Multivariate Powered Dirichlet-Hawkes Process

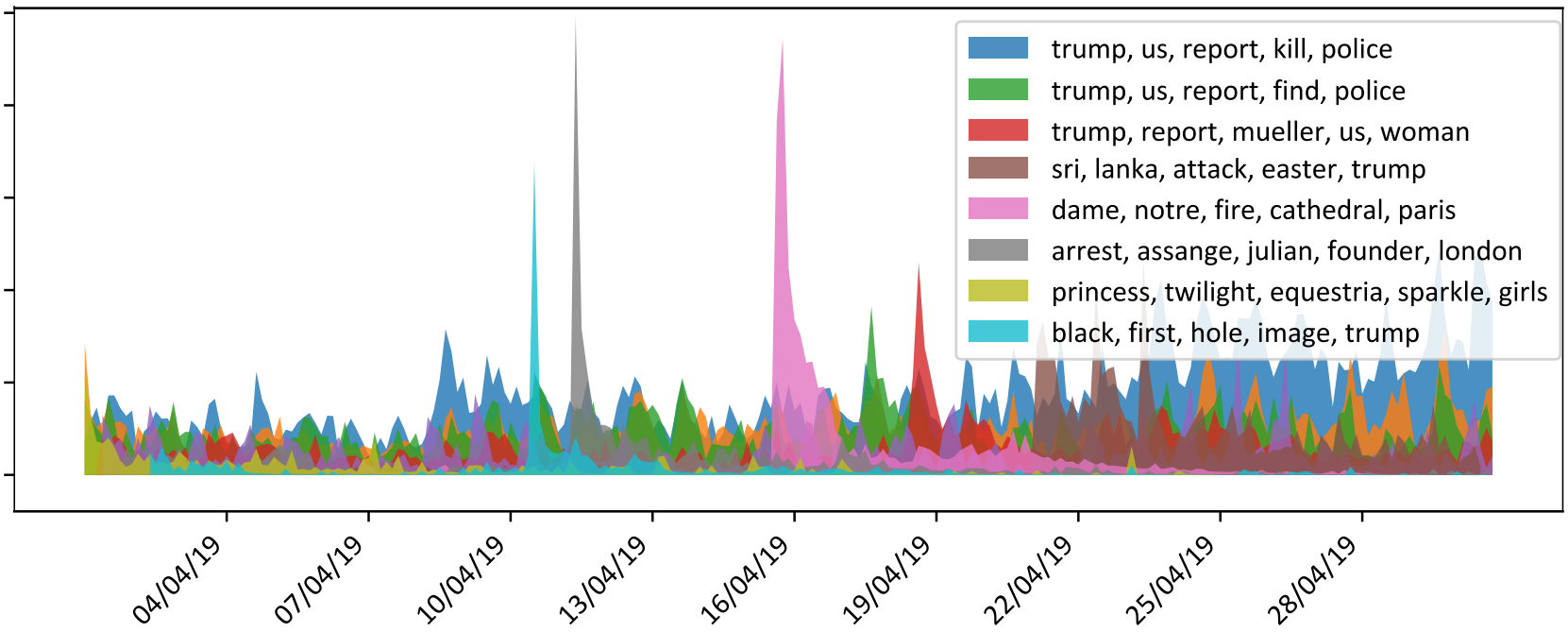
G.Poux-Médard, J. Velcin, S. Loudcher

Université de Lyon, France
Lyon 2, ERIC UR 3083

05 April 2023

Introduction

- Large flows of information on the internet
- Information can be split into clusters
- Two kinds of useful features: text and time
- Interaction between clusters has been little explored, in particular in a *continuous* manner



Contribution

- Previous work:
  - Combining Dirichlet-Hawkes Process and a simple language model (Du et al., 2015)
  - DHP has been improved with a more flexible prior: PDHP (Poux-Médard et al., 2021)
- Our contribution: Multivariate Powered Dirichlet-Hawkes Process
  - MPDHP: extension of previous work to deal with interacting clusters
  - Systematic test of MPDHP's application domain on synthetic data
  - Case study on real data from Reddit

Dirichlet Process

- Dirichlet Process: Bayesian clustering prior in sequential models

→ The probability of a cluster depends on its population

→ $\underbrace{P(\text{cluster}|\text{data, history } \mathcal{H})}_{\text{Posterior probability}} \propto \underbrace{P(\text{data}|\text{cluster})}_{\text{Likelihood}} \cdot \underbrace{P(\text{cluster}|\mathcal{H})}_{\text{Dirichlet prior}}$

$$DP(c|\mathcal{H},\alpha) = \begin{cases} \frac{N_c}{\alpha+N} & \text{if c = 1, ..., K} \\ \frac{\alpha}{\alpha+N} & \text{if c = K+1} \end{cases}$$

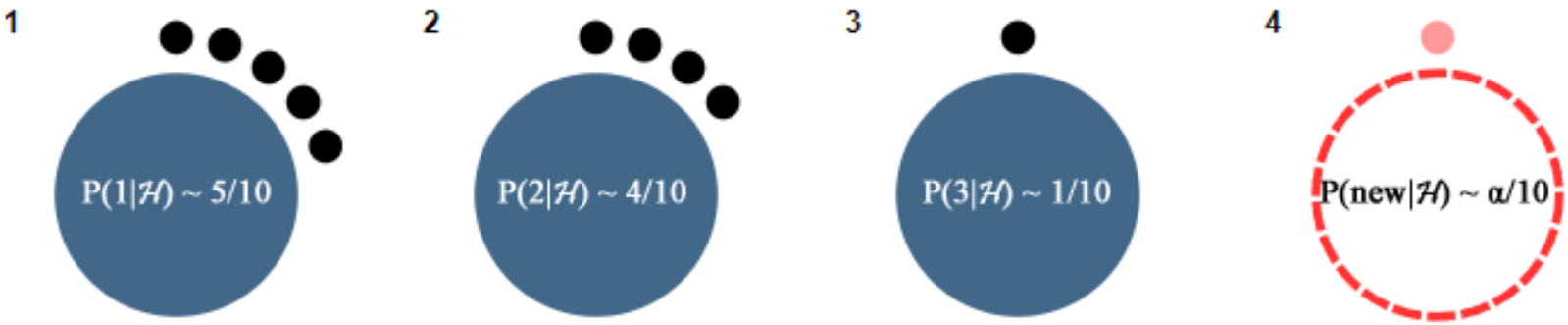


Figure 1: A Dirichlet Process after 10 steps

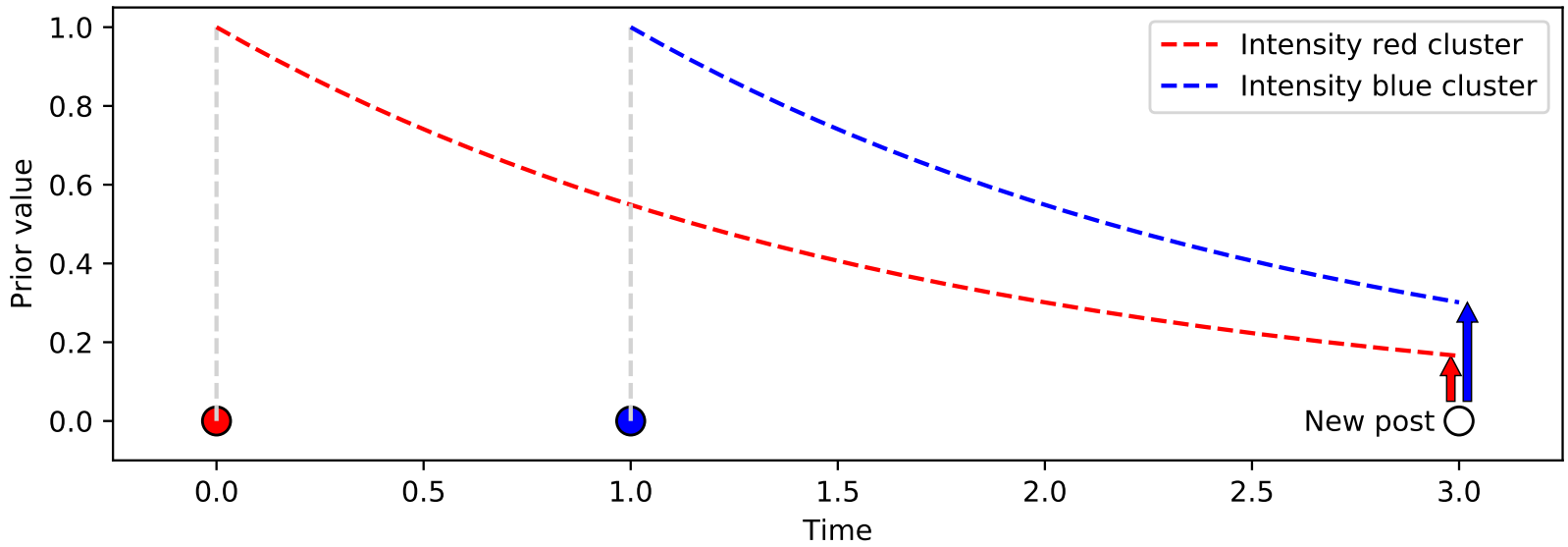
Dirichlet-Hawkes Process

- Dirichlet-Hawkes Process: history relevance should decay over time

→ Counts weighted over time as a Hawkes process of intensity $\lambda_c(t)$

→ Prob. of a cluster depends on its population at a given time

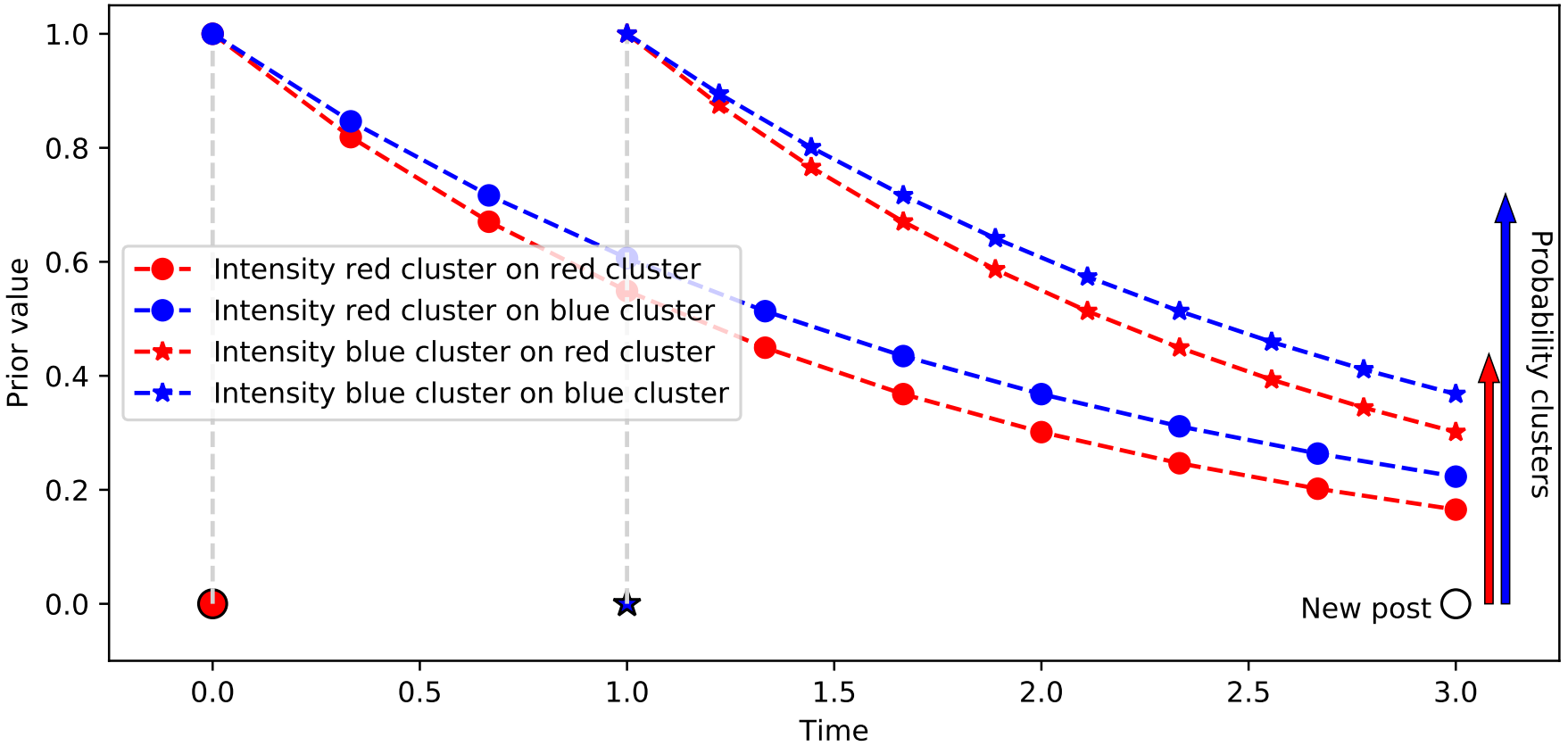
$$DHP(c|t,\lambda_0,\mathcal{H}) = \begin{cases} \frac{\lambda_c(t)}{\alpha_0+\sum_k \lambda_k(t)} & \text{if c = 1, ..., K} \\ \frac{\lambda_0}{\alpha_0+\sum_k \lambda_k(t)} & \text{if c = K+1} \end{cases}$$



Multivariate Powered Dirichlet-Hawkes Process

- Multivariate Powered Dirichlet-Hawkes Process: clusters interact

→ Prob. of a cluster also depends on the population of *other* clusters



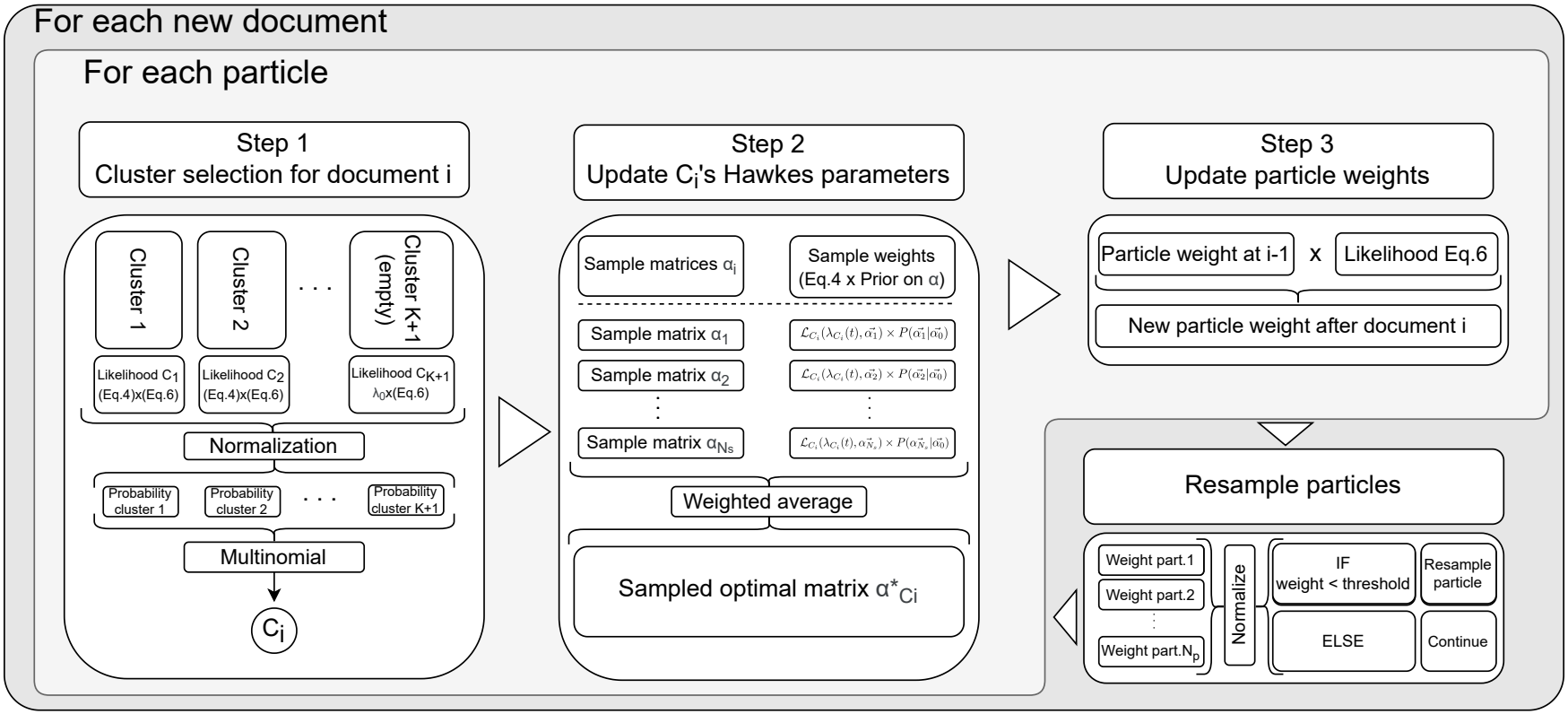
Our problem

Given a sequence of observations (text,time), find the best possible cluster c_i for each new observation, with:

$$\underbrace{P(\text{cluster}|\text{data, history } \mathcal{H})}_{\text{Posterior probability}} \propto \underbrace{P(\text{data}|\text{cluster})}_{\text{Likelihood}} \cdot \underbrace{P(\text{cluster}|\mathcal{H})}_{\text{MPDHP prior}}$$

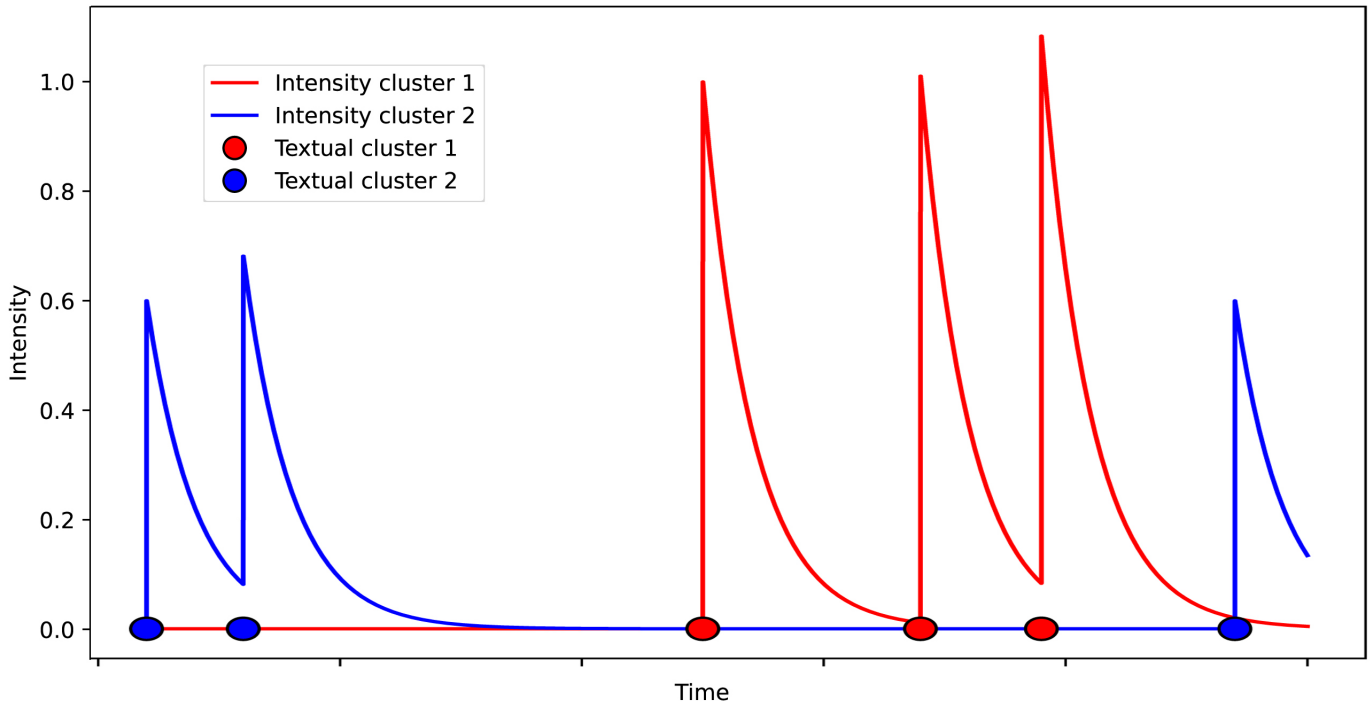
Optimization

- Sequential Monte Carlo algorithm: clusters and point processes intensities are updated as new data is added
- Step 2: params. updated by averaging likelihood-weighted samples
- Runs in $\mathcal{O}(N)$



Simulated data

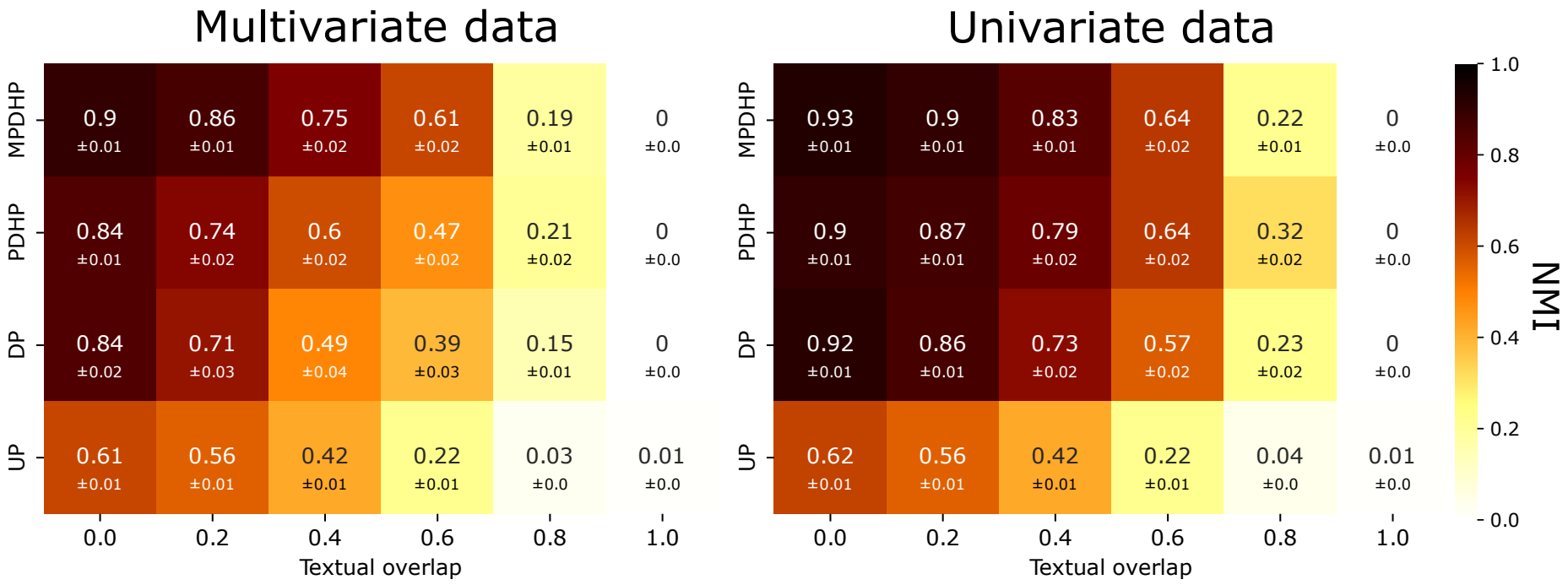
- Simulate a multivariate Hawkes process and associate words drawn from possibly overlapping vocabularies to each event
- Unless specified otherwise: 2 clusters, 20 words per event, drawn from non-overlapping vocabularies of 2000 words each.
- MPDHP coupled to a "Bag-of-words" language model (DirMult)



Numerical results

- Metric: Normalized Mutual Information (NMI)
  - 1 is perfect clustering, 0 is worst clustering
- 100 runs over 5 000 observations per situation

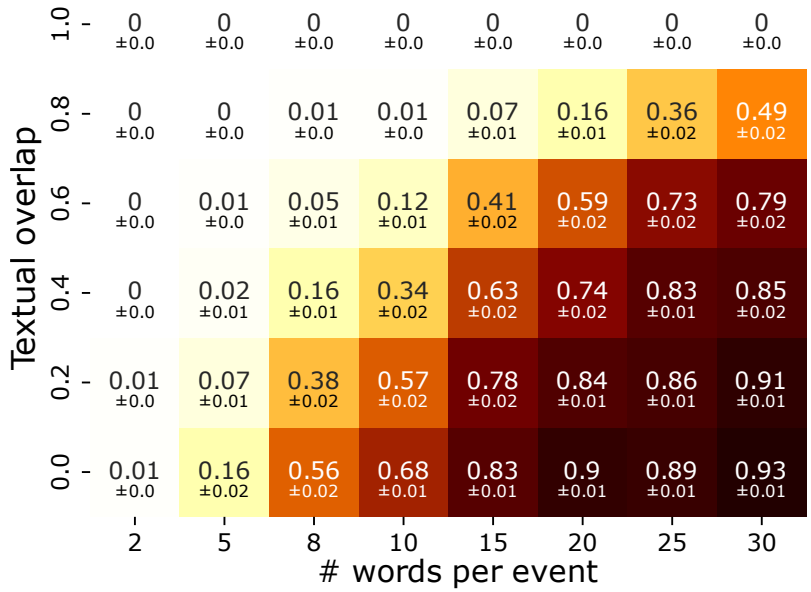
→ MPDHP outperforms other priors on both multi and univariate data



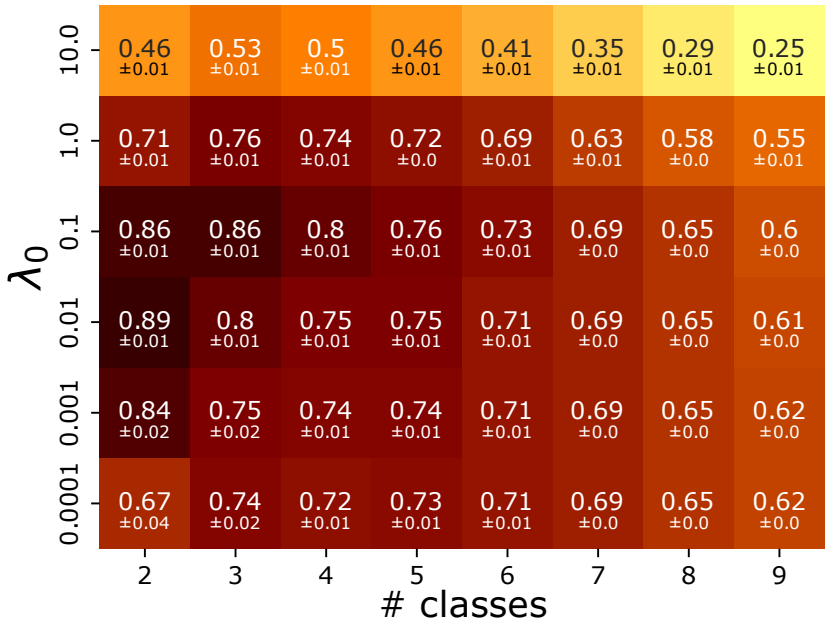
Application domain of MPDHP

- MPDHP:
  - Works with 15 words when vocabs do not overlap
  - Works with 25 words when vocabs overlap greatly
  - Is robust against the parameter λ_0 over 5 orders of magnitude

MPDHP works with few information



MPDHP works when several clusters coexist



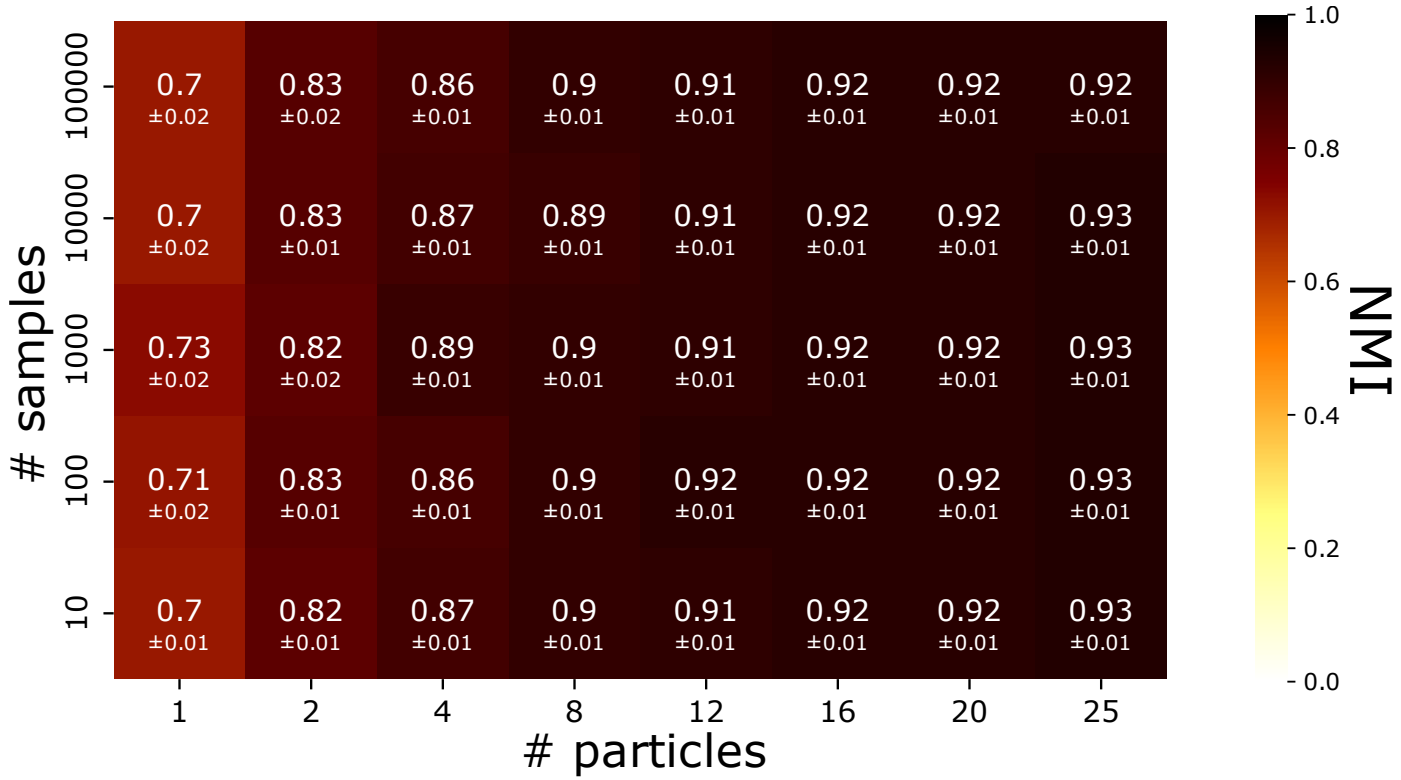
Needs few computational power

- SMC: several run in parallel, survival of the fittest

→ Few runs needed to get good results

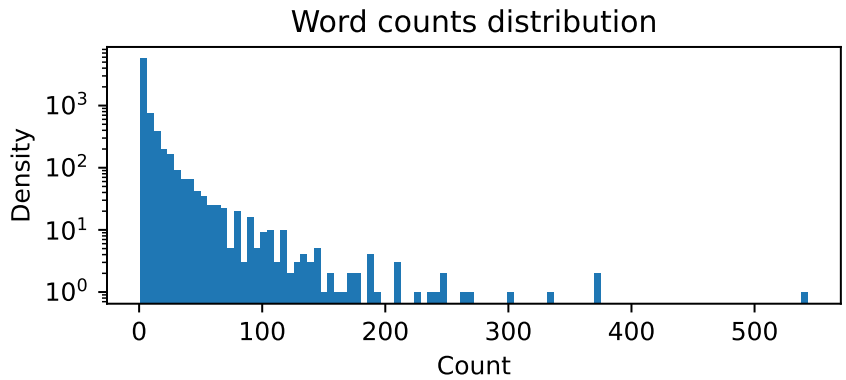
→ Few sample values to get good parameters approximation

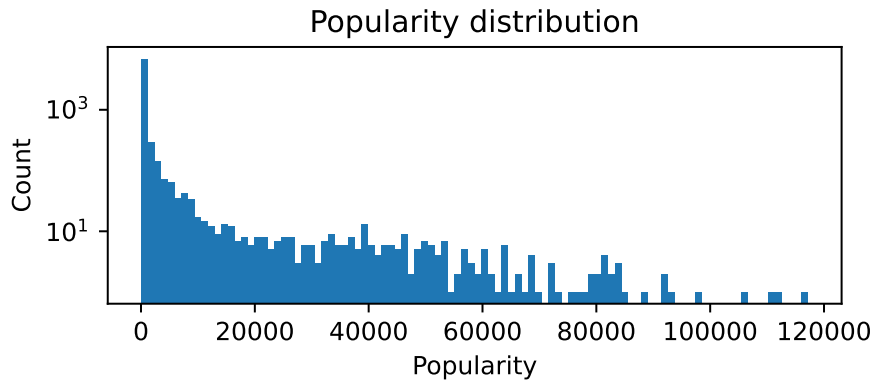
→ Needs few computational resources (optimized on a processor)

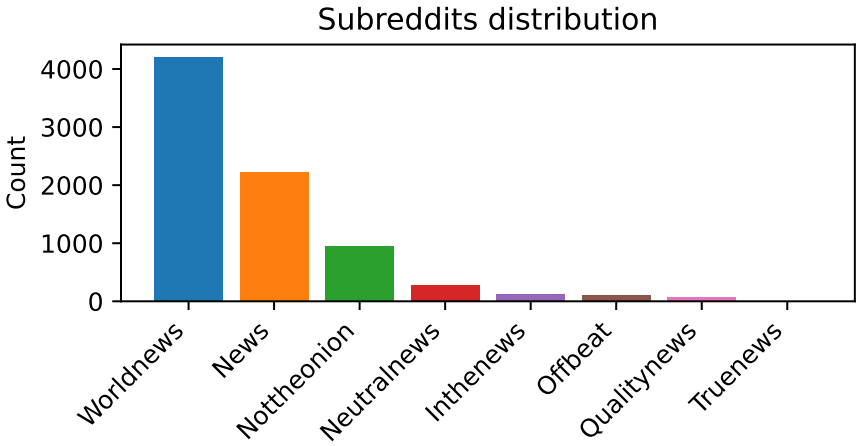


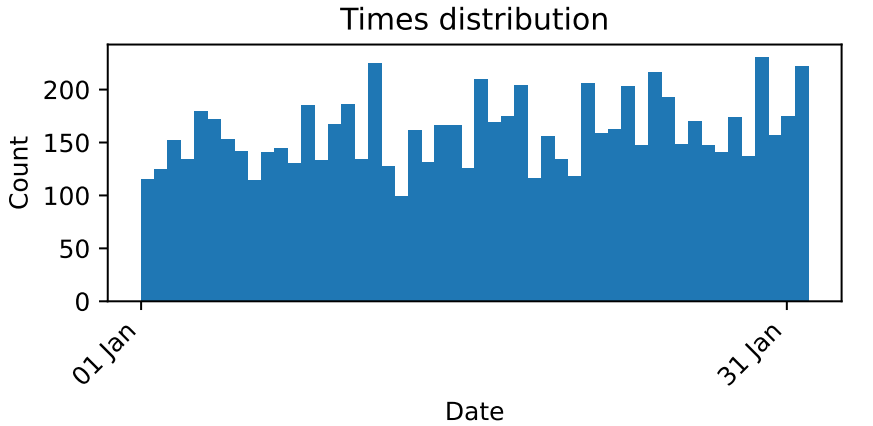
Reddit News dataset

- All 01/2019 posts from worldnews, news, nottheonion, inthenews, offbeat, qualitynews, truenews
  - $\sim$ 8 000 posts ; $\sim$ 65 000 tokens over vocab. of 7 000 words



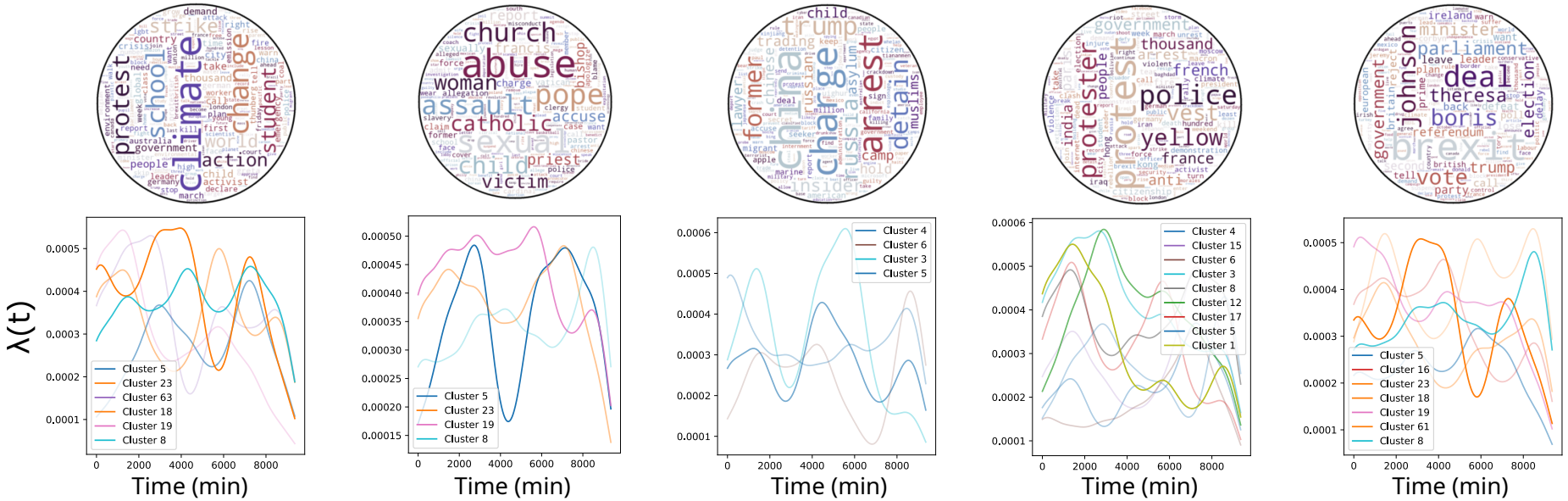






Sample of the results

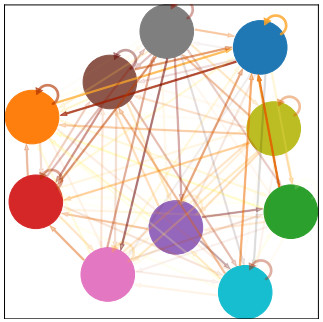
- Sample of our results for January 2019
- Uncover relevant topics and their dynamics



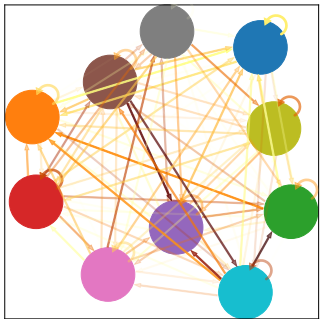
Sample of the results

- Sample of our results for January 2019
- Retrieve the network of topical influence at different times

Influence at 0h

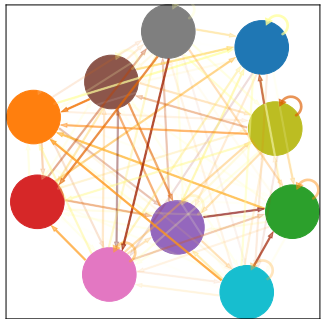


Influence at 2h



Influence at 4h

Influence at 6h



Influence at 8h

Conclusion

- Conclusion
  - Extension of Dirichlet-Hawkes processes to the multivariate case
  - Infers clusters *and* their dynamic interactions
  - Robust against challenging situations
    - Several coexisting clusters
    - Scarce textual data
    - Few resources needed for a $\mathcal{O}(N)$ SMC algorithm
- Perspectives
  - Improvements
    - Extensions to hierarchical/nested Dirichlet Processes ?
    - Couple with other (language or not) clustering models ?
  - Broad range of application
    - Summary generation
    - Moderation of online platforms / Buzz control
    - Study the dynamics of information spread

Thanks for your attention!

Slides & paper: `https://gaelpouxmedard.github.io/`
Code and data: `https://github.com/GaelPouxMedard/SIMSBM/`