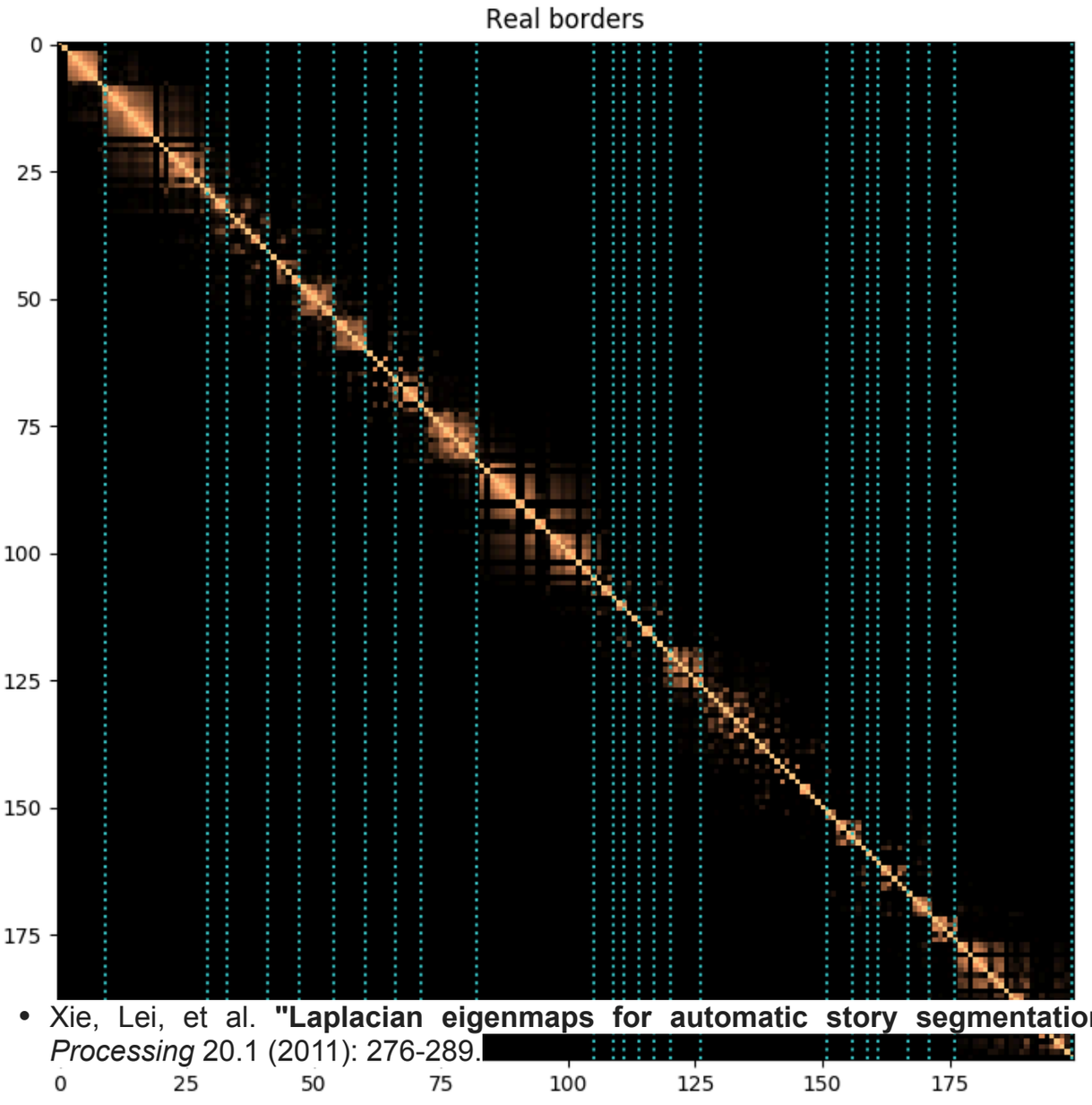
- réduction de la dimensionnalité de la matrice de similarité
- eigen-decomposition
- préserve la structure géométrique des données
- Graph laplacian : **L = C - S**

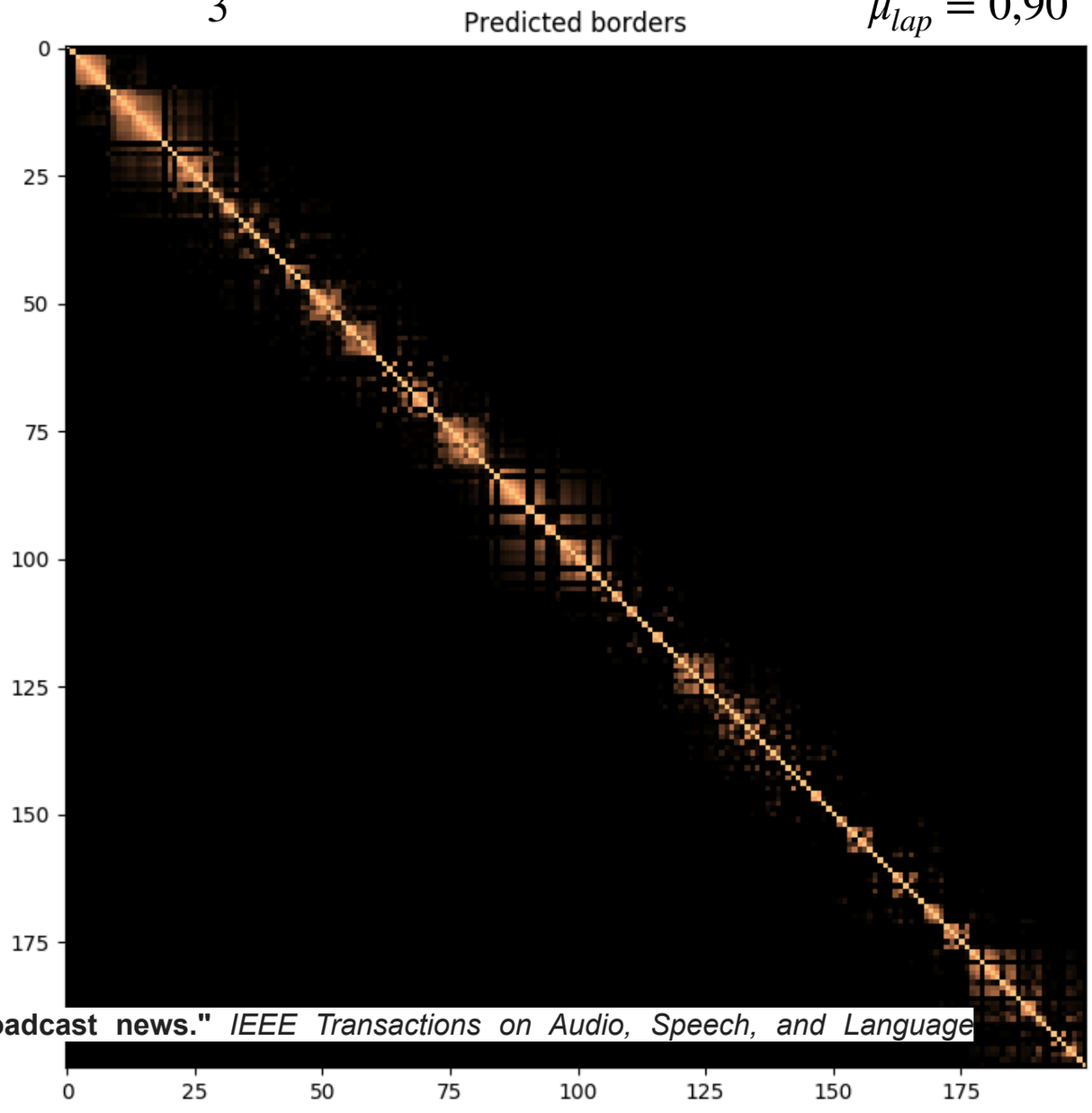
$$c_{i,i} = \sum_{j=1}^{N} s_{i,j}$$

$$\mathfrak{L} = \frac{1}{2} \sum_{i,j=1}^{N} ||\mathbf{y}_i - \mathbf{y}_j||^2 s_{i,j} = \sum_{q=1}^{Q} \mathbf{f}_q^T \mathbf{L} \mathbf{f}_q$$

$$Q = \frac{|\mathbf{S}|}{3}$$

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$



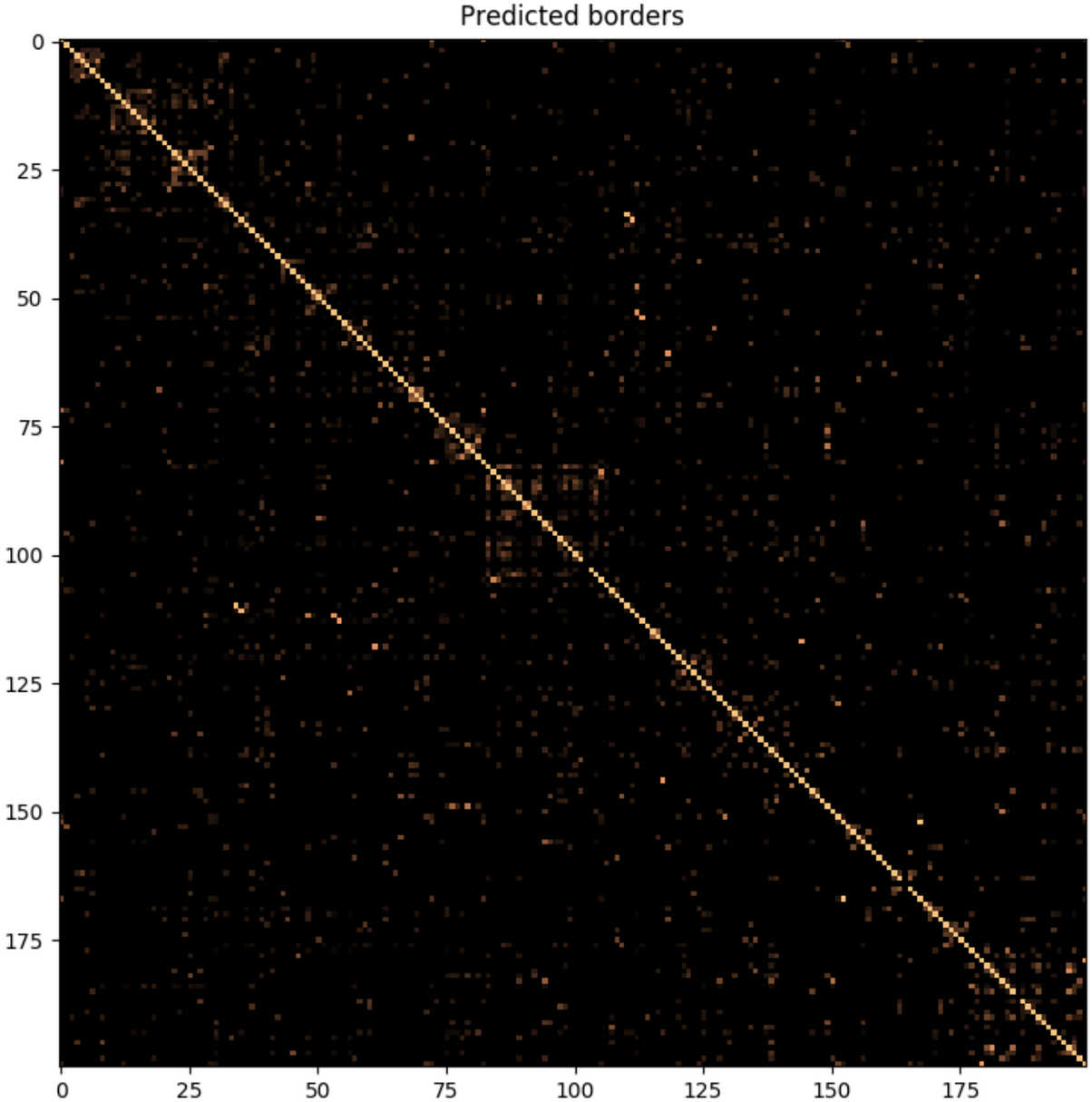


- Xie, Lei, et al. **"Laplacian eigenmaps for automatic story segmentation of broadcast news."** *IEEE Transactions on Audio, Speech, and Language Processing* 20.1 (2011): 276-289.

Partitionnement avec MAJORCLUST

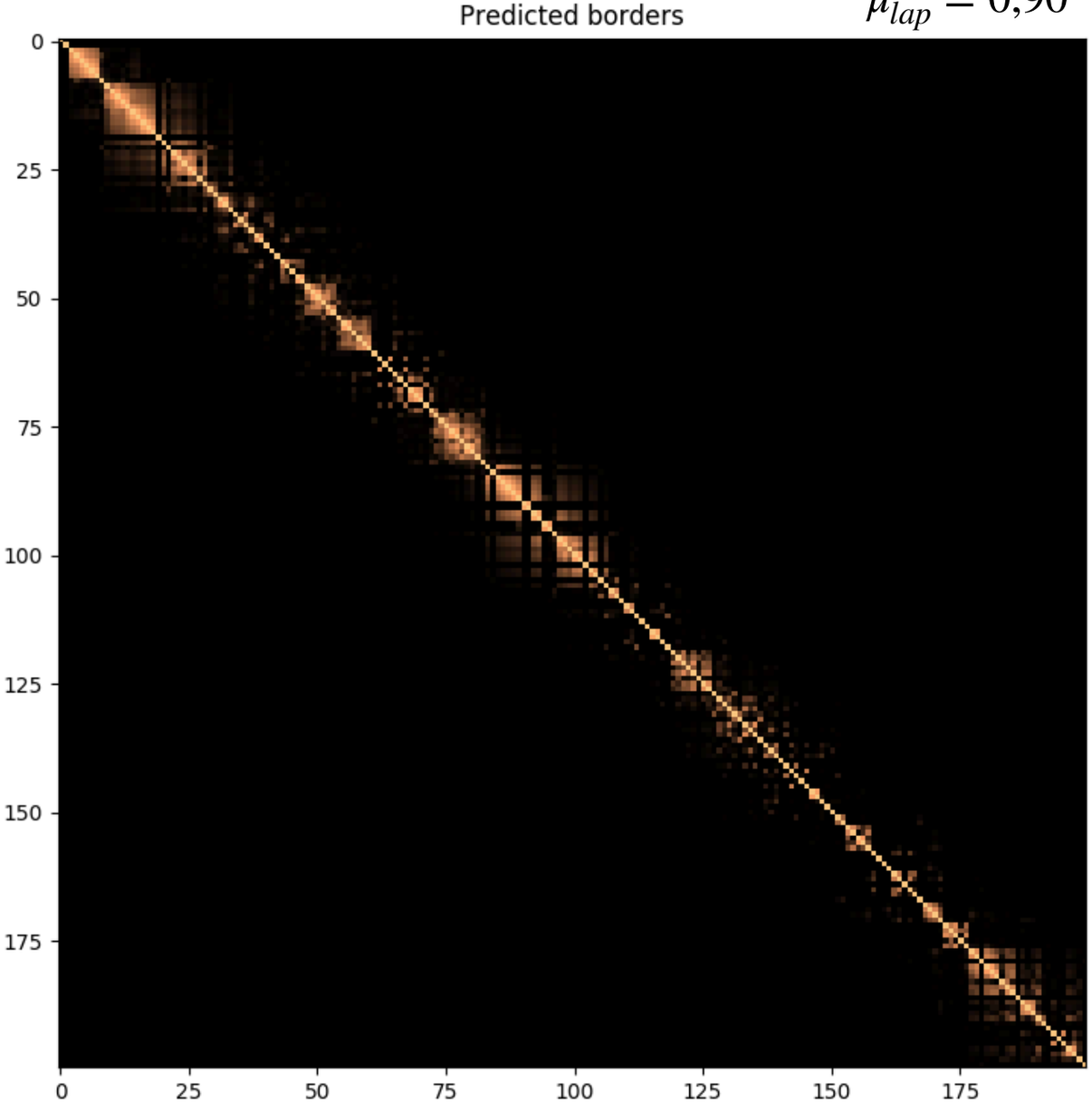
Documents dépendants de la temporalité

Similarité cosinus



Distance temporelle + Laplacian Eigenmaps

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$



Partitionnement avec MAJORCLUST

Exemple de clustering

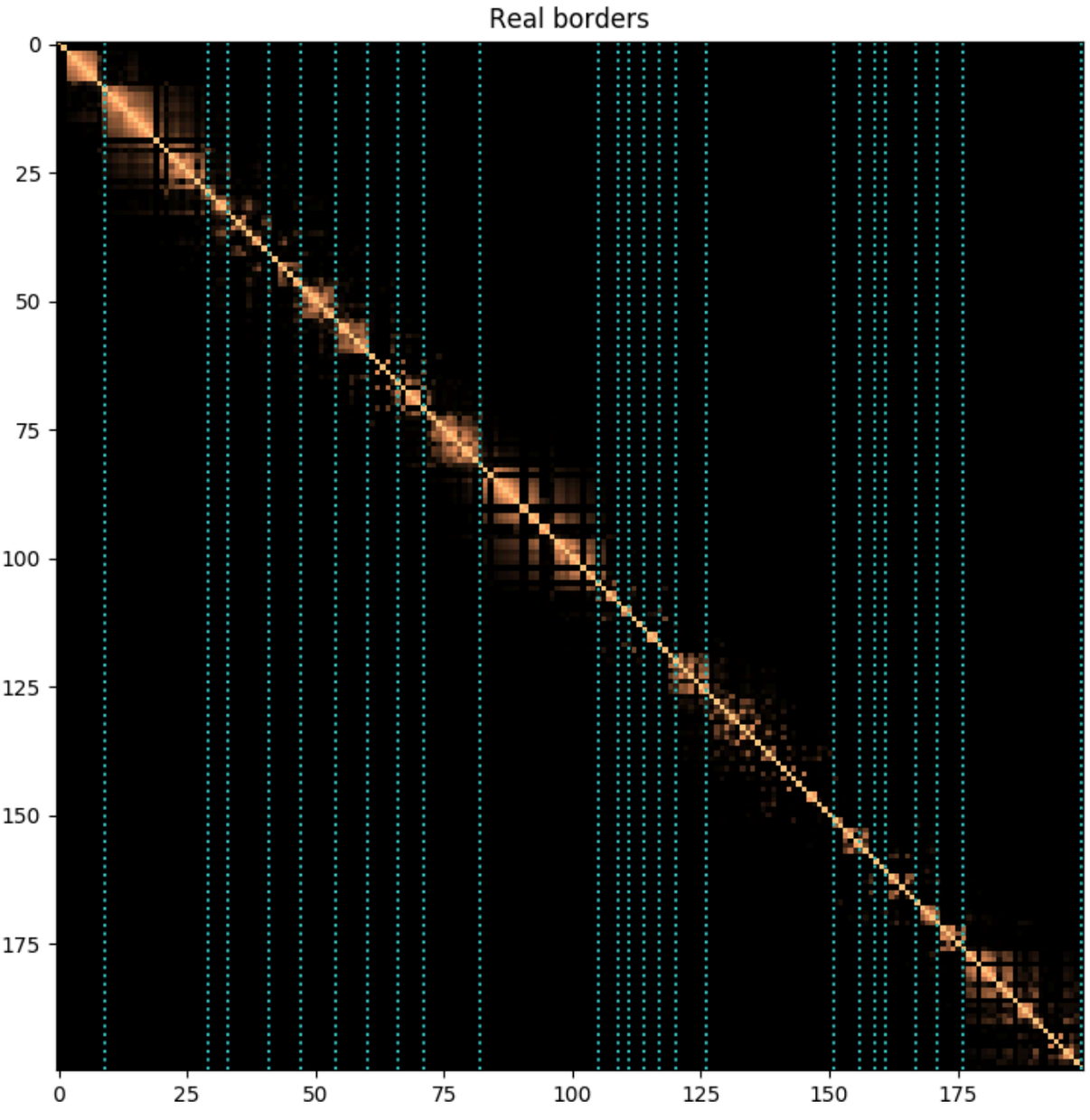
MAJORCLUST :

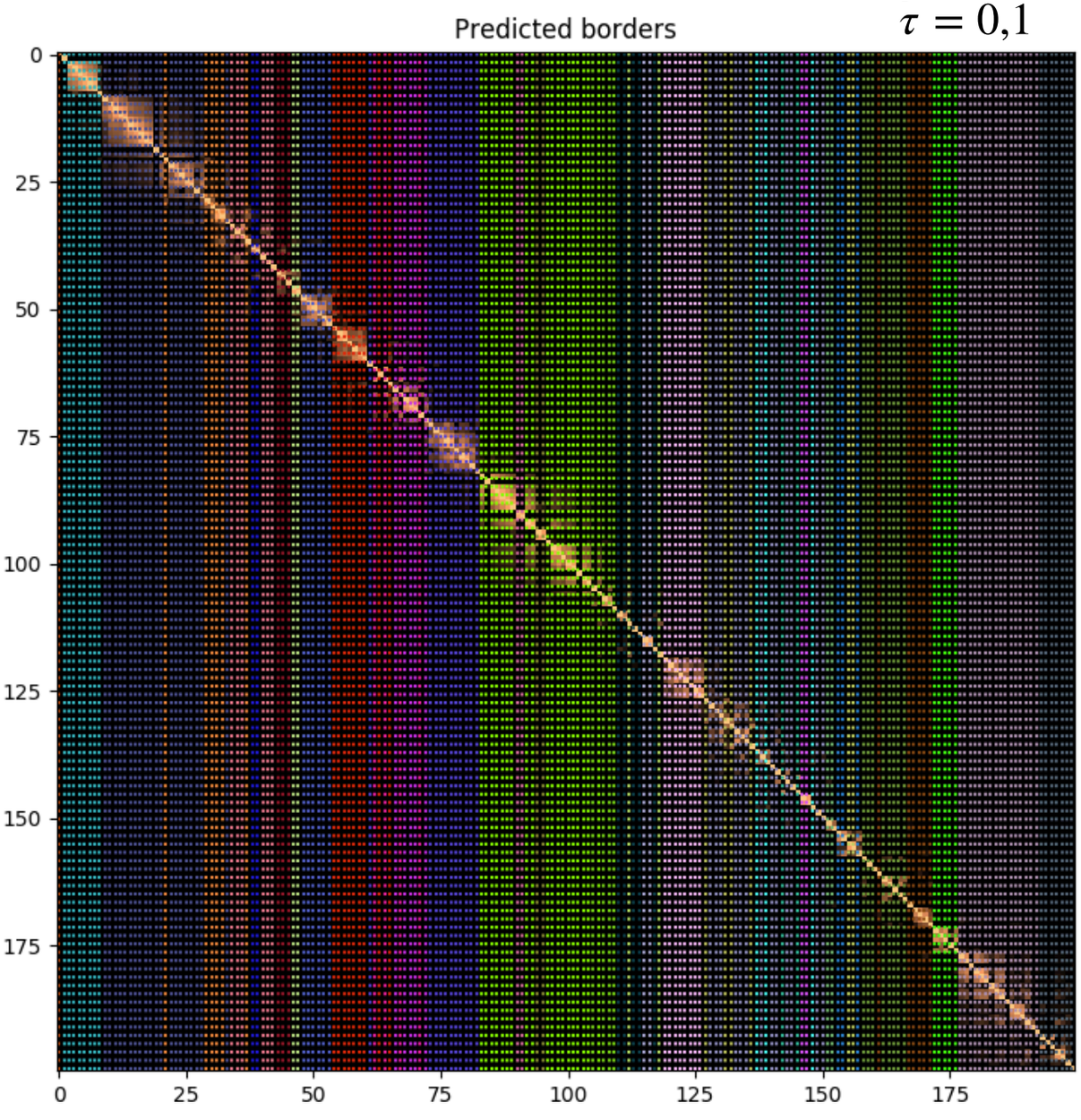
- Considère pas la temporalité des documents
- Exclusion de certain phrases

$$\varphi(d_i, d_j) = \cos(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{|\mathbf{d}_i||\mathbf{d}_j|} \; ; \; \varphi(d_i, d_j) \geq \tau$$

- Nombre de groupes découverts : 33
- Phrases exclues : 20

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$
$\tau = 0{,}1$

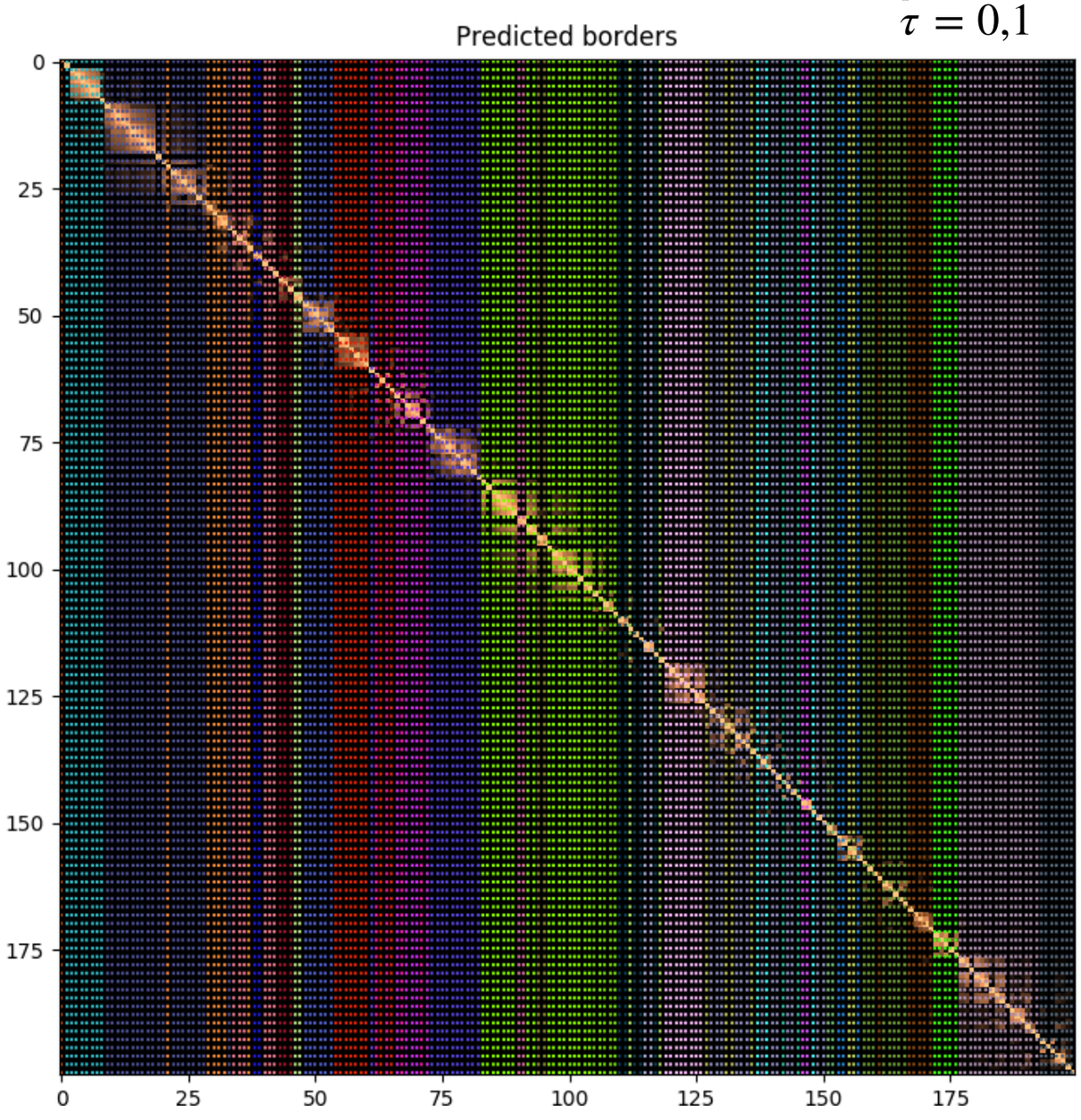




MAJORCLUSTemp

MAJORCLUST dépendant de la temporalité

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$
$\tau = 0{,}1$



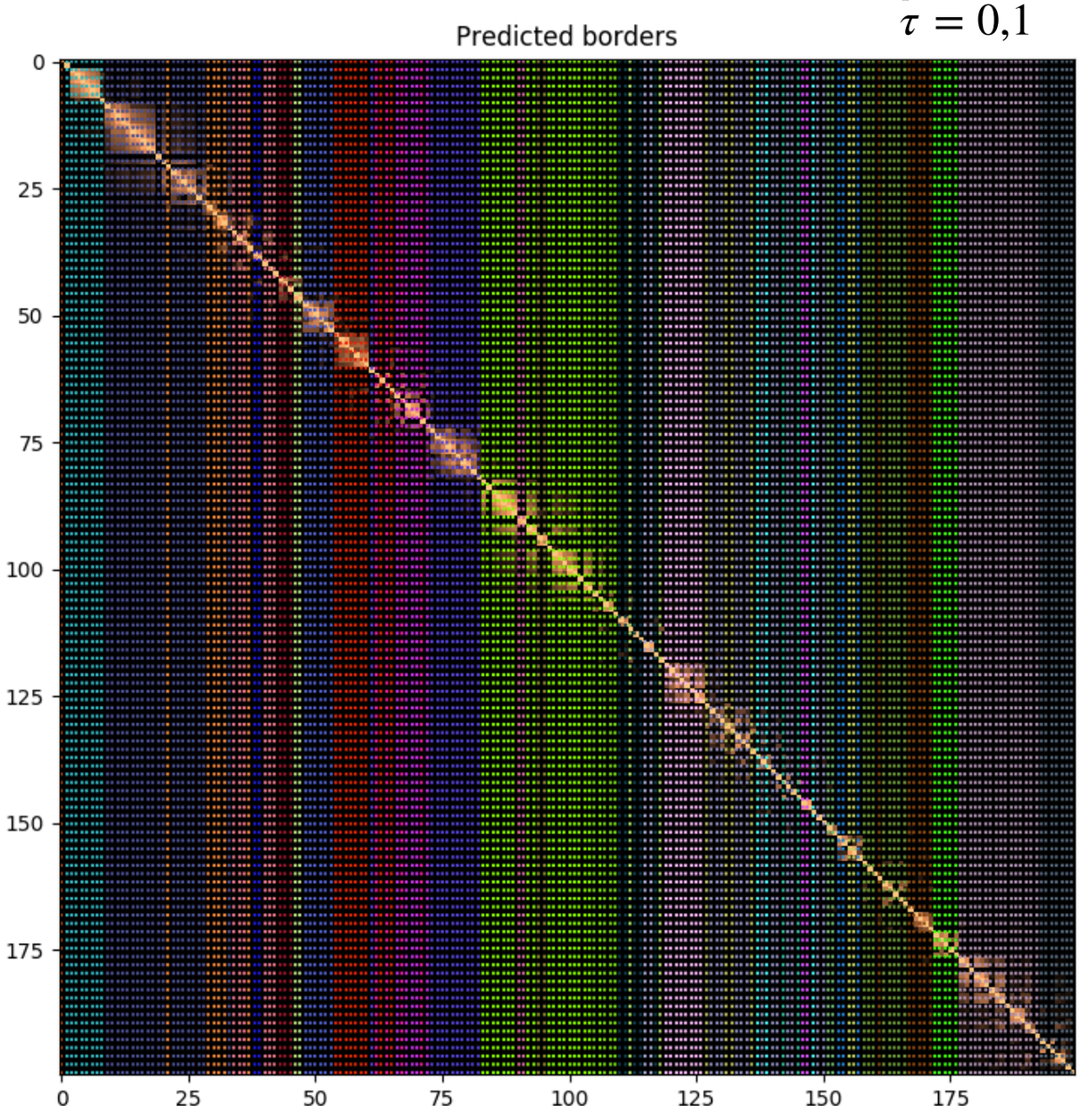
MAJORCLUSTemp:

1. Partitionner avec MAJORCLUST
2. Créer des intervalles à partir de clusters
3. Récupérer les phrases exclues avec la matrice de similarité
4. Placer dans les intervalles les phrases récupérées
5. Déplacer les intervalles isolés
6. Fusionner les clusters

MAJORCLUSTemp

MAJORCLUST dépendant de la temporalité

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$
$\tau = 0{,}1$



MAJORCLUSTemp:

1. Partitionner avec MAJORCLUST
2. Créer des intervalles à partir de clusters
3. Récupérer les phrases exclues avec la matrice de similarité
4. Placer dans les intervalles les phrases récupérées
5. Déplacer les intervalles isolés
6. Fusionner les clusters

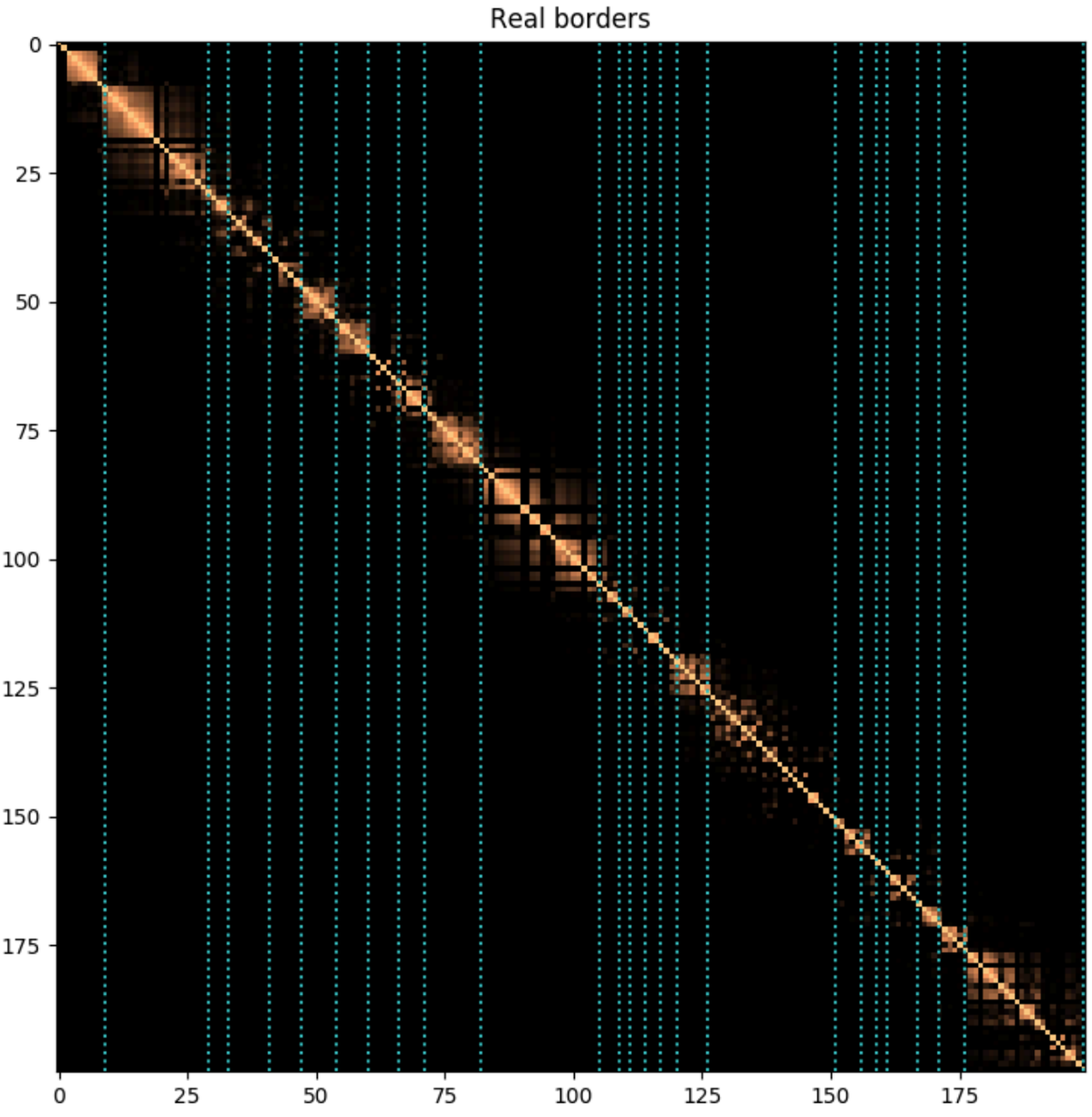
MAJORCLUSTemp

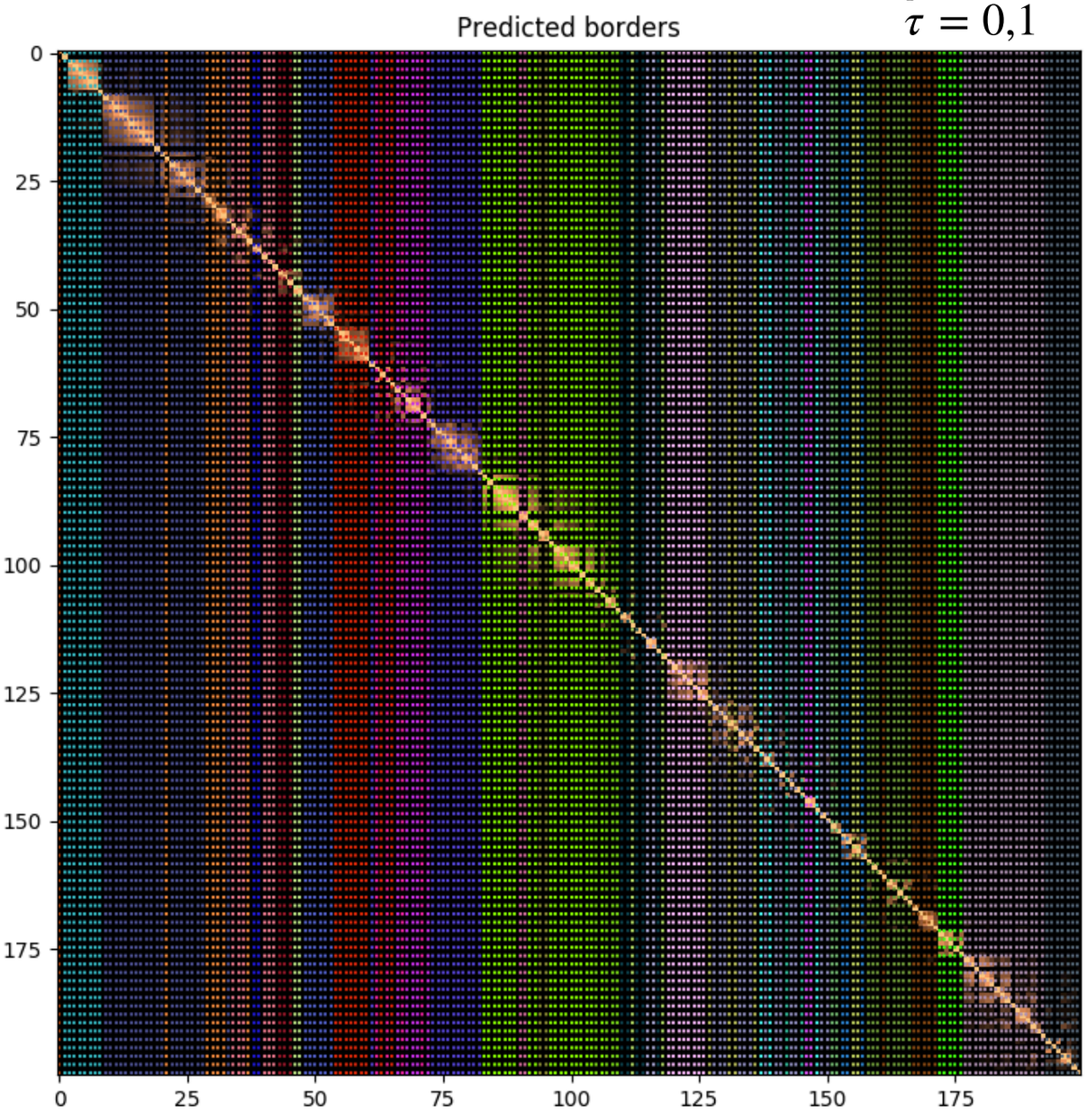
MAJORCLUST dépendant de la temporalité

MAJORCLUSTemp:

1. Partitionner avec MAJORCLUST
2. Créer des intervalles à partir de clusters
3. Récupérer les phrases exclues avec la matrice de similarité
4. Placer dans les intervalles les phrases récupérées
5. Déplacer les intervalles isolés
6. Fusionner les clusters

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$
$\tau = 0{,}1$





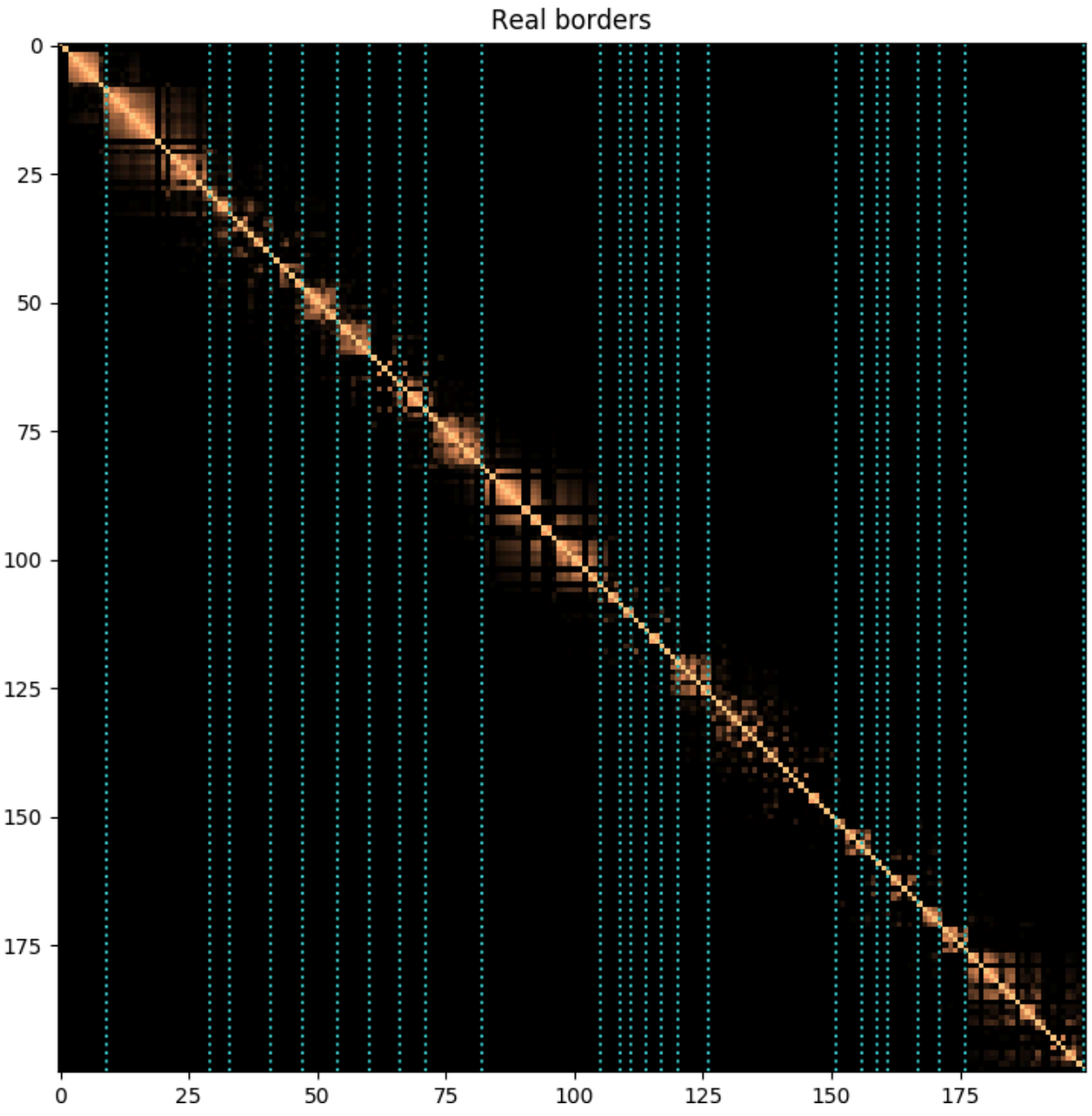
MAJORCLUSTemp

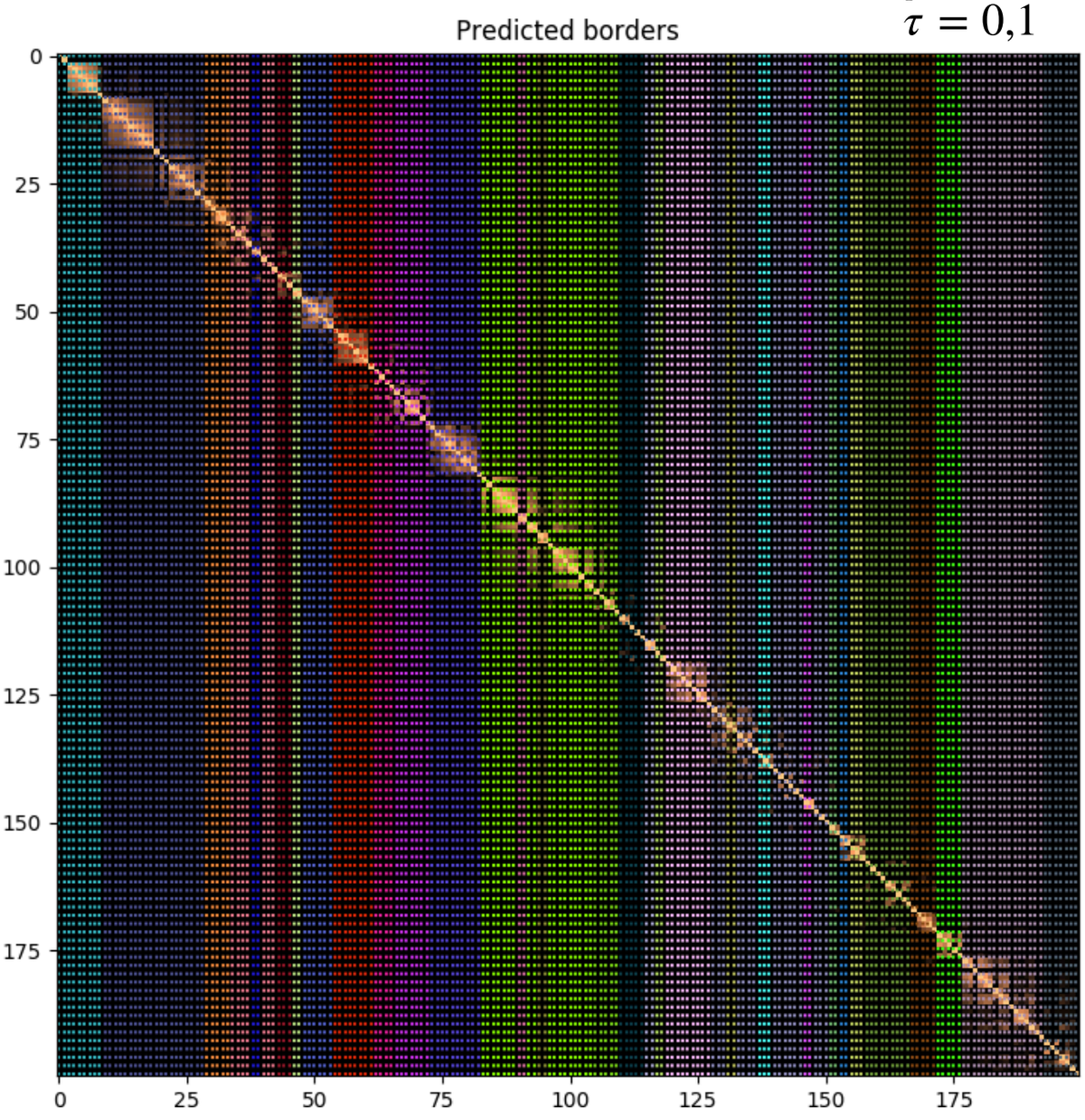
MAJORCLUST dépendant de la temporalité

MAJORCLUSTemp:

1. Partitionner avec MAJORCLUST
2. Créer des intervalles à partir de clusters
3. Récupérer les phrases exclues avec la matrice de similarité
4. Placer dans les intervalles les phrases récupérées
5. Déplacer les intervalles isolés
6. Fusionner les clusters

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$
$\tau = 0{,}1$





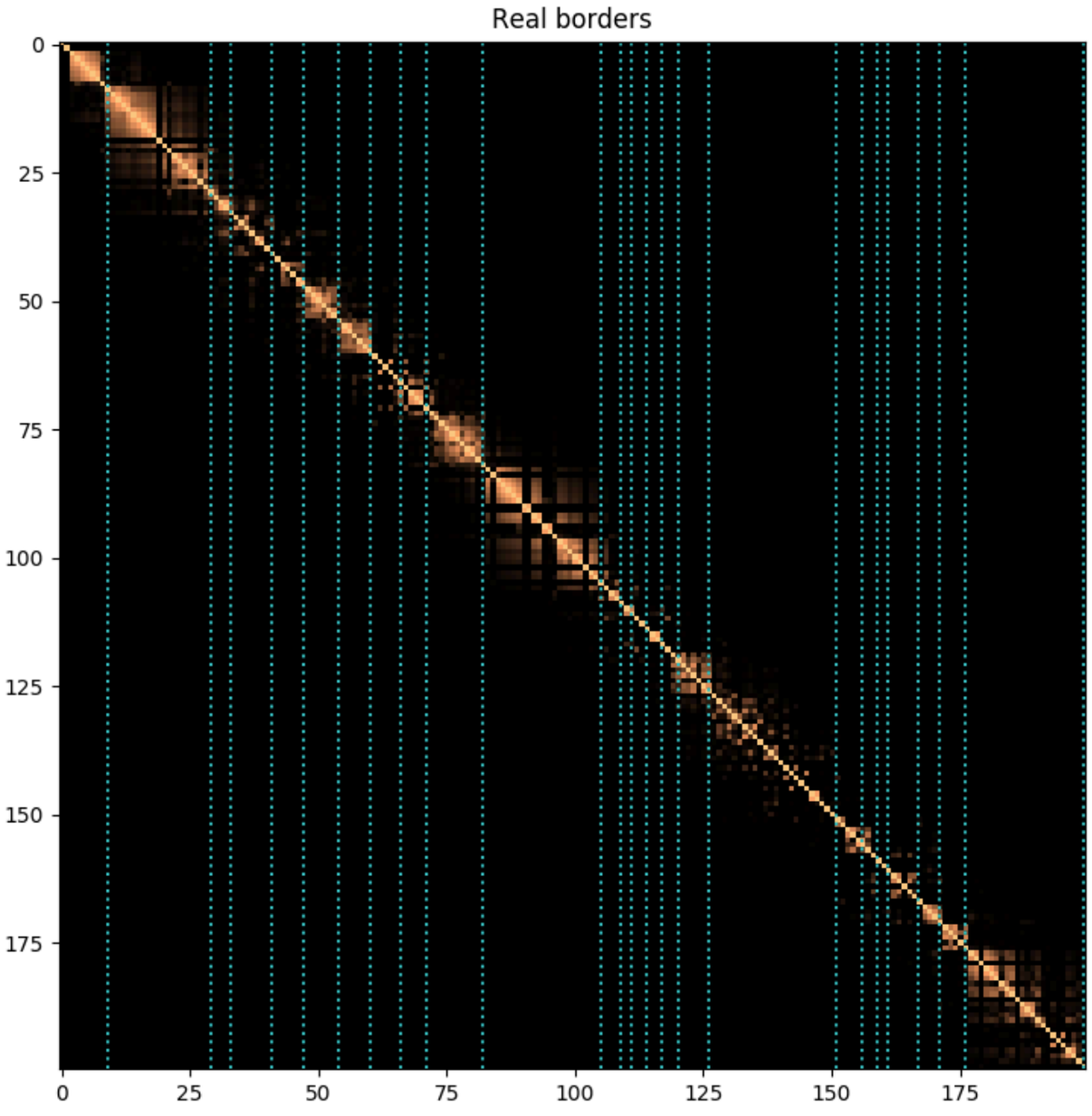
MAJORCLUSTemp

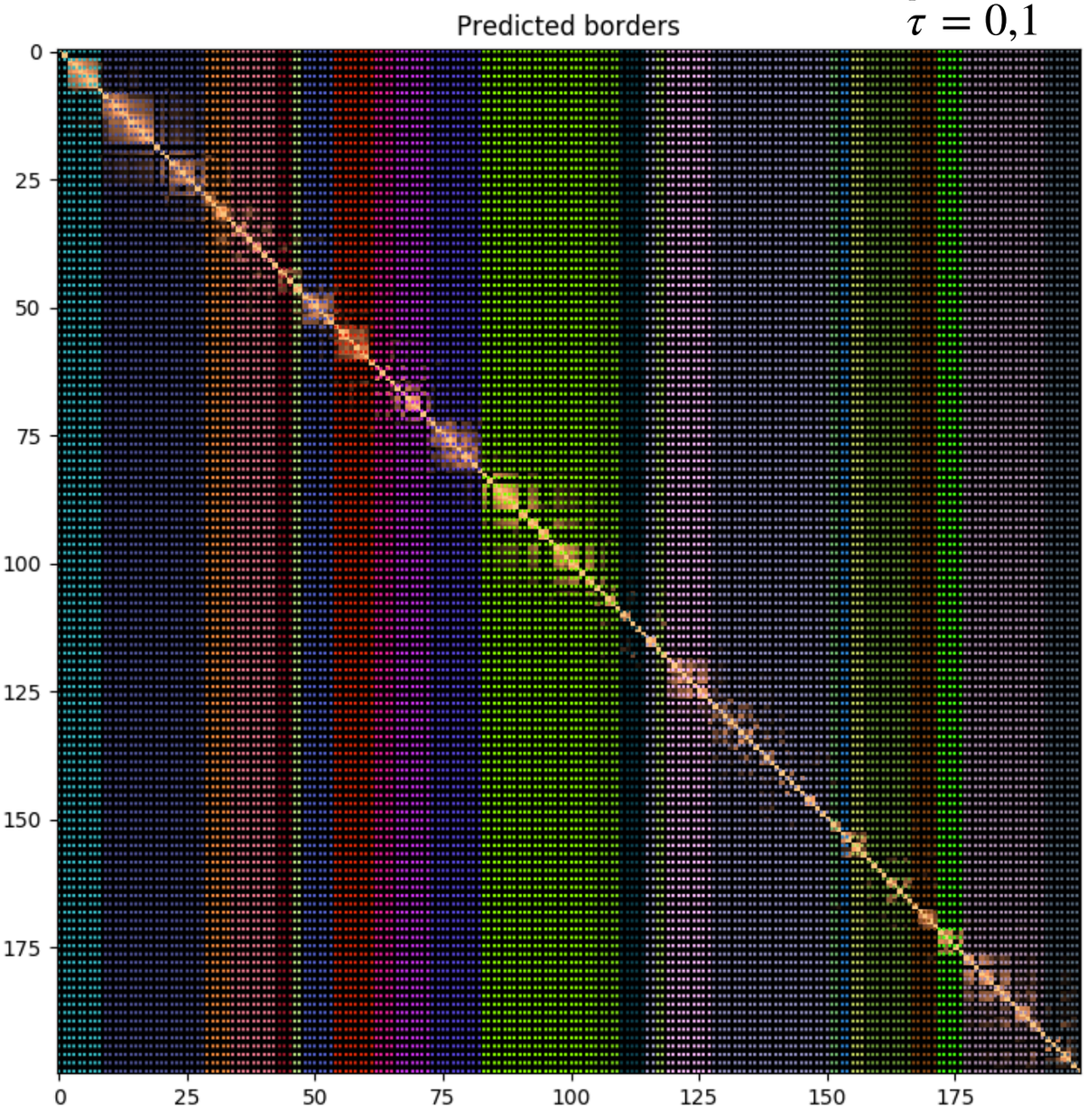
MAJORCLUST dépendant de la temporalité

MAJORCLUSTemp:

1. Partitionner avec MAJORCLUST
2. Créer des intervalles à partir de clusters
3. Récupérer les phrases exclues avec la matrice de similarité
4. Placer dans les intervalles les phrases récupérées
5. Déplacer les intervalles isolés
6. Fusionner les clusters

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$
$\tau = 0{,}1$





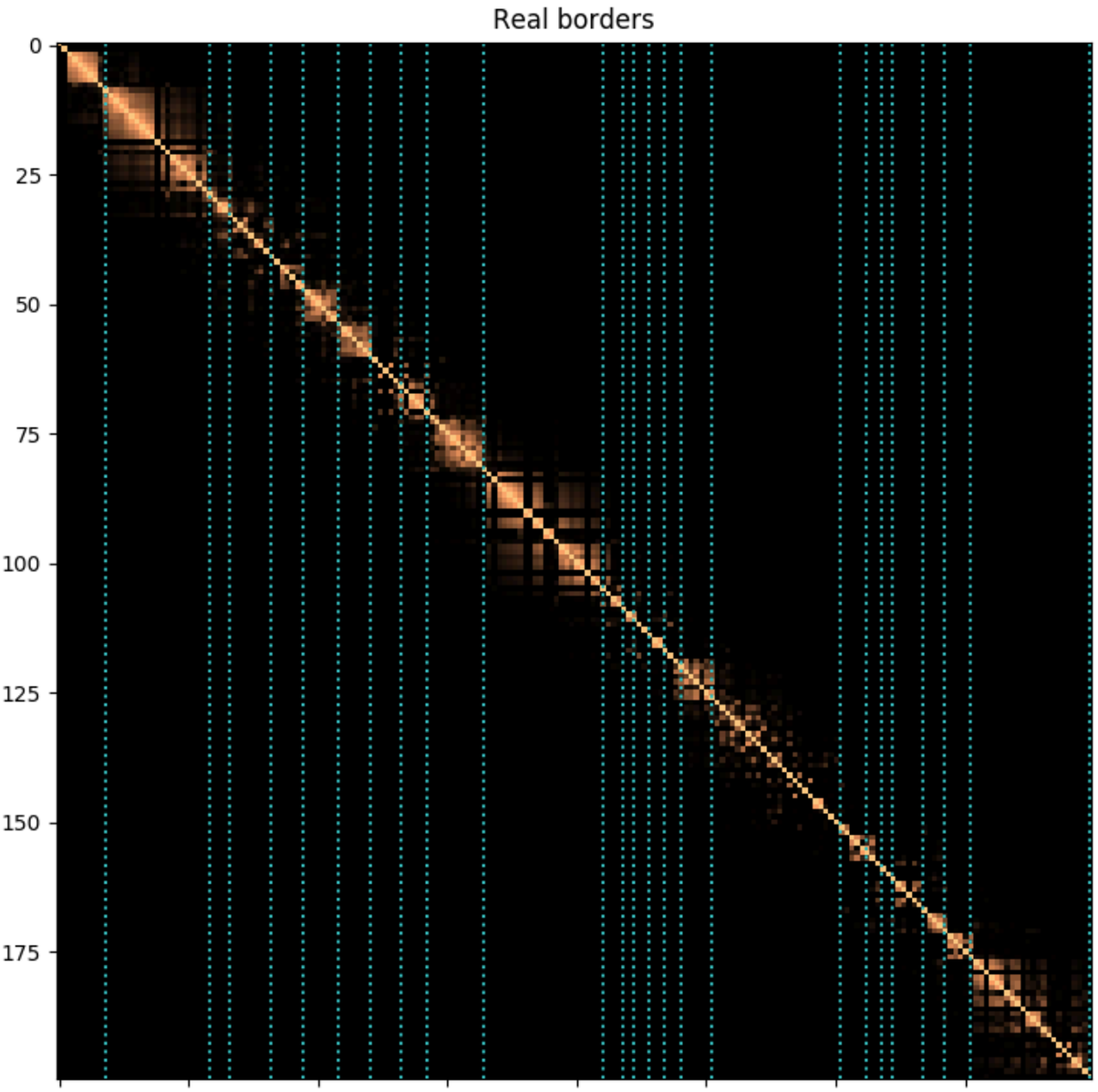
MAJORCLUSTemp

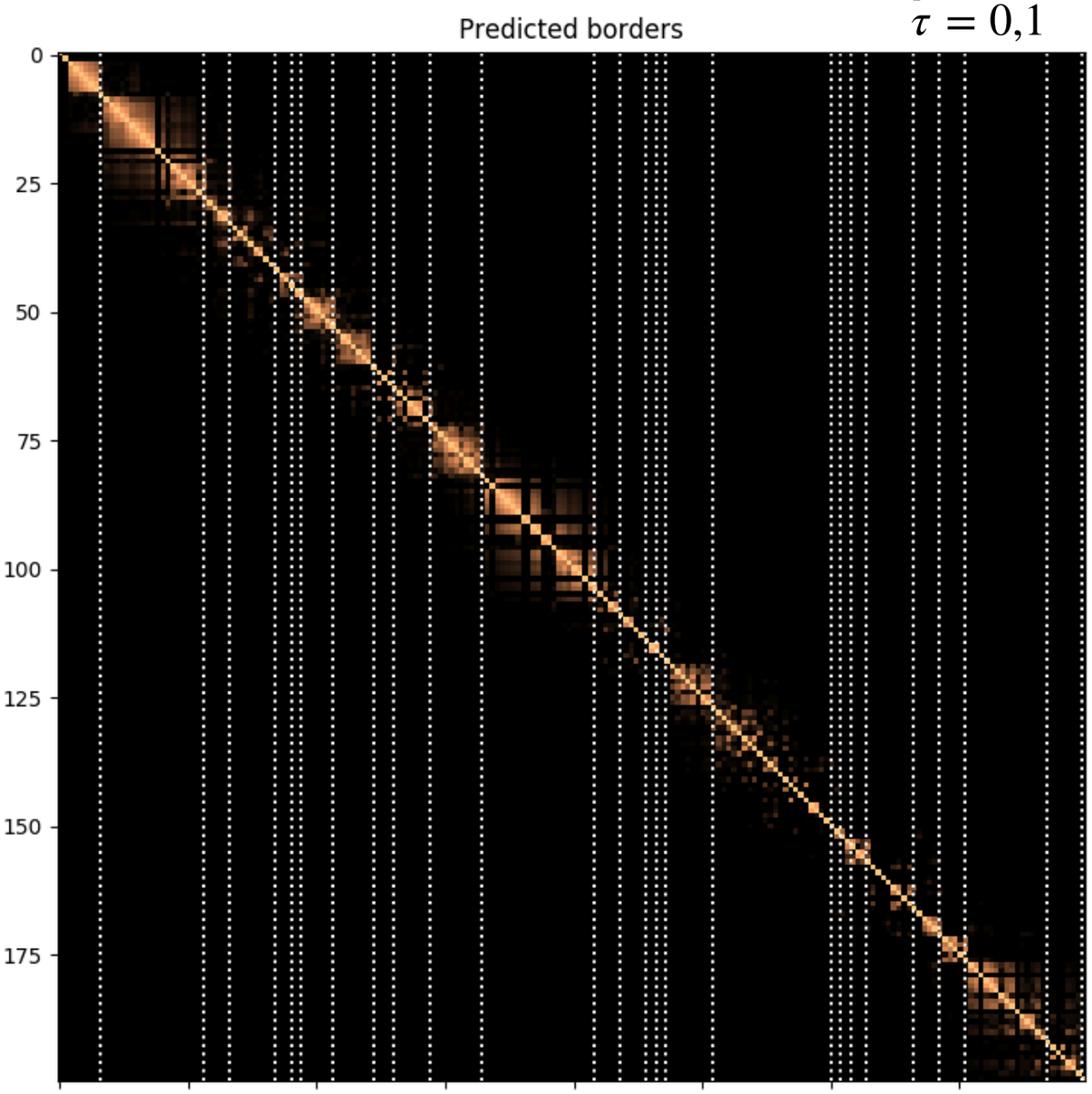
MAJORCLUST dépendant de la temporalité

MAJORCLUSTemp:

1. Partitionner avec MAJORCLUST
2. Créer des intervalles à partir de clusters
3. Récupérer les phrases exclues avec la matrice de similarité
4. Placer dans les intervalles les phrases récupérées
5. Déplacer les intervalles isolés
6. Fusionner les clusters

$\mu = 0{,}95$
$\mu_{lap} = 0{,}90$
$\tau = 0{,}1$





Expériences

Corpus :

- TDT2 English broadcast news corpus
- Transcriptions : manuelles
- Ensemble de 396 documents (Actualités)
  - ~ 283 phrases / document
  - ~ 27 sujets / document
  - ~ 11 phrases / sujet
- Éléments grammaticaux :
  - verbes
  - noms
  - adjectifs
  - adverbes

EX1 :

- 49 configurations

$$\mu = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$$

$$\tau = \{0.0, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$$

EX2 :

- 343 configurations

$$\mu = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$$

$$\mu_{lap} = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$$

$$\tau = \{0.0, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$$

MAJORCLUSTemp

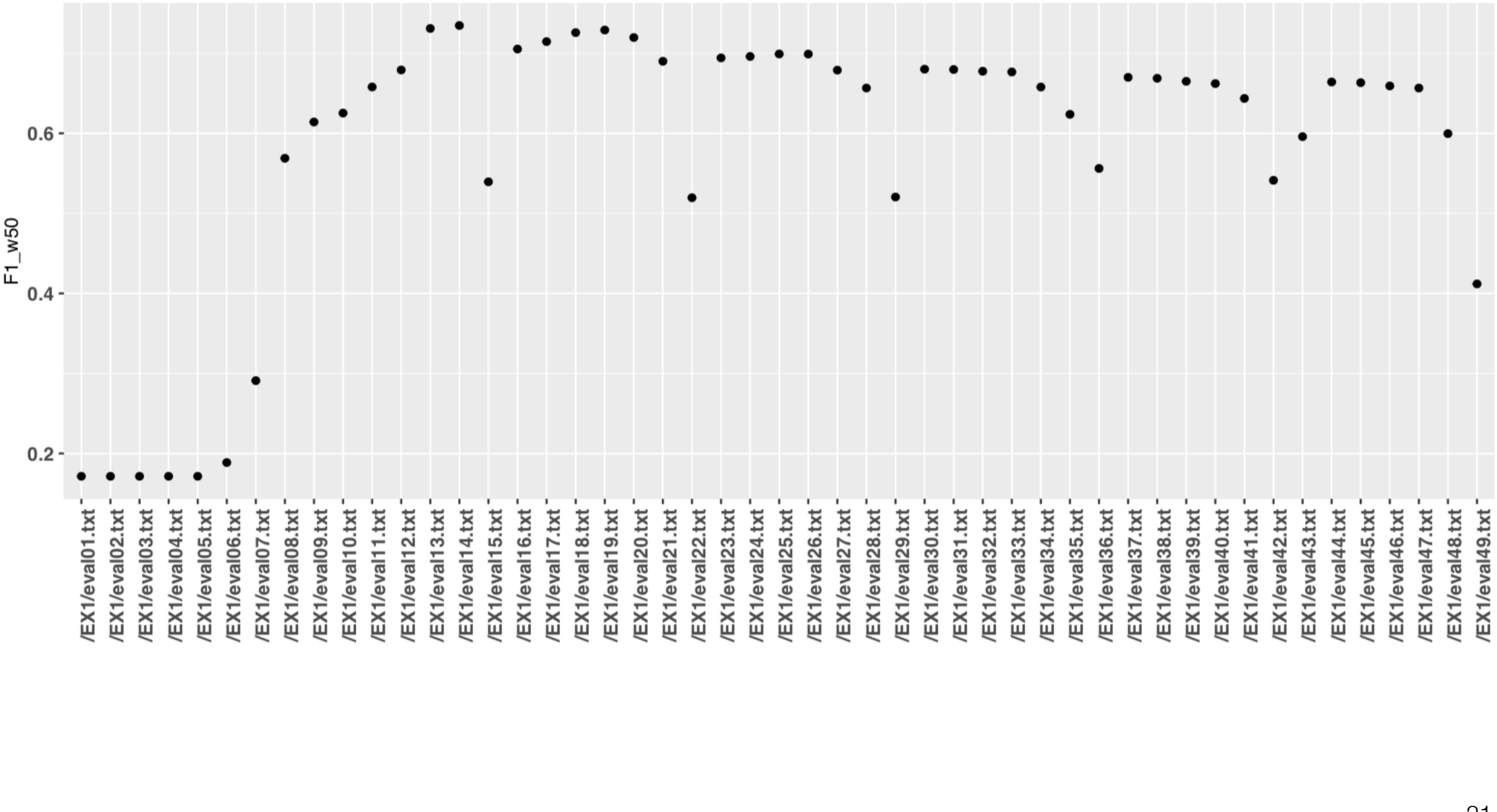
Résultats

EX1 : 49 configurations

$\mu = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$

$\tau = \{0.0, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$

	μ	$\mathcal{T}$	P	R	F-score
eval14	0,95	0,1	0,646	0,897	0,735
eval13	0,95	0,05	0,701	0,810	0,731
eval19	0,9	0,01	0,630	0,908	0,729



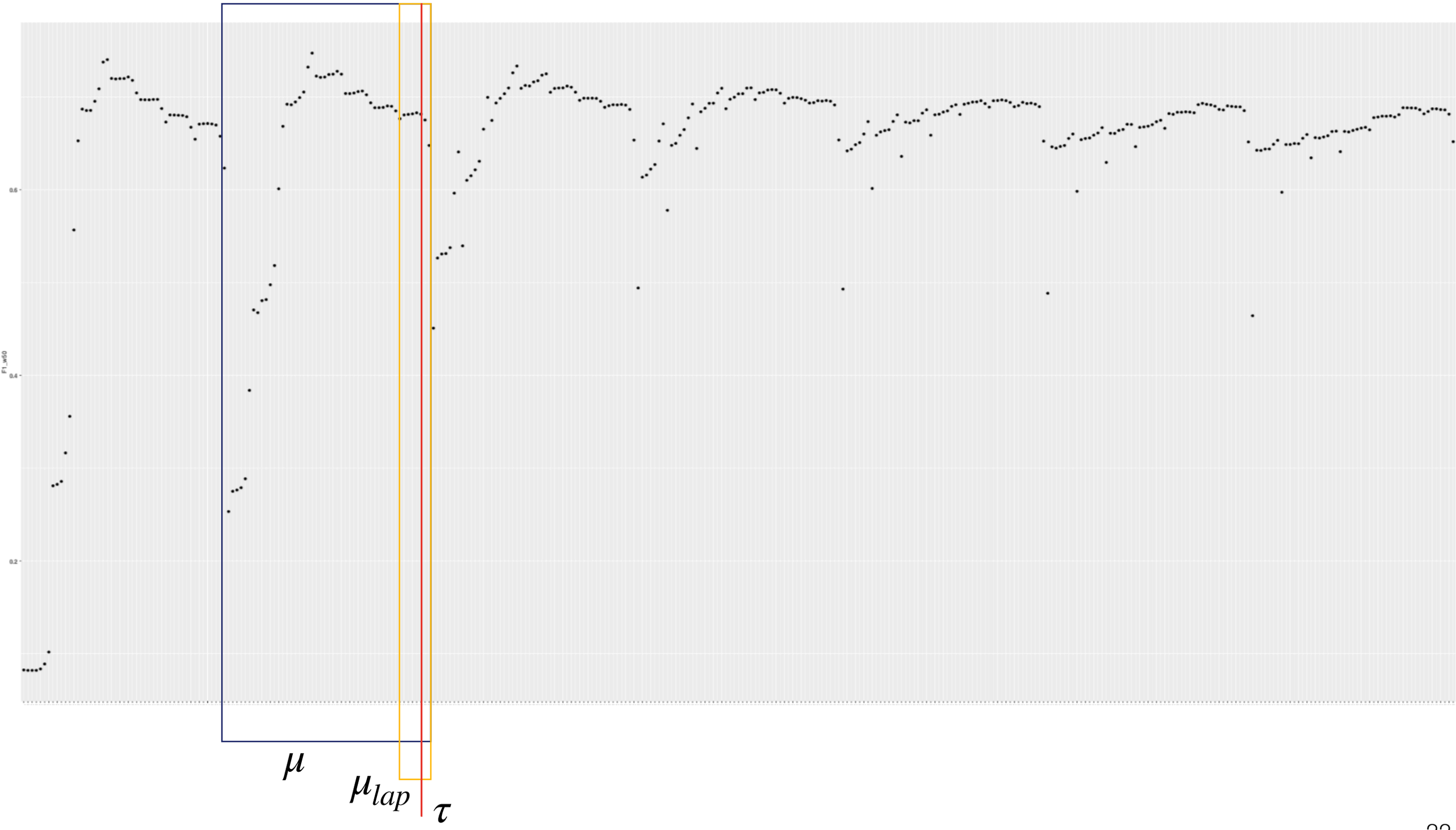
MAJORCLUSTemp

EX2 : 343 configurations

$\mu = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$

$\mu_{lap} = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$

$\tau = \{0.0, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$



MAJORCLUSTemp

Résultats

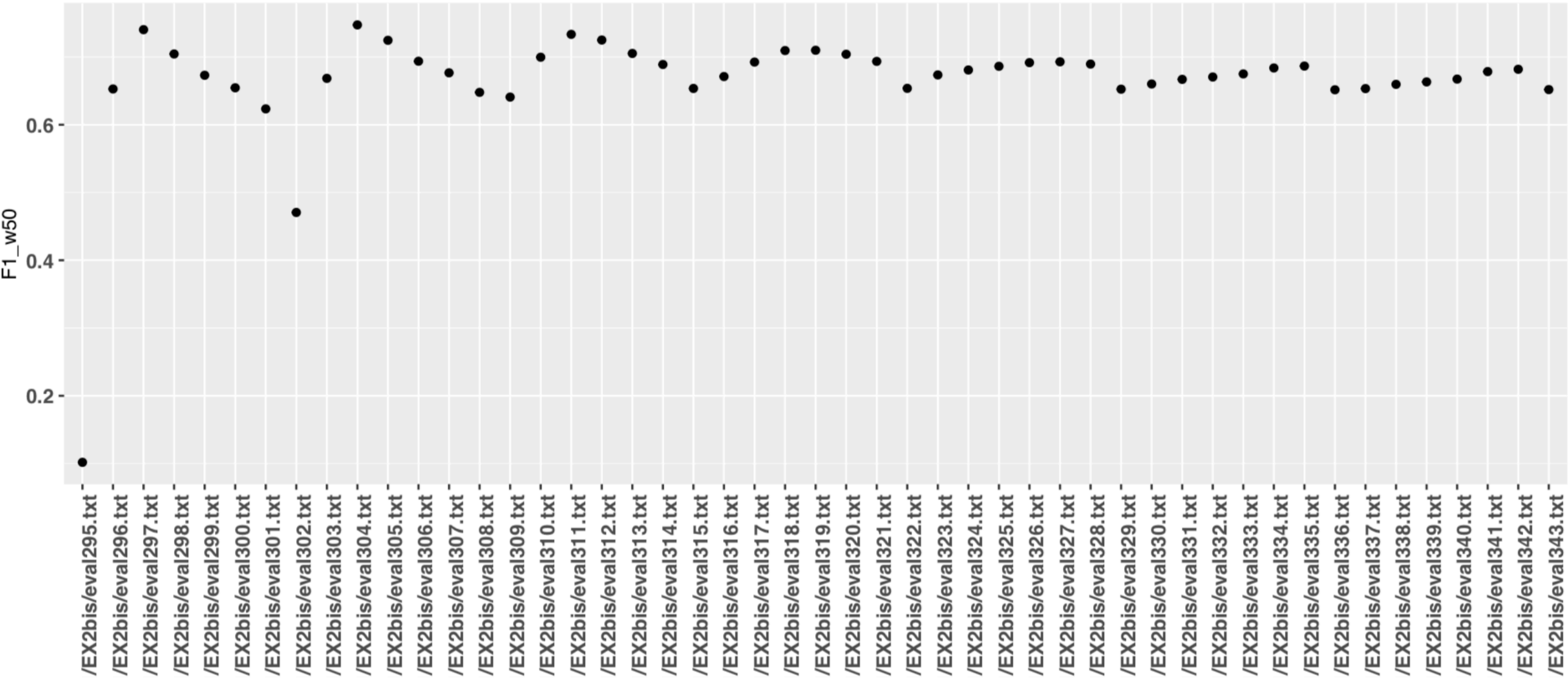
EX2 : 343 configurations

$\mu = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$

$\mu_{lap} = \{1, 0.95, 0.90, 0.85, 0.80, 0.75, 0.70\}$

$\tau = \{0.0, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$

	μ	μ_{lap}	$\mathcal{T}$	P	R	F-score
eval304	0,95	0,90	0,1	0,732	0,797	0,747
eval297	1,00	0,90	0,01	0,661	0,880	0,740
eval248	1,00	0,90	0,05	0,696	0,817	0,738



Conclusions

ACTUALITÉS / TÉLÉREPORTAGES

?

résumé + RI

multimédia & multilingue

carlos-emiliano.gonzalez-gallardo@sorbonne-universite.fr