Powered Dirichlet-Hawkes Process pour clustering temporel de textes

G. Poux-Médard, J. Velcin, S. Loudcher

Gaël Poux-Médard

Université de Lyon, France
Lyon 2, ERIC UR 3083

Janvier 2022

Introduction

- Chaque minute :
  - 400h de video
  - 350 000 tweets
  - 500 000 commentaires
  - 4 200 000 recherches

Introduction

- Chaque minute :
  - 400h de video
  - 350 000 tweets
  - 500 000 commentaires
  - 4 200 000 recherches
- Comment rendre *ceci* intelligible ?

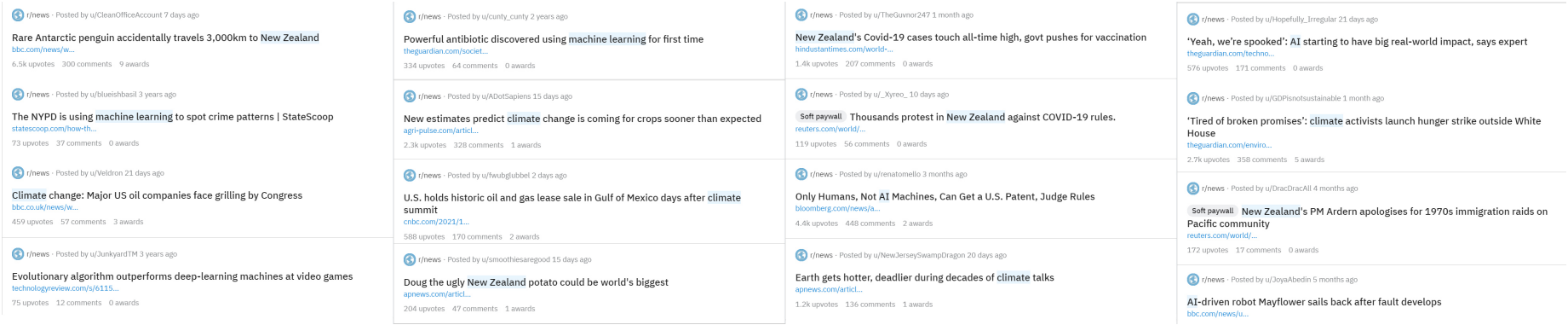


Figure 1 – Un flux typique sur r/news

Introduction

- Chaque minute :
  - 400h de video
  - 350 000 tweets
  - 500 000 commentaires
  - 4 200 000 recherches
- Comment rendre ceci intelligible *automatiquement* ?

Figure 1 – Un flux typique sur r/news – avec topics

Information disponible

- Indices :
  - Information textuelle

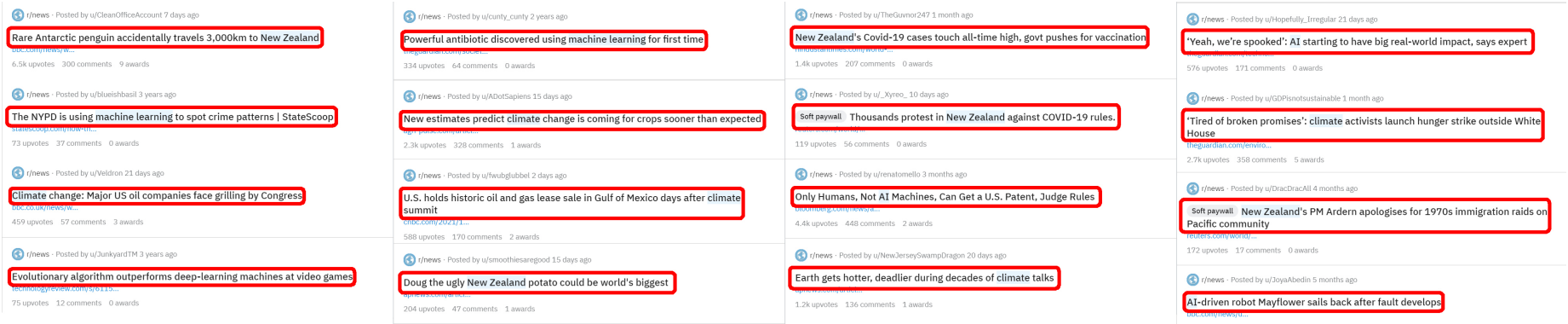


Figure 2 – Nous pouvons utiliser l'information textuelle

Information disponible

- Indices :
  - Information textuelle
  - Information temporelle

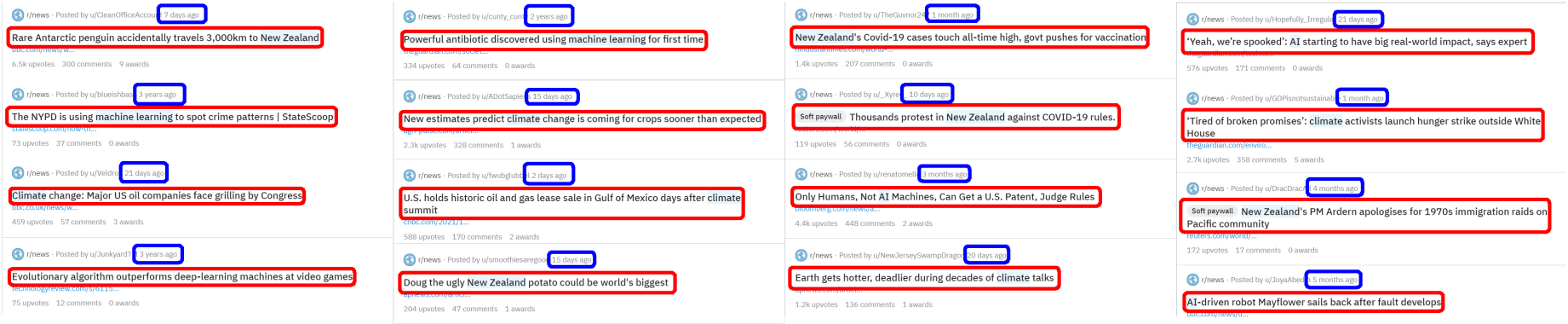


Figure 2 – Nous pouvons utiliser l'information textuelle et temporelle

Flux de documents

- Les données se présentent donc sous forme de flux

Etat de l'art

- Temps souvent "modélisé" en échantillonnant les informations (DTM, TOT, RCRP, DDCRP, etc.)
  - ◇ Problème : découpage des données, quel échantillonnage, poids des observations.. ?
- On aimerait plutôt modéliser le temps explicitement.

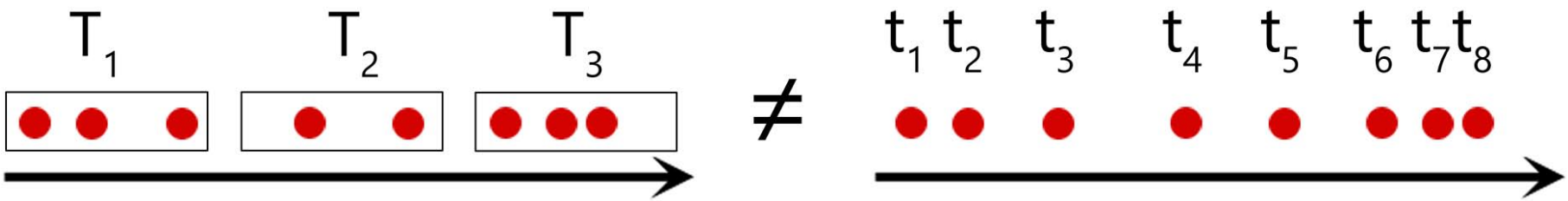
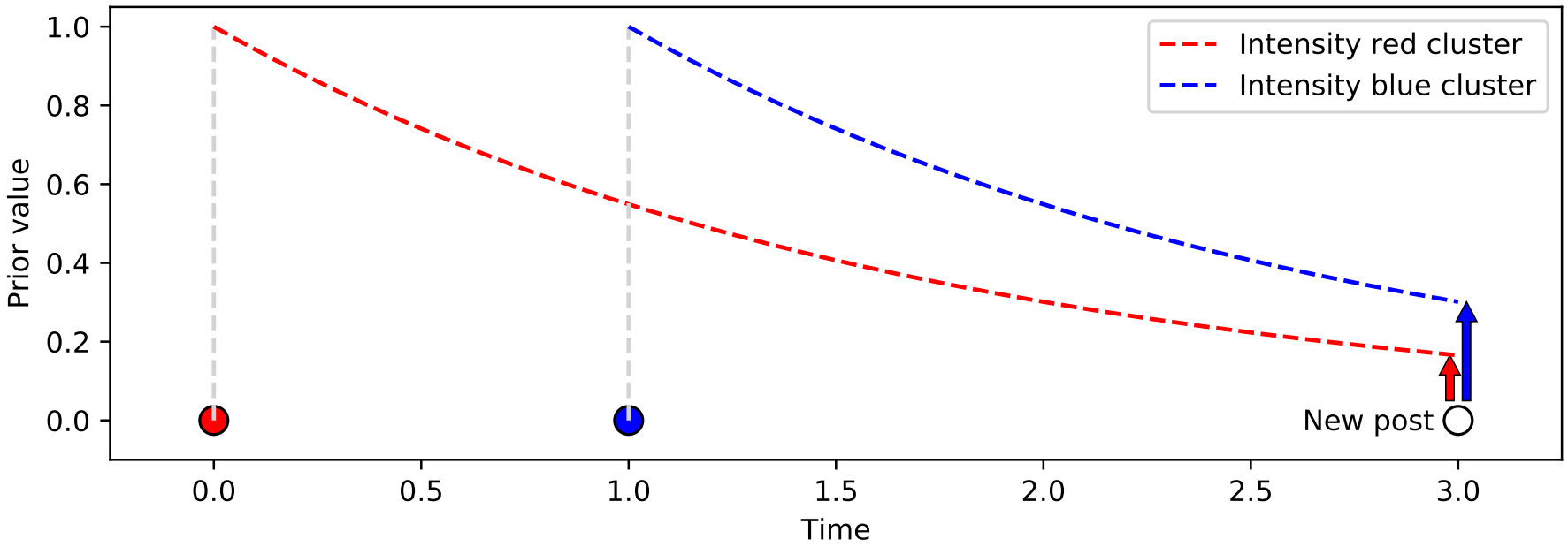


Figure 3 – Échantillonnage/découpage induisent des approximations

Etat de l'art

- (Du *et al.*, KDD 2015) : Dirichlet-Hawkes prior (inférence Bayésienne)

$$P(\text{cluster}|\text{text, time}) \propto \underbrace{P(\text{text}|\text{cluster})}_{\substack{\text{Textual likelihood} \\ \text{(Dirichlet-Multinomial)}}} \times \underbrace{P(\text{cluster}|\text{time, history})}_{\substack{\text{Temporal prior} \\ \text{(Dirichlet-Hawkes)}}}$$



Etat de l'art

- Le modèle a cette forme :

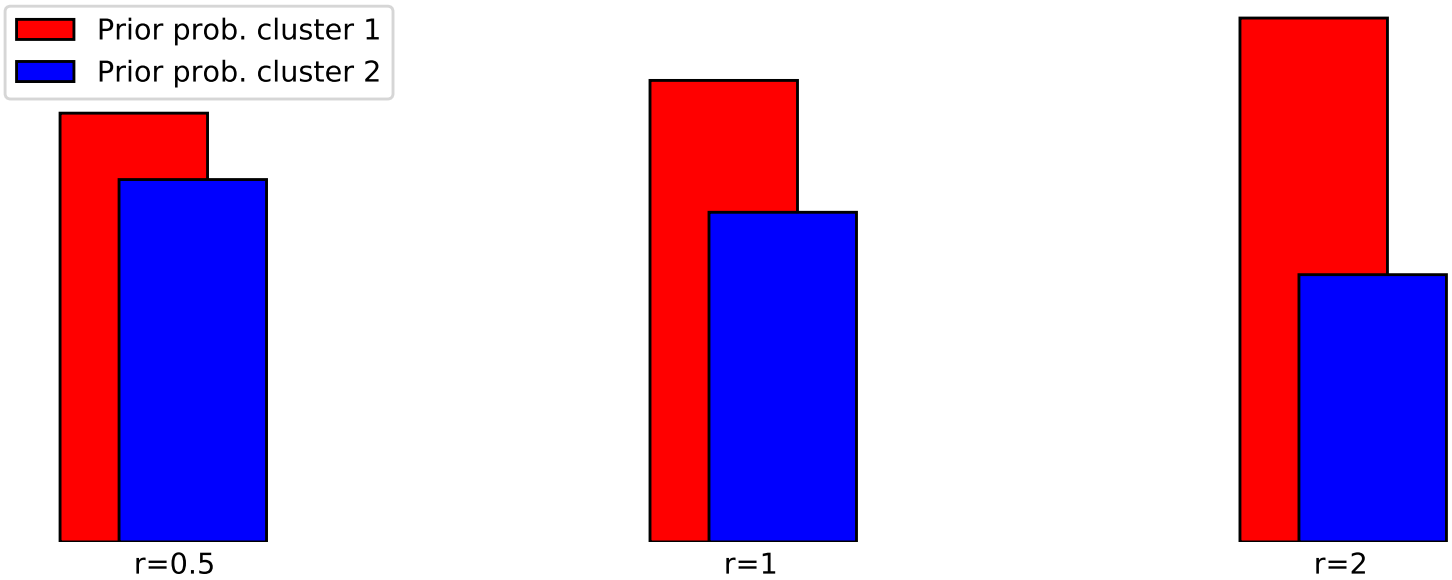
$$P(\text{cluster}|\text{text, time}) \propto \underbrace{P(\text{text}|\text{cluster})}_{\substack{\text{Likelihood textuelle} \\ \text{(Dirichlet-Multinomial)}}} \times \underbrace{P(\text{cluster}|\text{time, history})}_{\substack{\text{Prior temporel} \\ \text{(Dirichlet-Hawkes)}}}$$

- Pourquoi la probabilité *a priori* devrait évoluer linéairement avec l'intensité ? (Welling 2006 ; Wallach *et al.* 2009 & 2010)

Powered Dirichlet Hawkes process

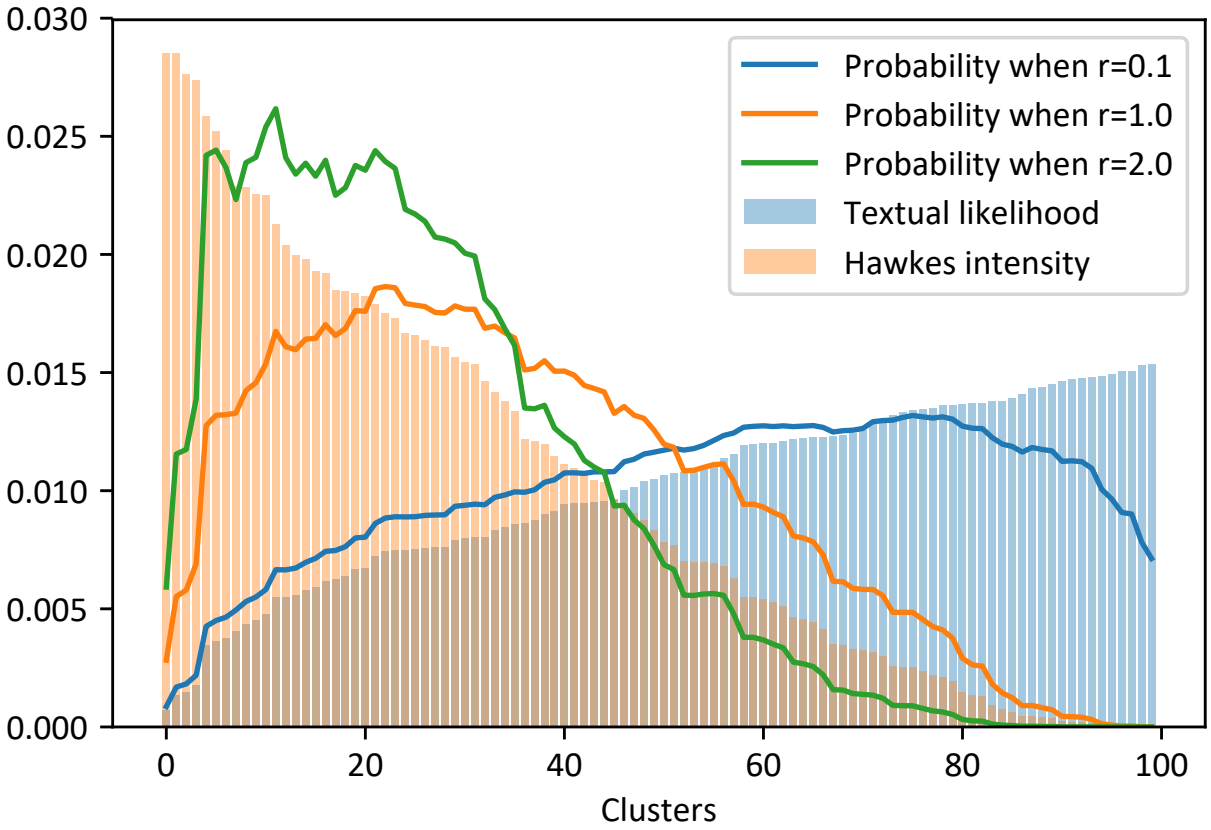
- $P(c|t,\mathcal{H})$: probabilité *a priori* du cluster c au temps t sachant $\mathcal{H}$
- $\lambda_c(t)$: intensité du cluster c au temps t
- On définit le Powered Dirichlet-Hawkes process :

$$P(c|t,\mathcal{H},r) = \begin{cases} \frac{\lambda_c(t)^r}{\alpha_0+\sum_k \lambda_k(t)^r} & \text{if c = 1, ..., K} \\ \frac{\alpha_0}{\alpha_0+\sum_k \lambda_k(t)^r} & \text{if c = K+1} \end{cases}$$



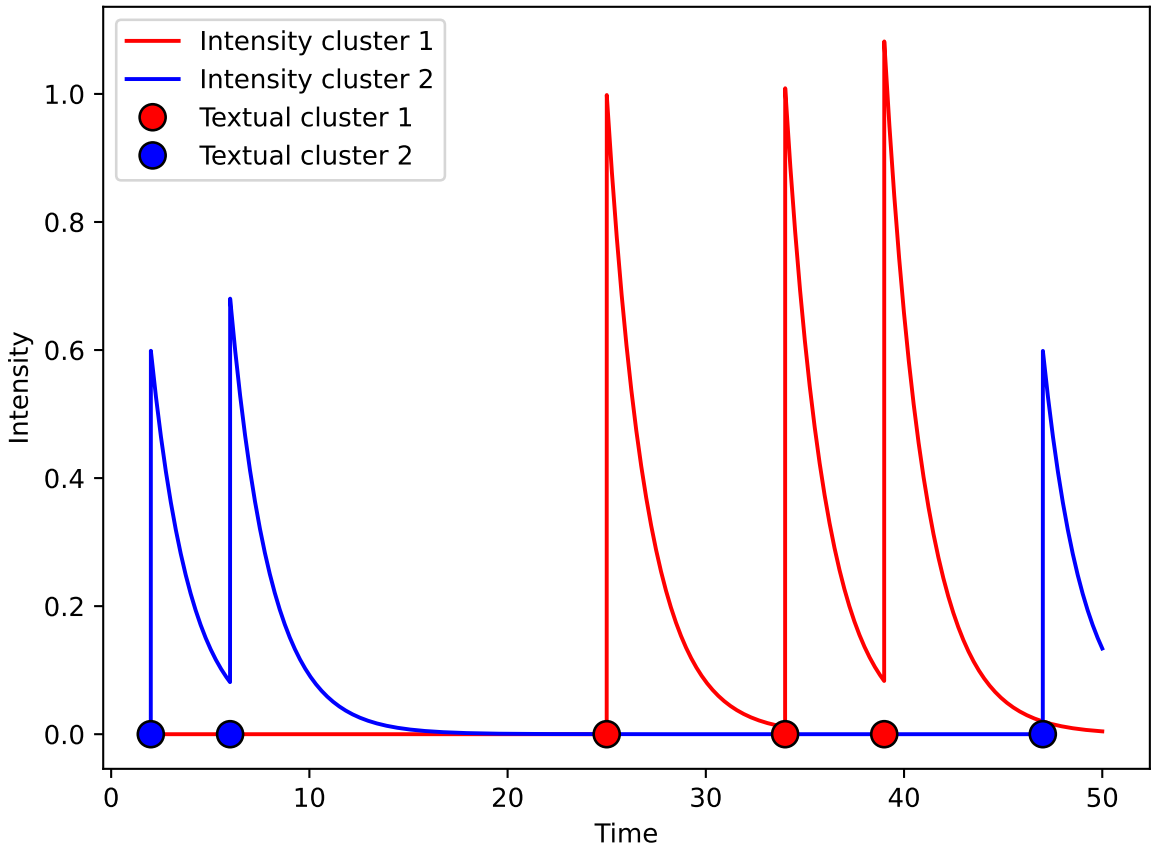
Changements induits par PDHP

$$P(\text{cluster}|\text{text, time}) \propto \underbrace{P(\text{text}|\text{cluster})}_{\text{Likelihood textuelle}} \times \underbrace{P(\text{cluster}|\text{time, r, history})}_{\text{PDHP } \textit{a priori}}$$



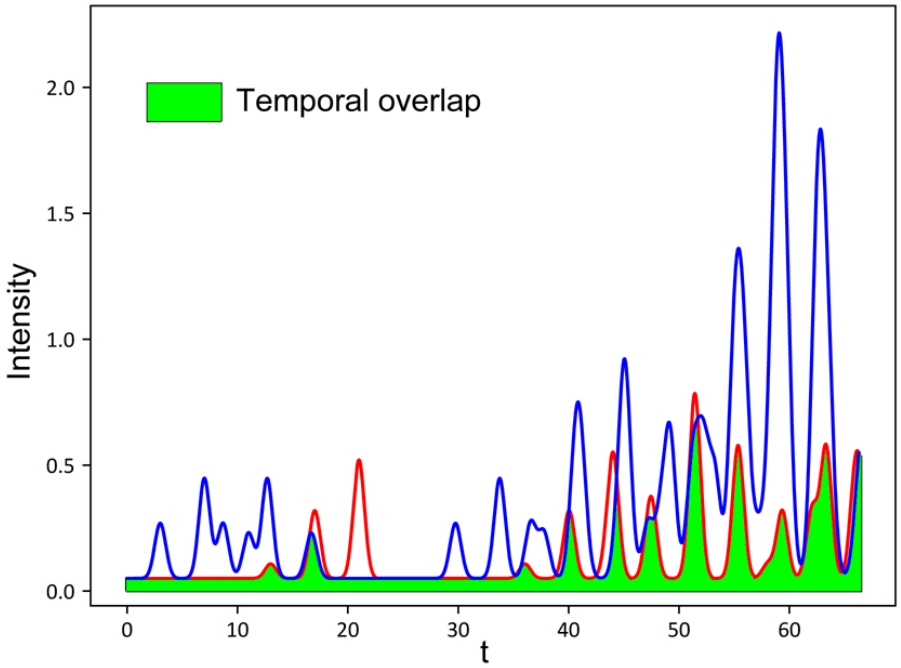
Datasets

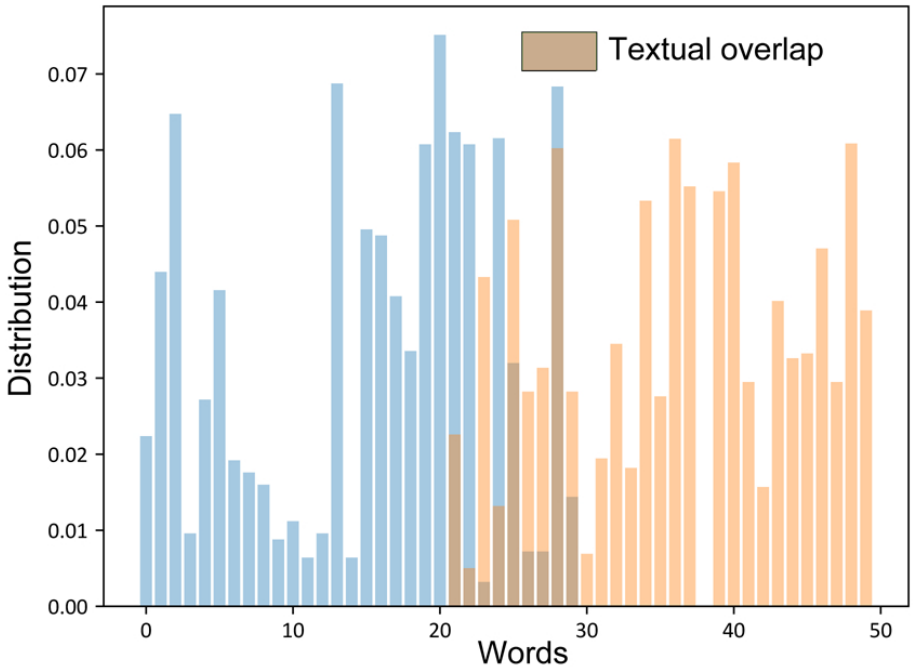
- 300 datasets synthétiques
  - 10 pour chaque combinaison de recouvrement textuel et temporel
  - 10 pour chaque valeur de décorrélation (discuté plus tard)



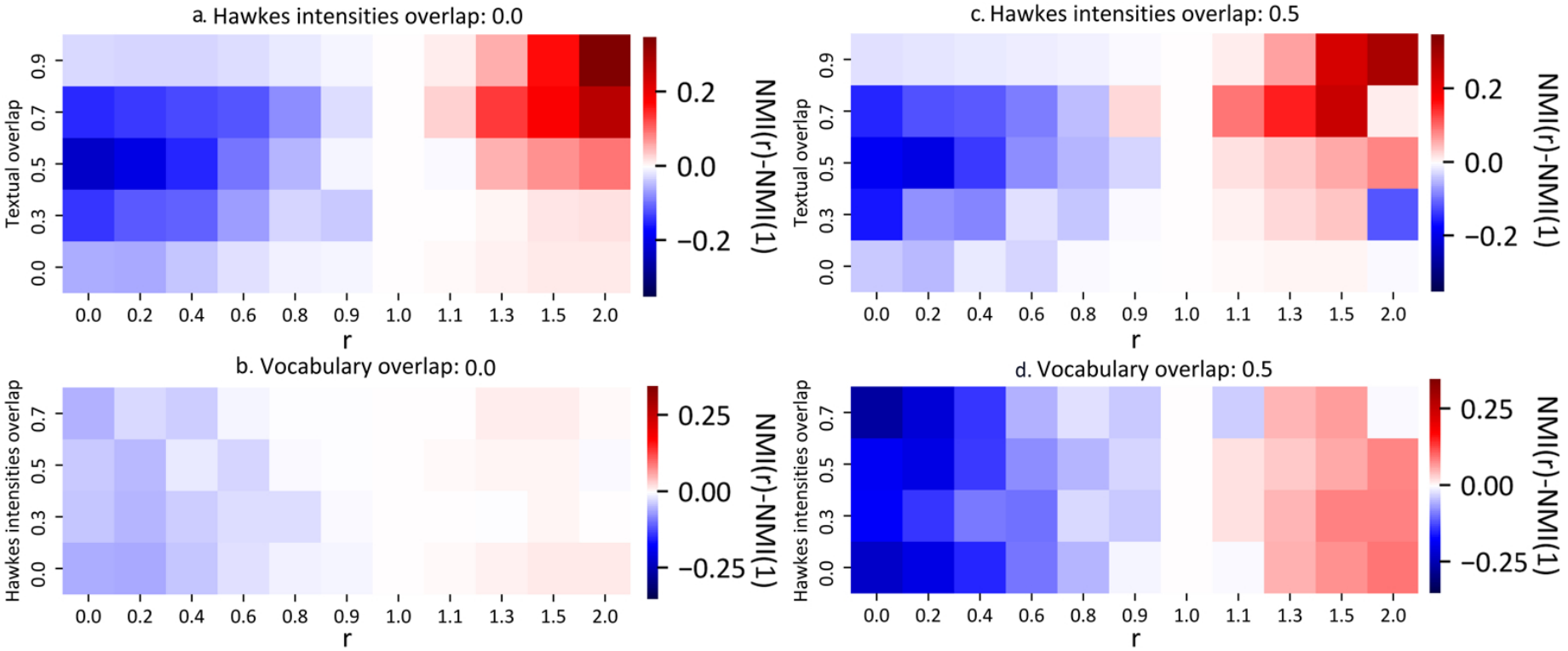
Recouvrements

- Les recouvrements sont définis comme suit :





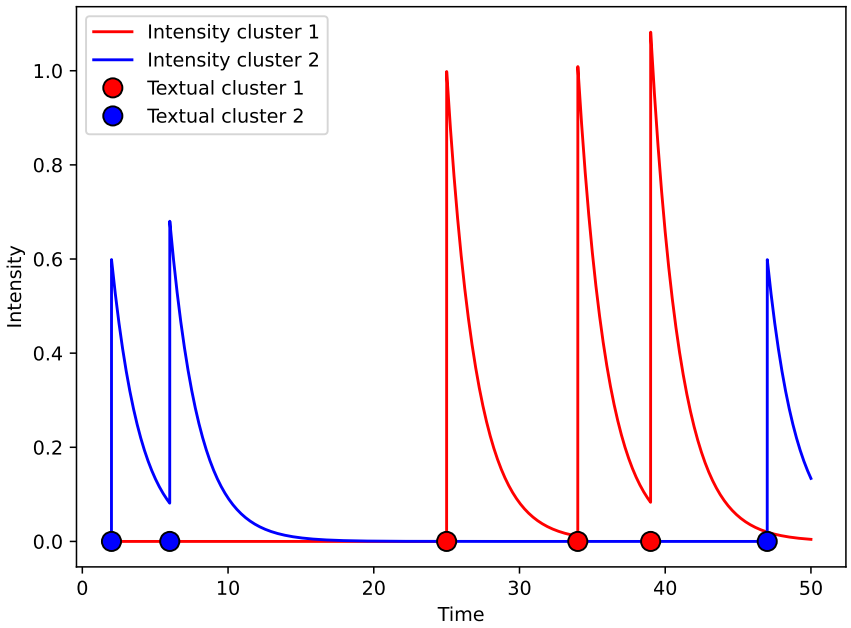
NMI difference avec l'état de l'art

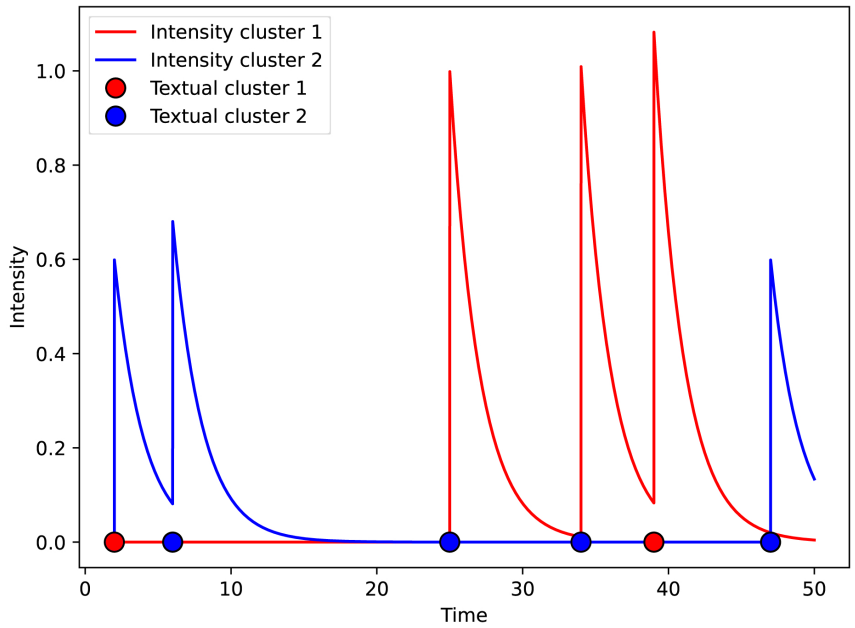


- PDHP permet de s'adapter aux diverses situations (+0.3 NMI) :
  - Grand recouvrement de vocabulaire
  - Grand recouvrement d'intensités
  - Pas de recouvrement

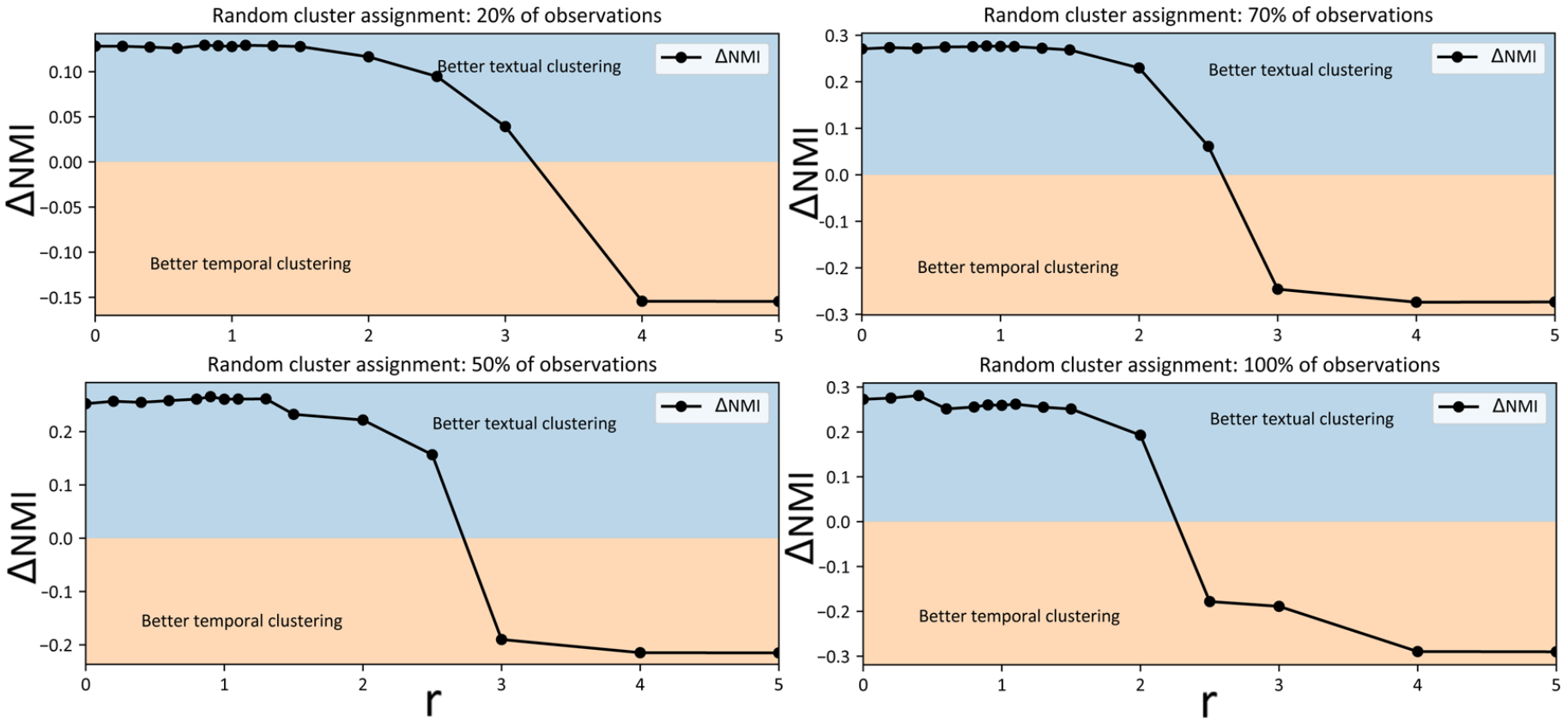
Décorrelations

- Décorrelations :



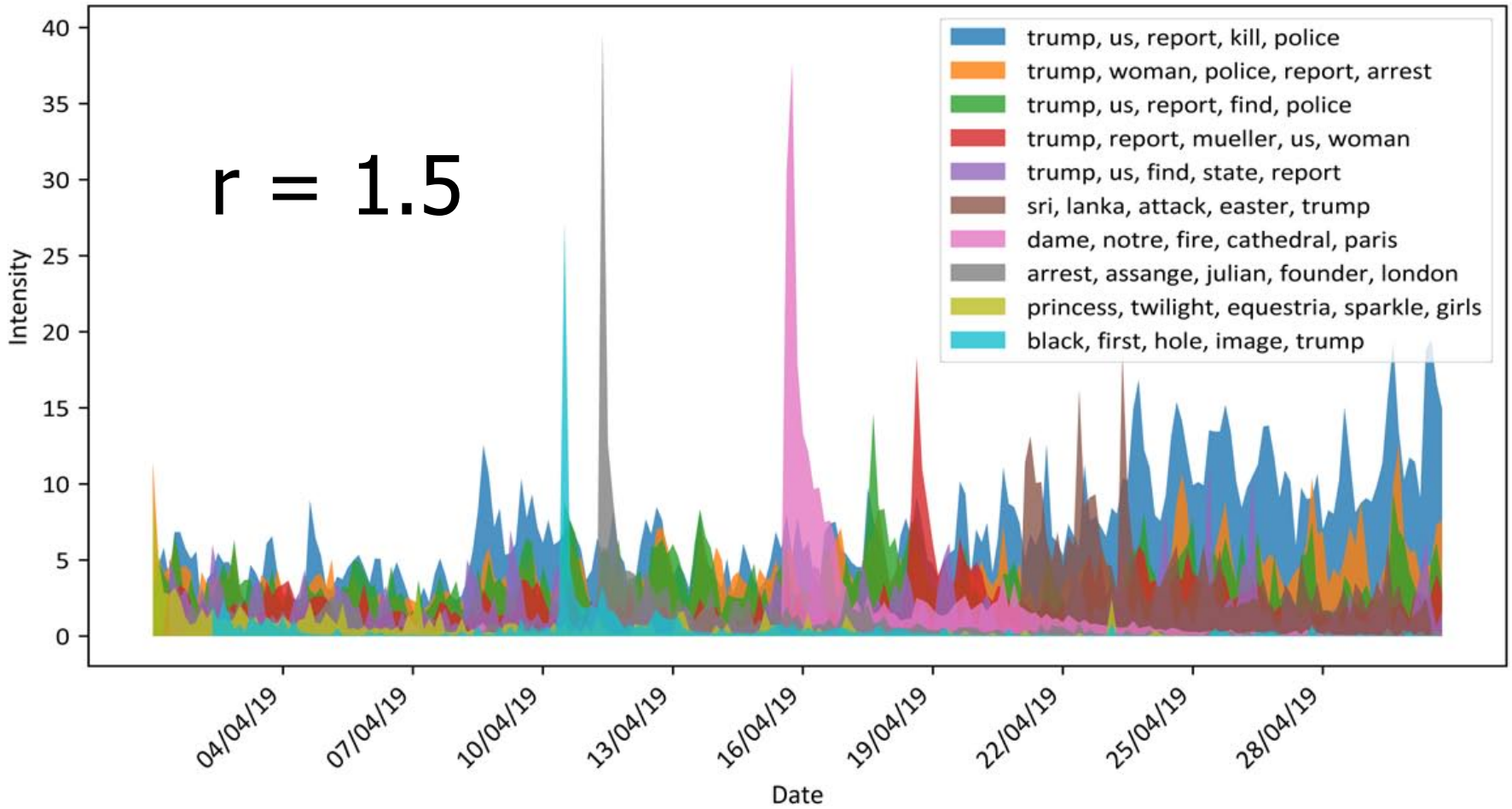


Différence entre NMI textuelle et temporelle

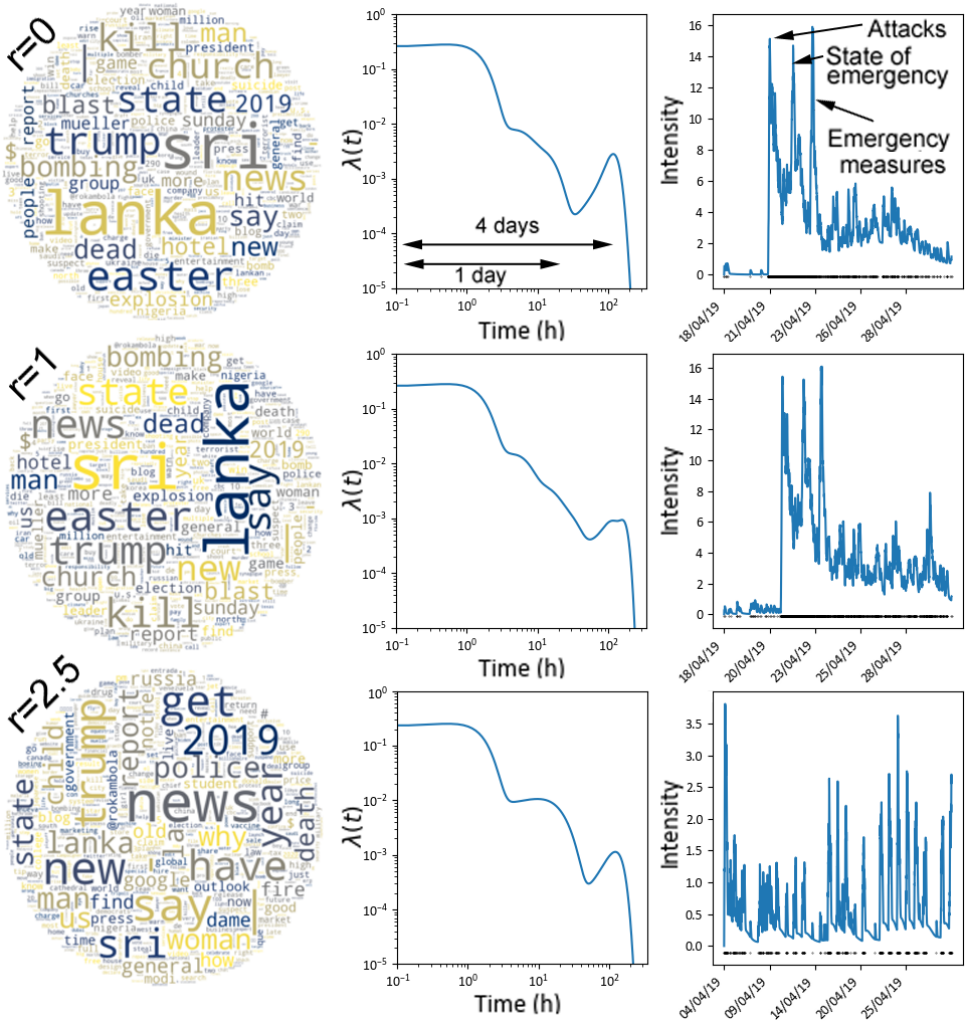


- PDHP permet de recouvrer un types de clusters ou l'autre
  - Petit r : bons clusters textuels
  - Grand r : bons clusters temporels

Merci de votre attention !



Application sur un jeu de données réel



- Données réelles : r/news
- Différents clusters et dynamiques en fonction de r
  - Petit r : vocabulaires similaires
  - Grand r : dynamiques particulières

Autres métriques

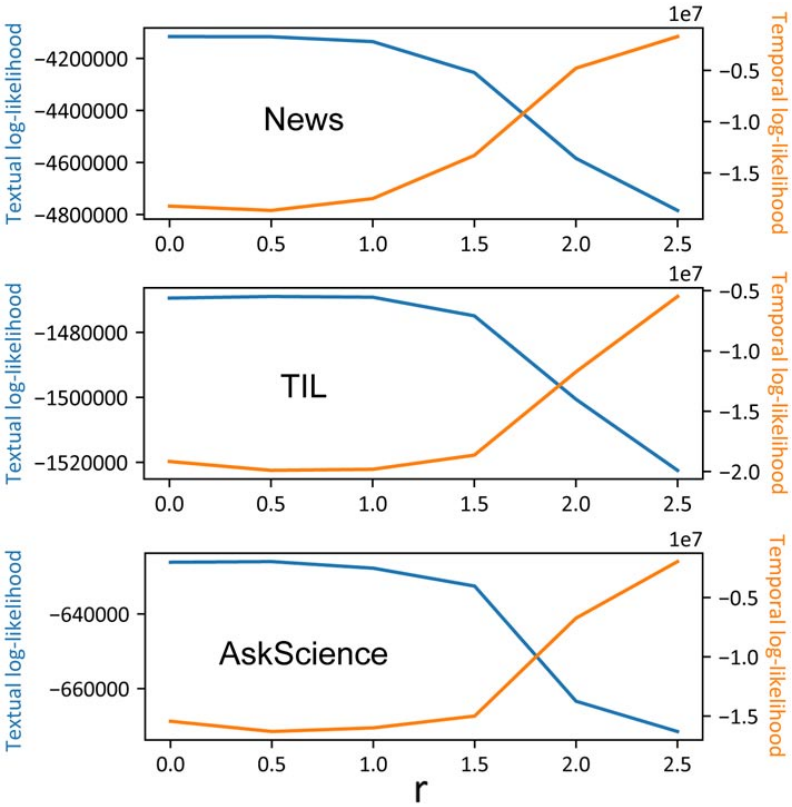


Figure 5 – Likelihoods textuelle et temporelle vs r

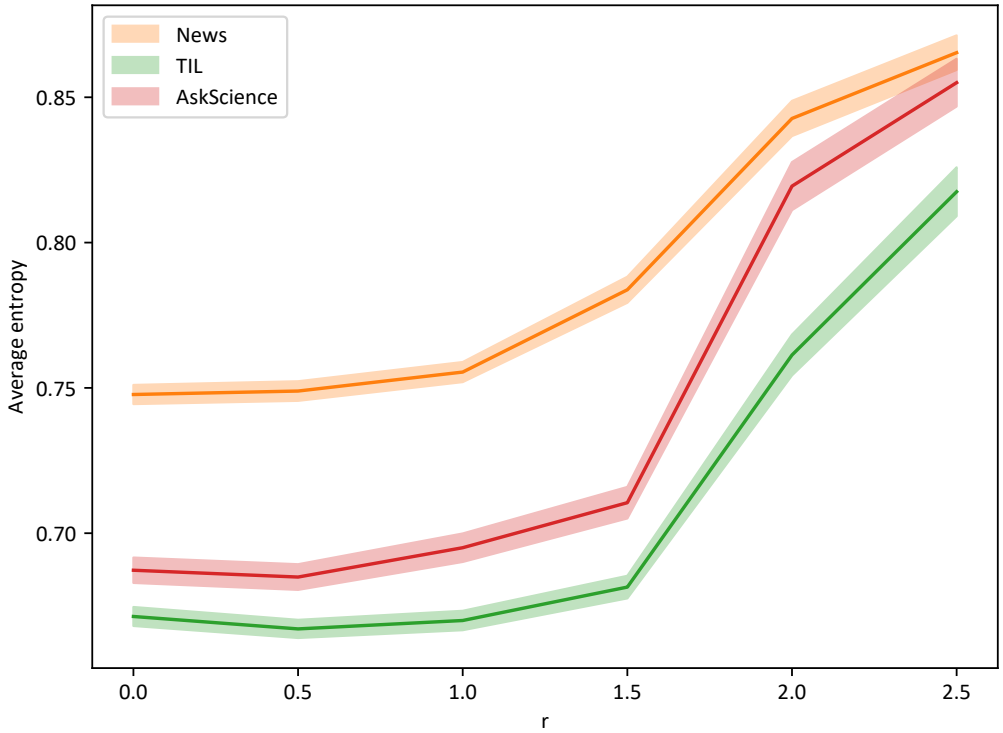
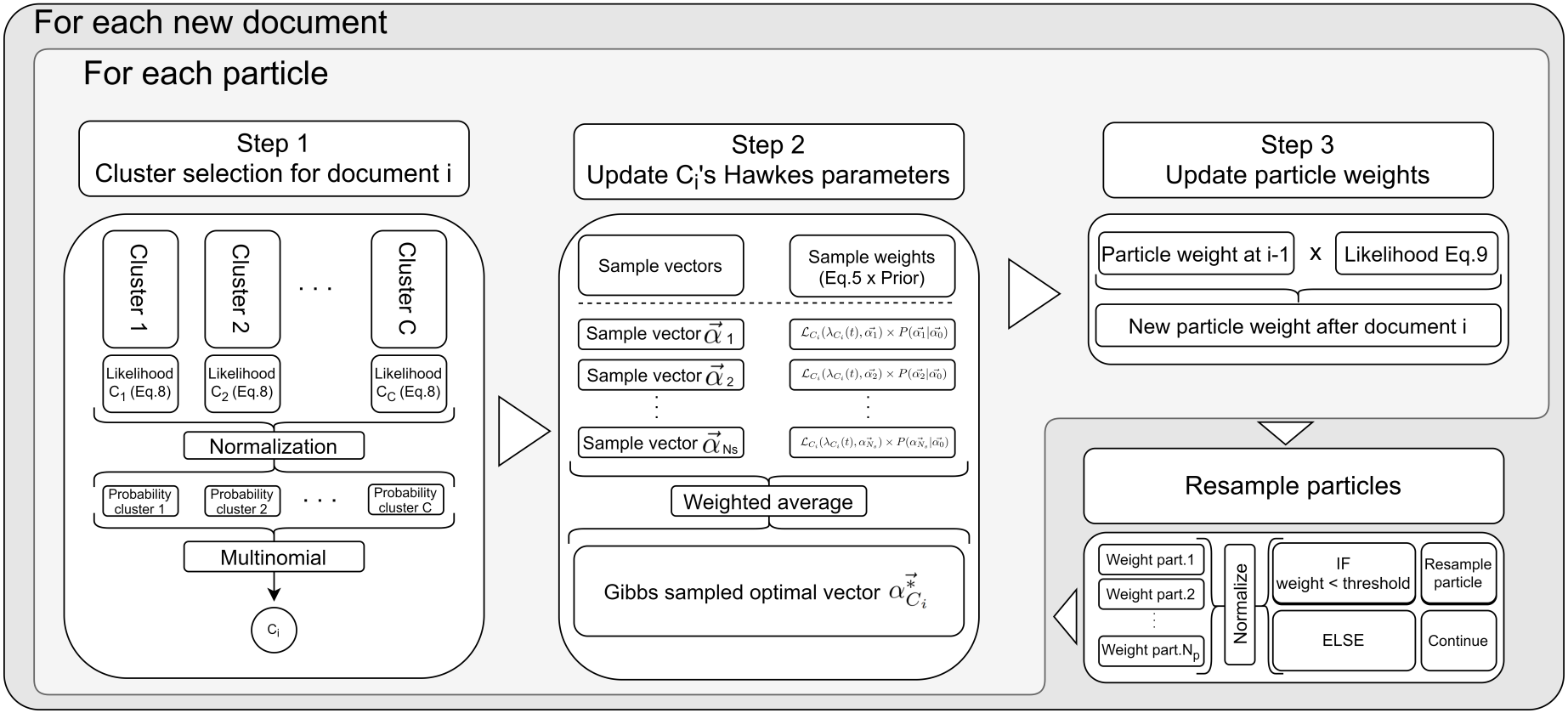
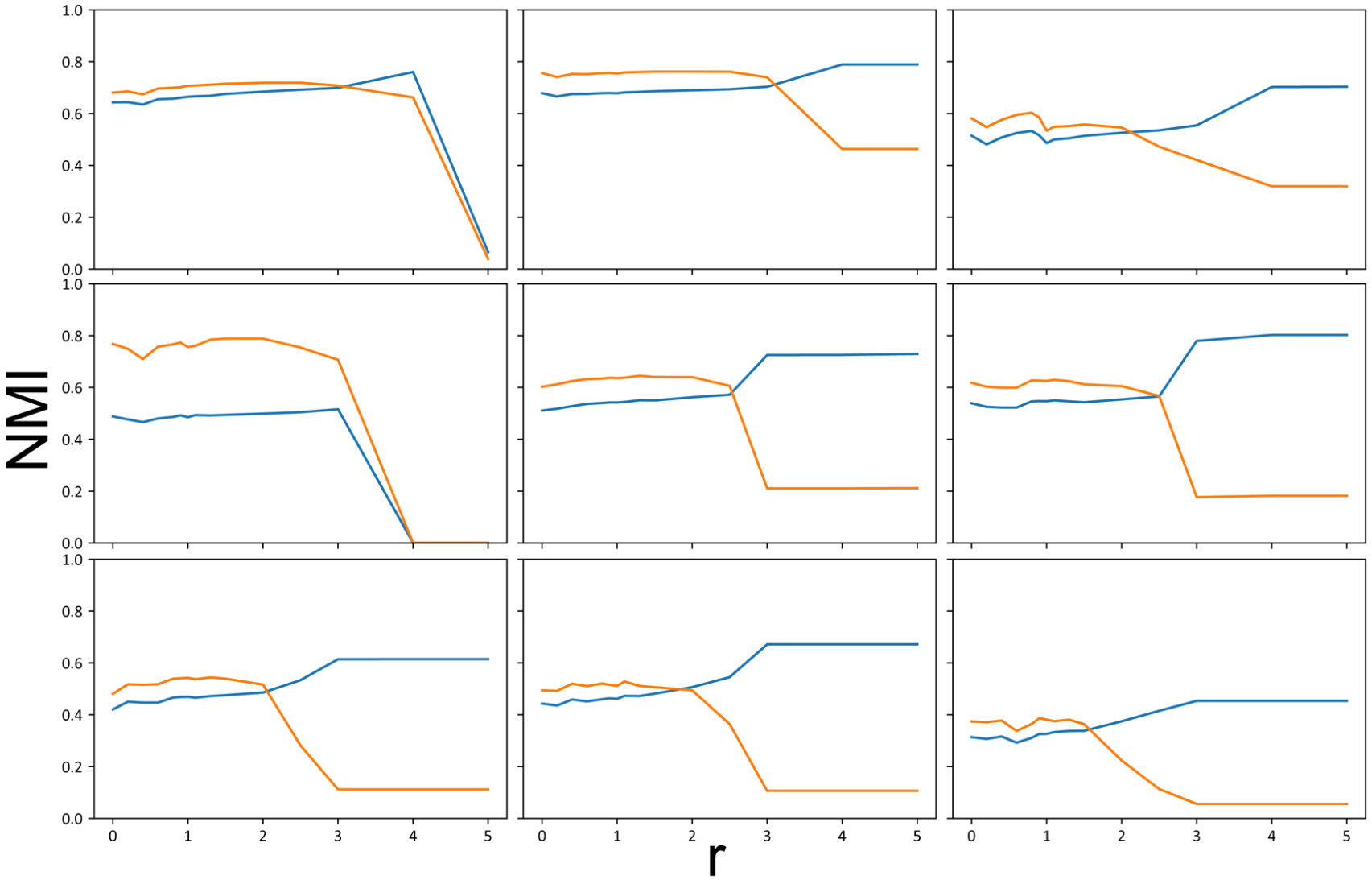


Figure 6 – Entropy des clusters textuels : ils sont plus concentrés pour de faibles r

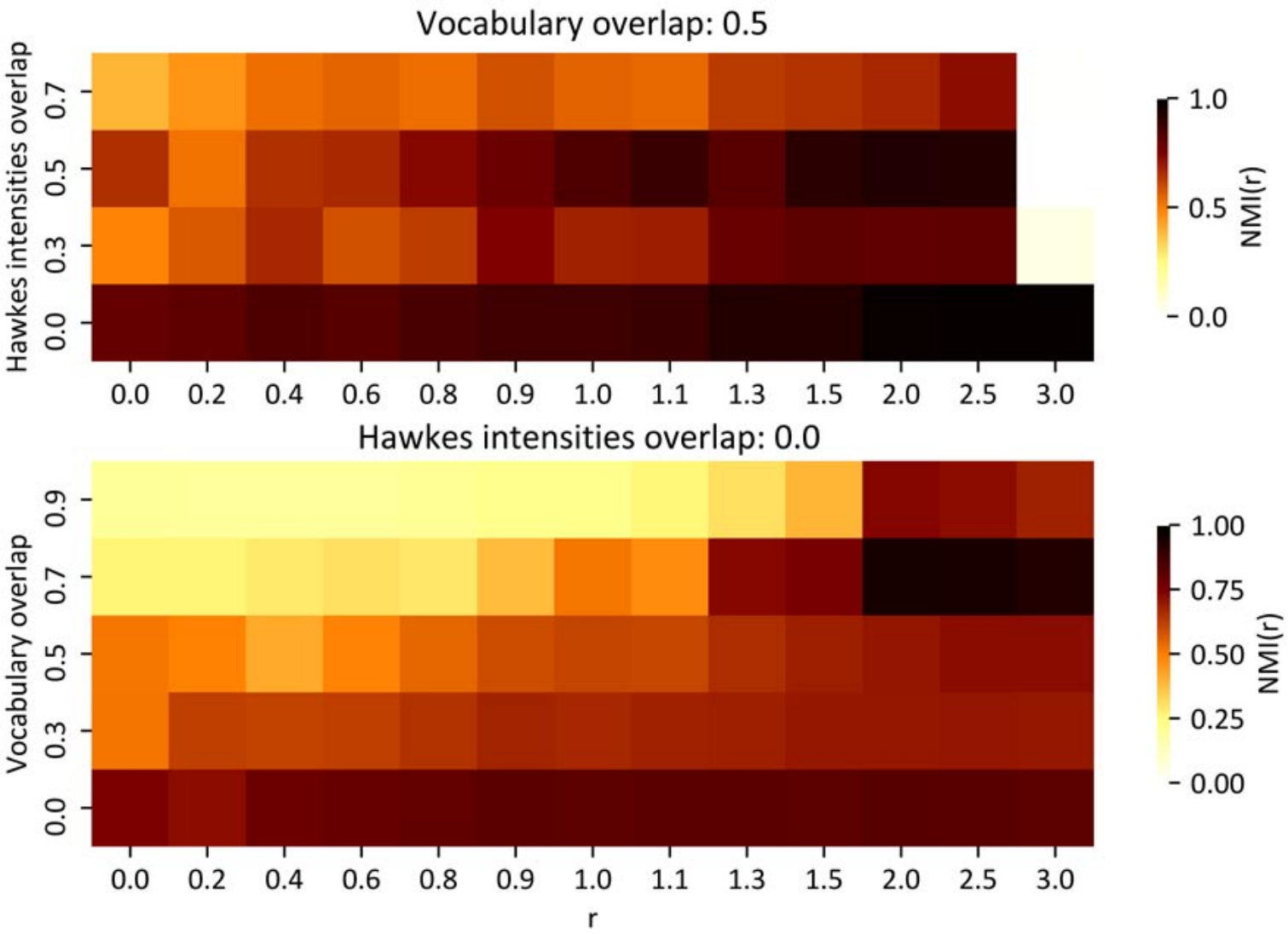
Optimization



Decorrelation (1 run)



Raw NMI


Vocabulary overlap: 0.5
Hawkes intensities overlap
0.0
0.3
0.5
0.7
0.0
0.2
0.4
0.6
0.8
0.9
1.0
1.1
1.3
1.5
2.0
2.5
3.0
1.0
0.5
0.0
NMI(r)
Hawkes intensities overlap: 0.0
Vocabulary overlap
0.0
0.3
0.5
0.7
0.9
r
1.00
0.75
0.50
0.25
0.00
NMI(r)