Powered Hawkes-Dirichlet Process: Challenging Textual Clustering using a Flexible Temporal Prior

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

Gaël Poux-Médard

Université de Lyon, France
Lyon 2, ERIC UR 3083

December 2021

UNIVERSITÉ DE LYON

IEEE ICDM 2021

Intro

- Every minute:
  - 400h of video
  - 350 000 tweets
  - 500 000 comments
  - 4 200 000 searches

Intro

- Every minute:
  - 400h of video
  - 350 000 tweets
  - 500 000 comments
  - 4 200 000 searches
- How to make sense out of *that*?

Figure 1: A typical stream from r/news

Intro

- Every minute:
  - 400h of video
  - 350 000 tweets
  - 500 000 comments
  - 4 200 000 searches
- How to *automatically* make sense out of that?

Figure 1: A typical stream from r/news – with topics

Information available

- Main clues:
  - ◇ Textual information

Figure 2: We can use textual information

Information available

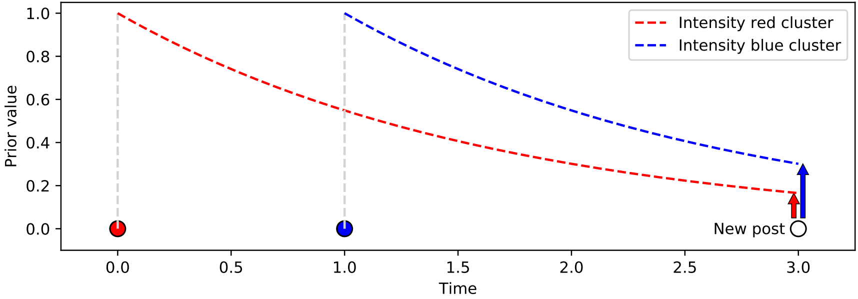
- Main clues:
  - ◇ Textual information
  - ◇ Temporal information

Figure 2: We can use textual information and temporal information

State of the art

- Lots of works consider time by sampling observations
- (Du *et al.*, KDD 2015): Dirichlet-Hawkes prior (Bayesian inference)

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



State of the art

- The model takes this form:

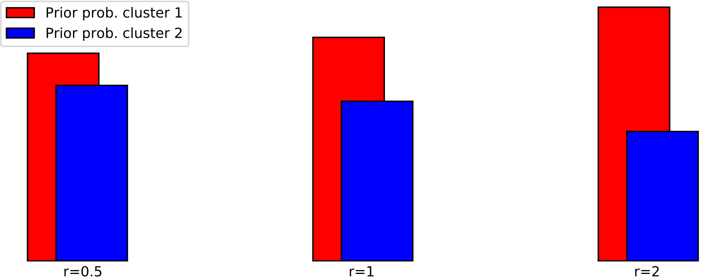
$$P(\text{cluster}|\text{text, time}) \propto \underbrace{P(\text{text}|\text{cluster})}_{\substack{\text{Textual likelihood} \\ \text{(Dirichlet-Multinomial)}}} \times \underbrace{P(\text{cluster}|\text{time, history})}_{\substack{\text{Temporal prior} \\ \text{(Dirichlet-Hawkes)}}}$$

- But does the prior probability have to evolve linearly with the intensity?

Powered Dirichlet Hawkes process

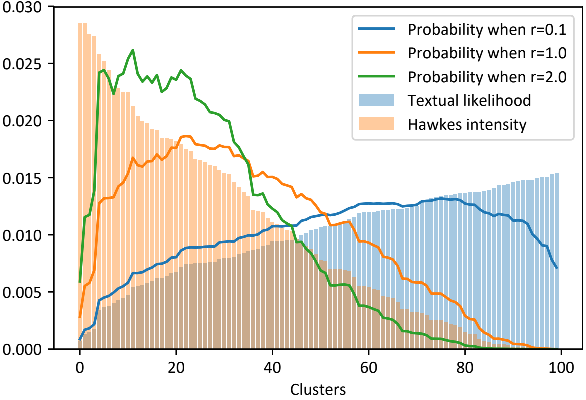
- $P(c|t,\mathcal{H})$: prior probability of cluster c at time t given history $\mathcal{H}$
- $\lambda_c(t)$: intensity of cluster c at time t
- We define the Powered Dirichlet-Hawkes process:

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



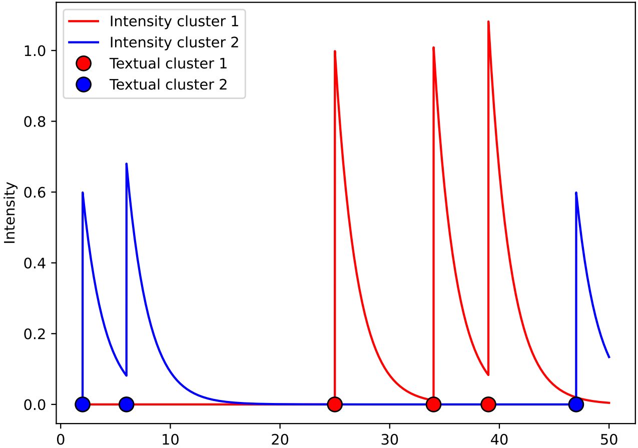
Changes induced by PDHP

$$P(\text{cluster}|\text{text, time}) \propto \underbrace{P(\text{text}|\text{cluster})}_{\text{Textual likelihood}} \times \underbrace{P(\text{cluster}|\text{time, r, history})}_{\text{PDHP temporal prior}}$$



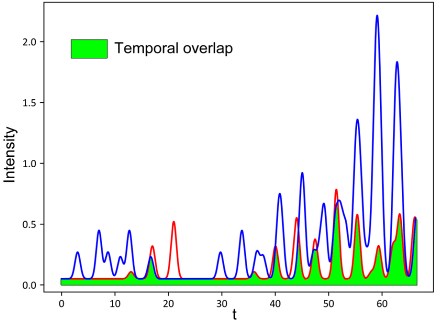
Datasets

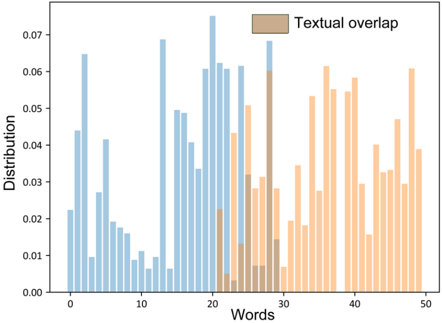
- 300 synthetic datasets
  - 10 for each value of temporal and textual overlaps
  - 10 for each value of decorrelation



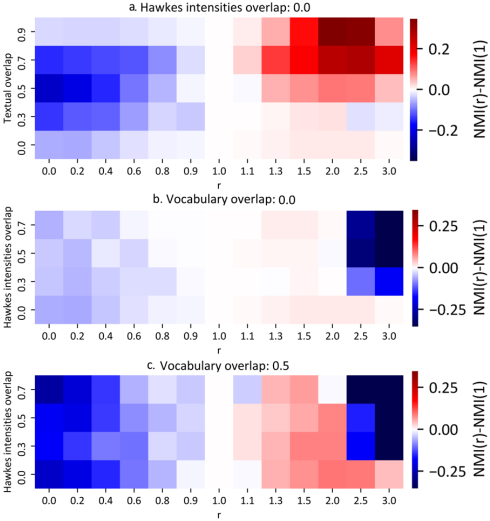
Overlaps

- Overlaps are defined as follow:





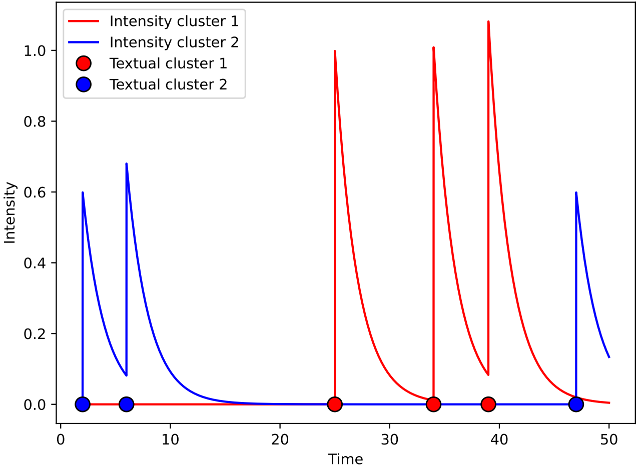
NMI difference wrt SotA



- PDHP adapts to various situations better than DHP:
  - Large textual overlap
  - Large temporal overlap
  - No overlap
- Up to +0.3 NMI in our case

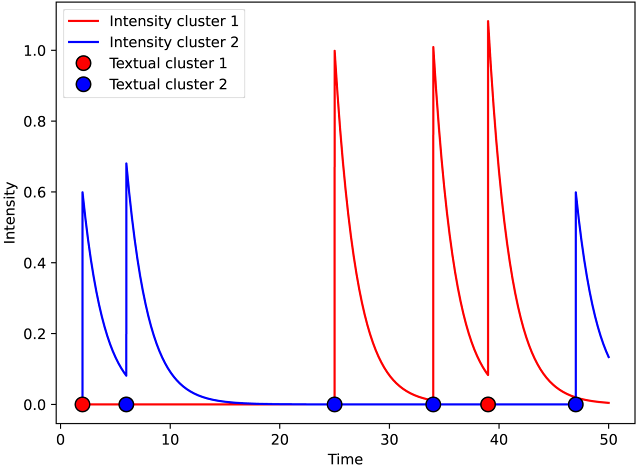
Datasets

- We decorrelate text and time

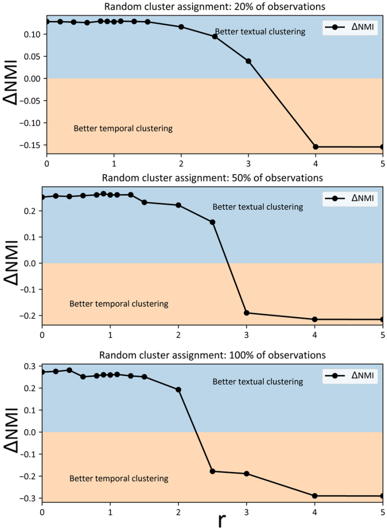


Datasets

- We decorrelate text and time

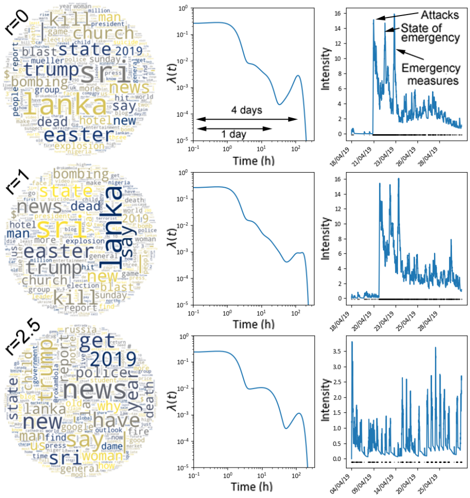


Difference between textual and temporal NMI



- PDHP retrieves either temporal or textual clusters
  - Small r: good textual clusters
  - Large r: good temporal clusters

Application to a real-world dataset



- Real world data: r/news
- Different clusters and dynamics for different r
  - Small r: similar vocabulary
  - Large r: specific dynamics

Other metrics

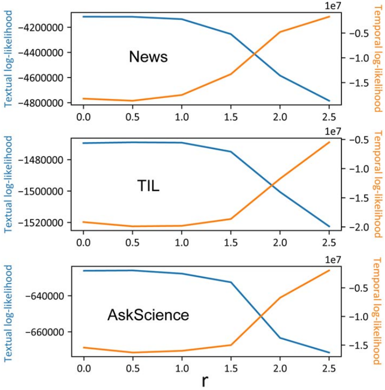


Figure 4: Textual and temporal likelihood vs r

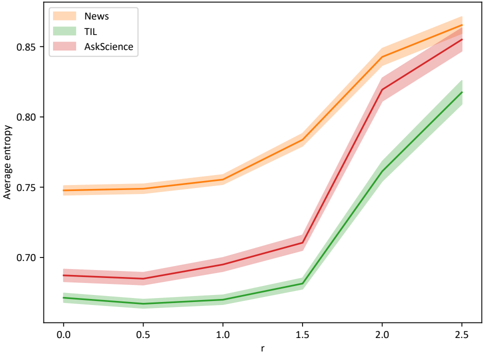
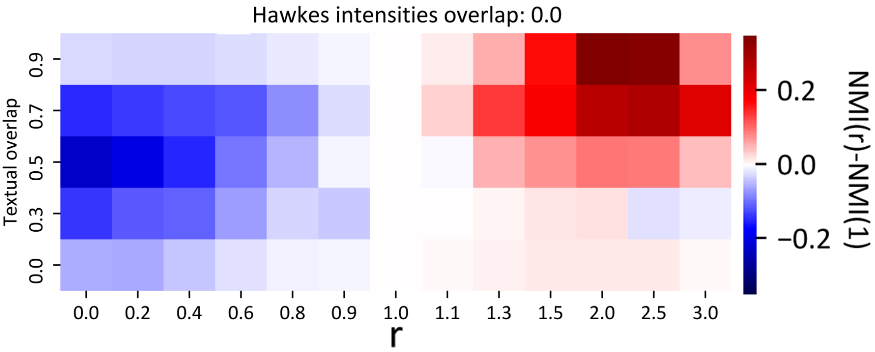
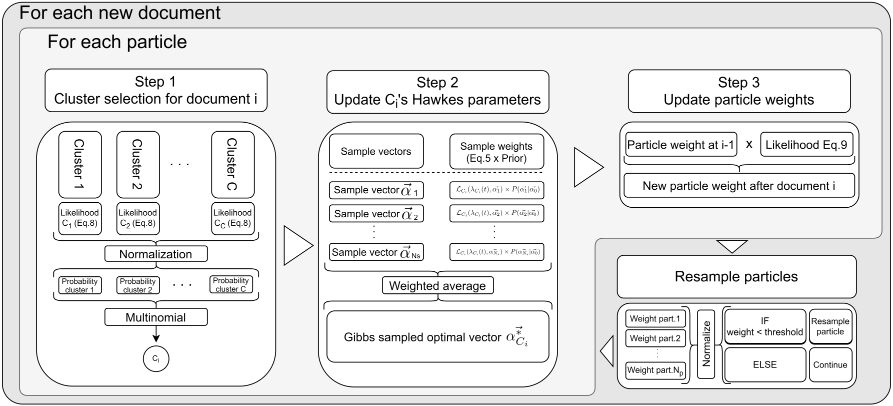


Figure 5: Entropy of textual clusters: sharper textual clusters for low r

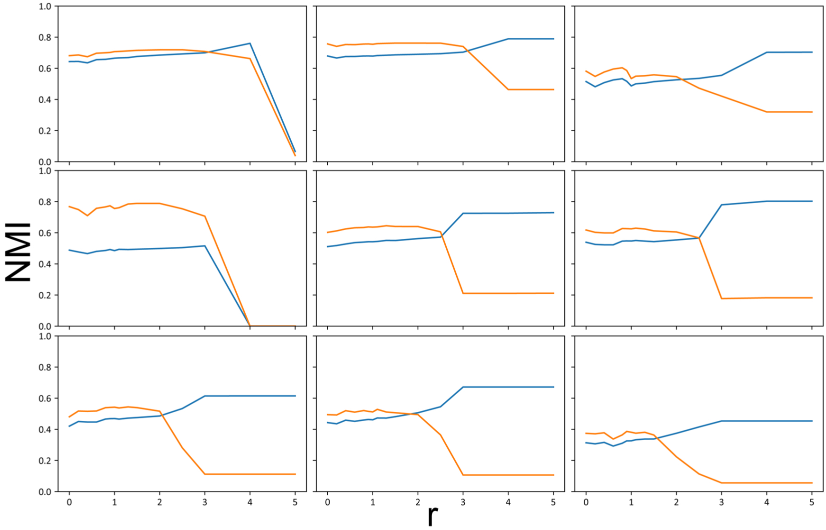
Thanks for your attention!



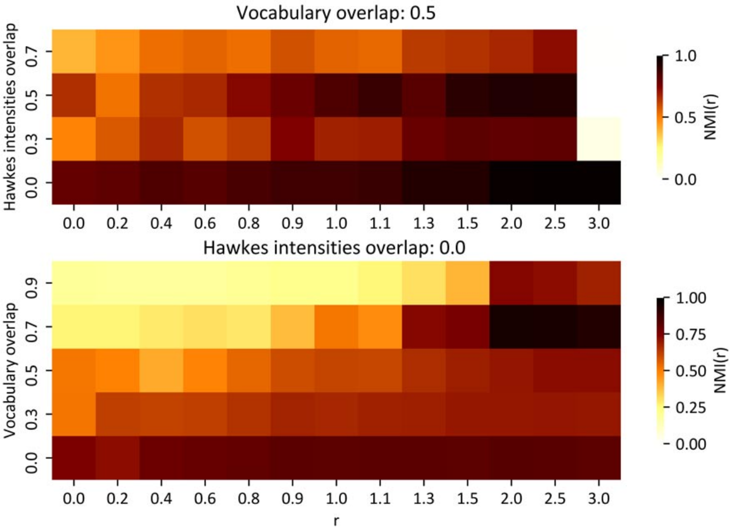
Optimization



Decorrelation (1 run)



Raw NMI



Vocabulary overlap: 0.5
Hawkes intensities overlap
0.7
0.5
0.3
0.0
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
0.9
0.7
0.5
0.3
0.0
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
r
1.00
0.75
0.50
0.25
0.00
NMI(r)