Different ways for modeling time with textual data

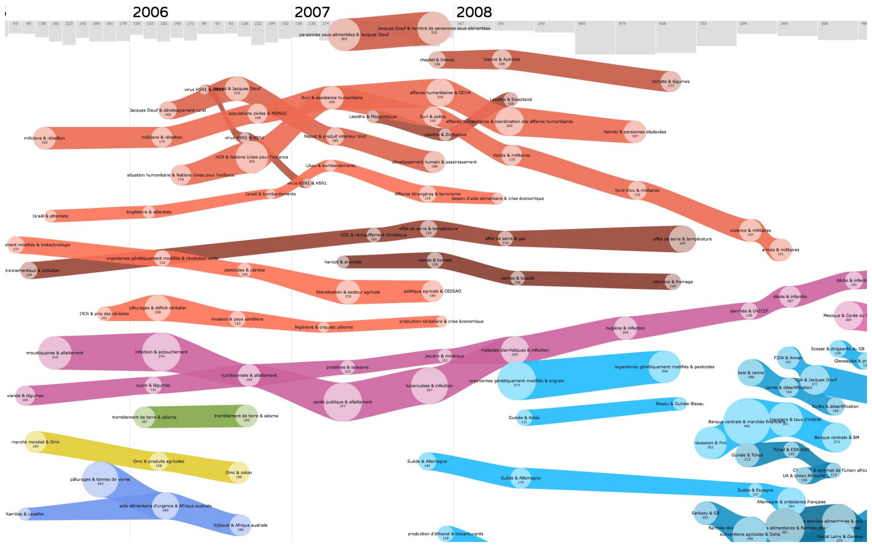
Julien Velcin and Gaël Poux-Médard
eric.univ-lyon2.fr/~jvelcin - gaelpouxmedard.github.io

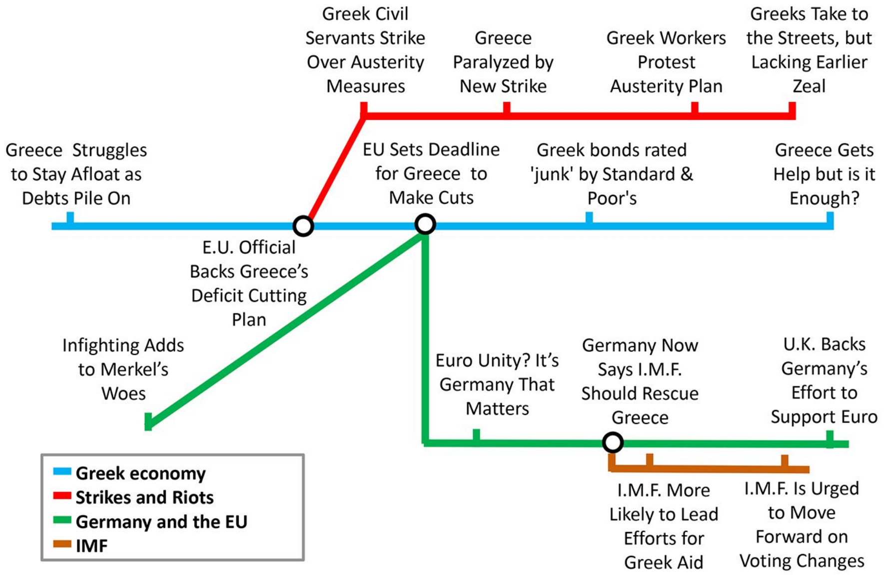
ERIC Lab, Université Lyon 2

Some context

- ERIC Lab: Univ. Lyon 1 + Lyon 2 http://eric.msh-lse.fr
  (some keywords: data science, machine learning, business intelligence, social media analysis, digital humanities…)
- The lab is a member of MSH-LSE https://www.msh-lse.fr
- Many applications to Social Sciences and Humanities
  (projects in Literature, political sciences, Archeology…)
- Two teams: SID and DMD
  (but today we will focus on DMD)

Dealing with temporal data: what for?



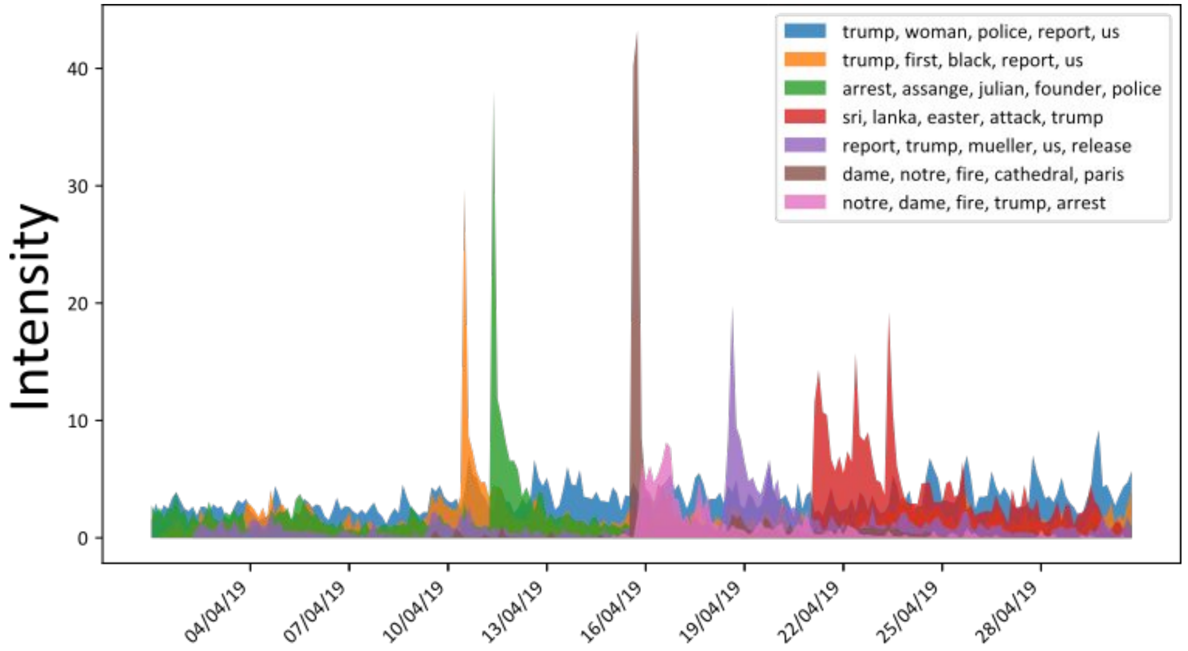


http://pulseweb.cortext.net

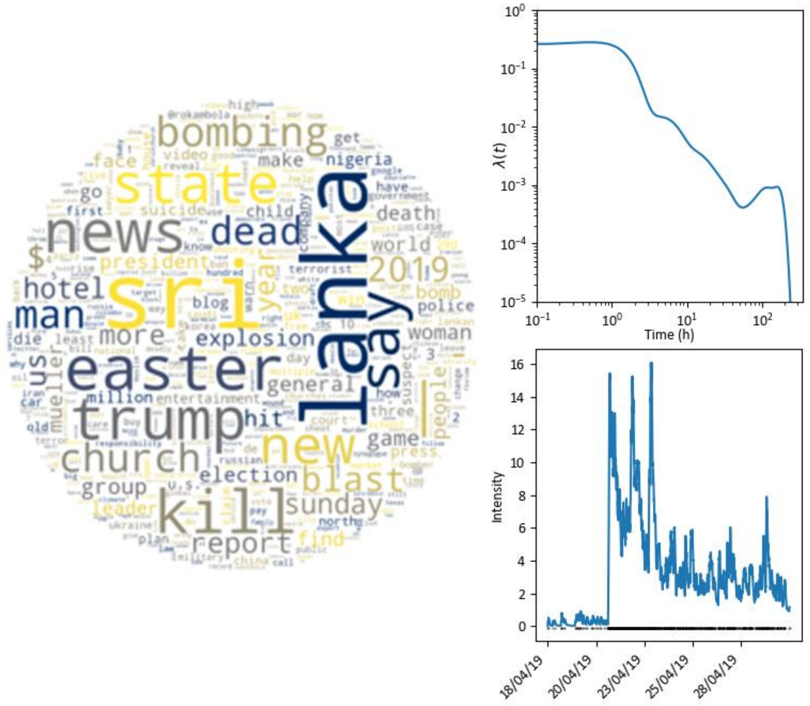
Projet Pulseweb
(Cointet, Chavalarias…)

Metromaps
(Shahaf et al., 2015)

Dealing with temporal data: what for?



Summary generation
(Reddit r/news - April 2019)



Understanding publication
dynamics (Sri Lanka
bombings, 2019)

Outline of the talk

- Previous works for clustering temporal data
- Clustering textual data over time
  - Representation learning for author embedding
  - Dynamic stochastic block models
  - Dirichlet-point processes
- Conclusion and future works

Previous works for clustering temporal data

Temporal Mixture Model

Each time step is associated
to a mixture model (clusters)

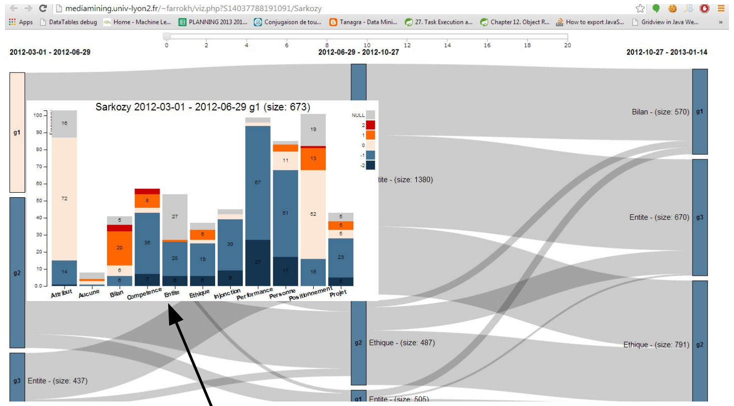
z = cluster associated to one
observation (here, a Twitter user)

Our model TMM
(Kim et al., ECIR 2015)

w = observed features (here, an
opinion expressed by the user)

t

Expected output
(project ImagiWeb)



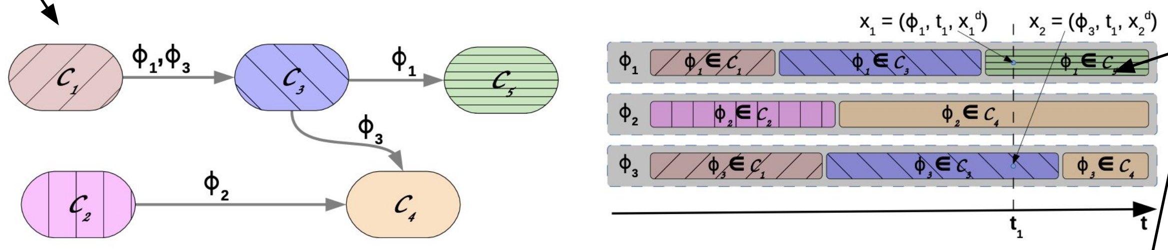
zoom on one cluster (distribution over features)

2

ClusPath

Each cluster is associated to a time span automatically based on a measure using **both** the similarity in terms of **features and time**

Our model ClusPath (Rizoiu et al., DMKM 2016)

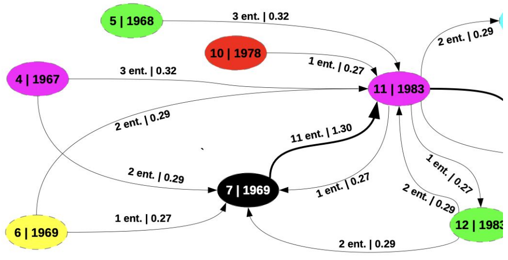


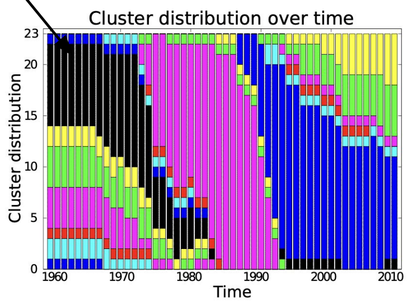
(a) (b)

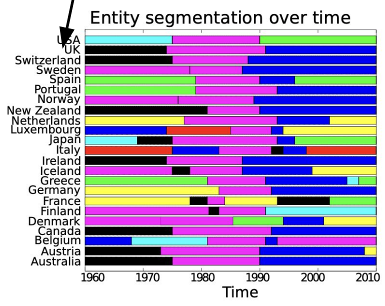
An entity (e.g., a country) φ passes from one cluster to another over time

Number of entities (here, countries) in that cluster over time

Expected output







t

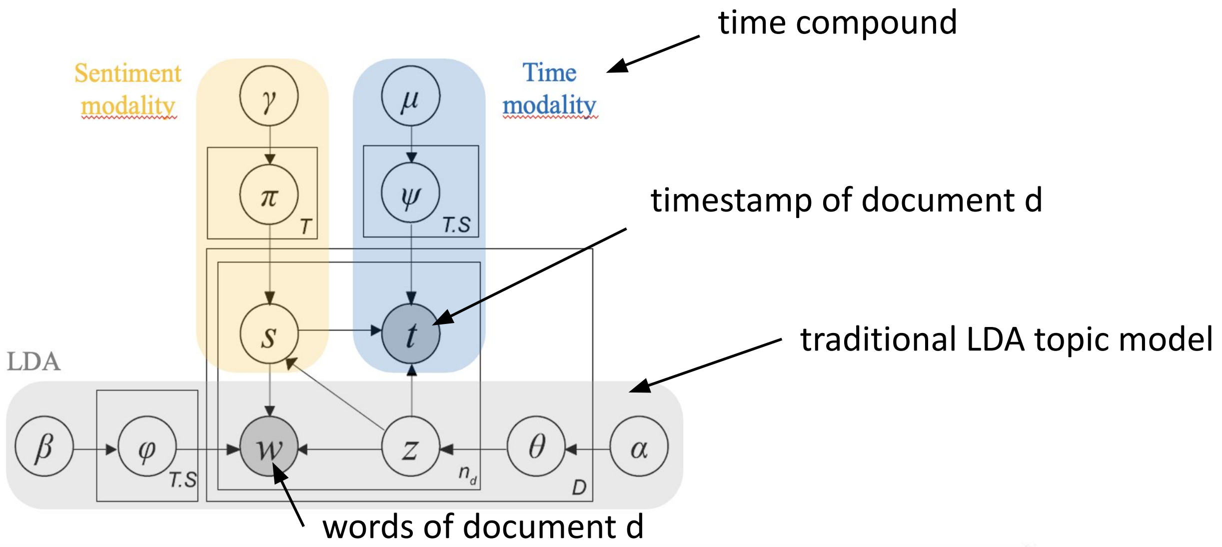
t

Excerpt of the cluster network
On a given cas study with countries

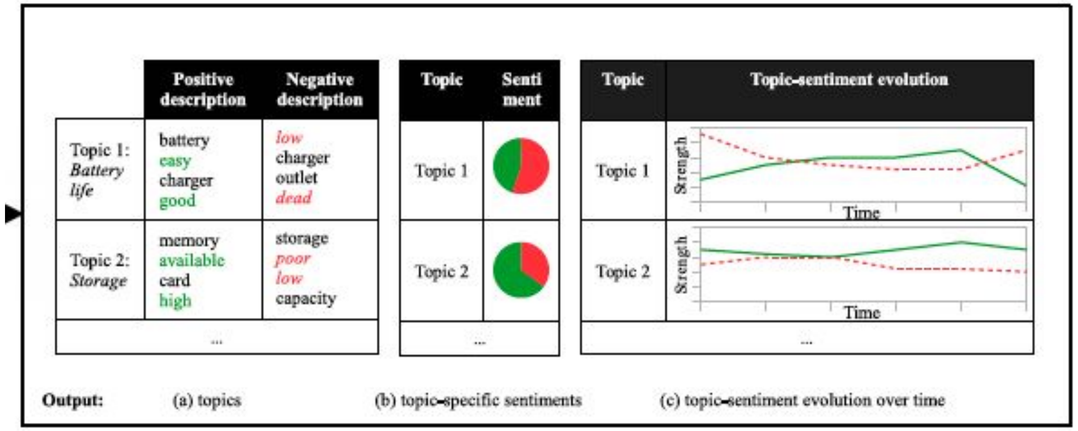
3

Time-Aware Topic Sentiment Model

Our model TTS
(Dermouche et al.,
ICDM 2014)



Expected output

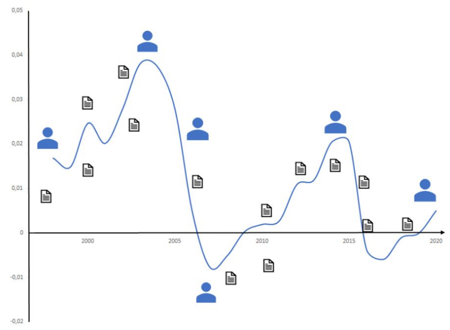


t

Clustering textual data over time

Recent work at the ERIC Lab

- Representation learning for author embedding
- **(…TBC…)**
- **(…TBC…)**



4

Representation learning for author embedding

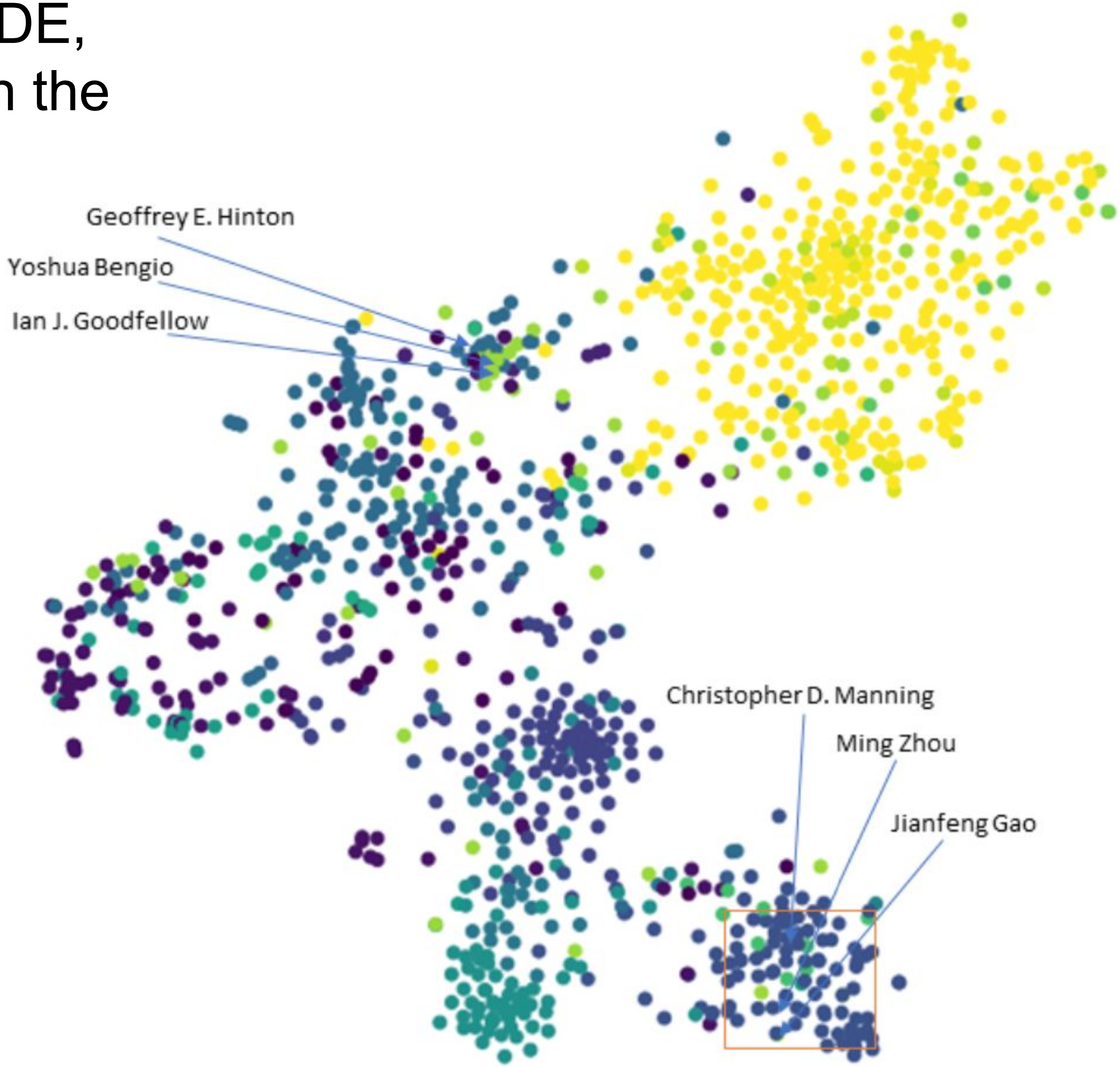
- Objective: build a joint vector space where we can place authors and documents
- Previous models of the literature include:
  ATM (Rosen-Zvi et al., 2004), aut2vec (Ganguly et al., 2016), D-CODE (Sarkar et al., 2007), DAR (Delasalles et al., 2019)
- Main ideas:
  - leverage existing pretrained sentence embeddings (e.g., USE)
  - model the dispersion around a mean vector
- Applications to scientific watch, recommandation, identification of most likely authors…

VADE

Static model based on the VIB framework

First contribution: VADE, a static model based on the VIB framework

(Gourru et al., EGC 2021)

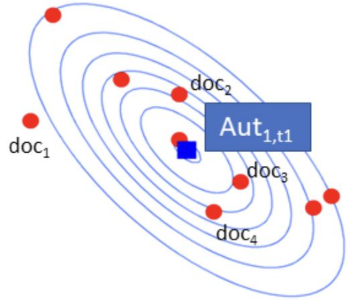


Dynamic Gaussian embedding

Dynamic Gaussian Embedding of Authors

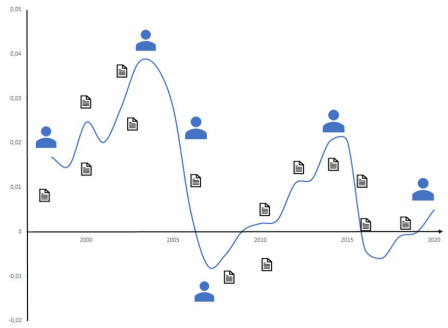
(Gourru et al., to appear in WWW 2022)

- Two main hypotheses:
  - Vector v_d for document d written by author a is **drawn from a Gaussian** $G_a = (\mu_a; \Sigma_a)$



  - There is a **temporal dependency** between G_a at time t, noted $G_a(t)$, and the history $G_a(t\text{-}1, t\text{-}2\ldots)$:
    - probabilistic dependency based on t-1 only (K-DGEA)
    - functional dependency based on the full history (R-DGEA)

  (note that the two versions of the model need different optimization techniques: Kalman filters for K-DGEA, and gradient-based for R-DGEA)



Evaluation of DGEA

Articles from the New York Times

Abstract of scientific papers

	NYT	S2G
ATM	34.4 (1.3)	33.5 (1.6)
Aut2Vec	34.5 (1.6)	24.6 (1.9)
Usr2Vec	43.8 (2.4)	36.9 (1.4)
DAR (statique)	43.3 (1.7)	40.4 (1.1)
DAR (dynamique)	27.1 (3.0)	33.2 (1.5)
DAR (concat)	44.1 (1.5)	39.8 (1.5)
N-DGEA_I	44.9 (2.2)	35.2 (1.3)
K-DGEA_I	53.1 (2.3)	40.5 (1.2)
R-DGEA_I	53.2 (1.8)	**42.7** (1.9)
N-DGEA_S	36.5 (2.6)	27.5 (1.4)
K-DGEA_S	44.8 (2.0)	30.9 (1.9)
R-DGEA_S	49.5 (2.5)	33.7 (1.5)
N-DGEA_U	45.1 (2.4)	31.4 (1.4)
K-DGEA_U	**54.2** (2.4)	37.2 (1.3)
R-DGEA_U	52.4 (3.0)	39.7 (1.6)

Task 1: classification of authors (Micro-F1, to be maximized)

We tested several sentence embedding techniques

ATM		40.78	
Aut2Vec		30.36	
Usr2Vec		23.22	
DAR (statique)		**21.49**	
DAR (dynamique)		42.25	
DAR (concat)		33.17	
Method	SBERT	InferSent	USE
N-DGEA	31.39	33.82	30.69
K-DGEA	24.91	26.97	23.96
R-DGEA	22.67	22.89	21.55

Task 2: link prediction on the graph of collaborations between authors (covering error, to be minimized)

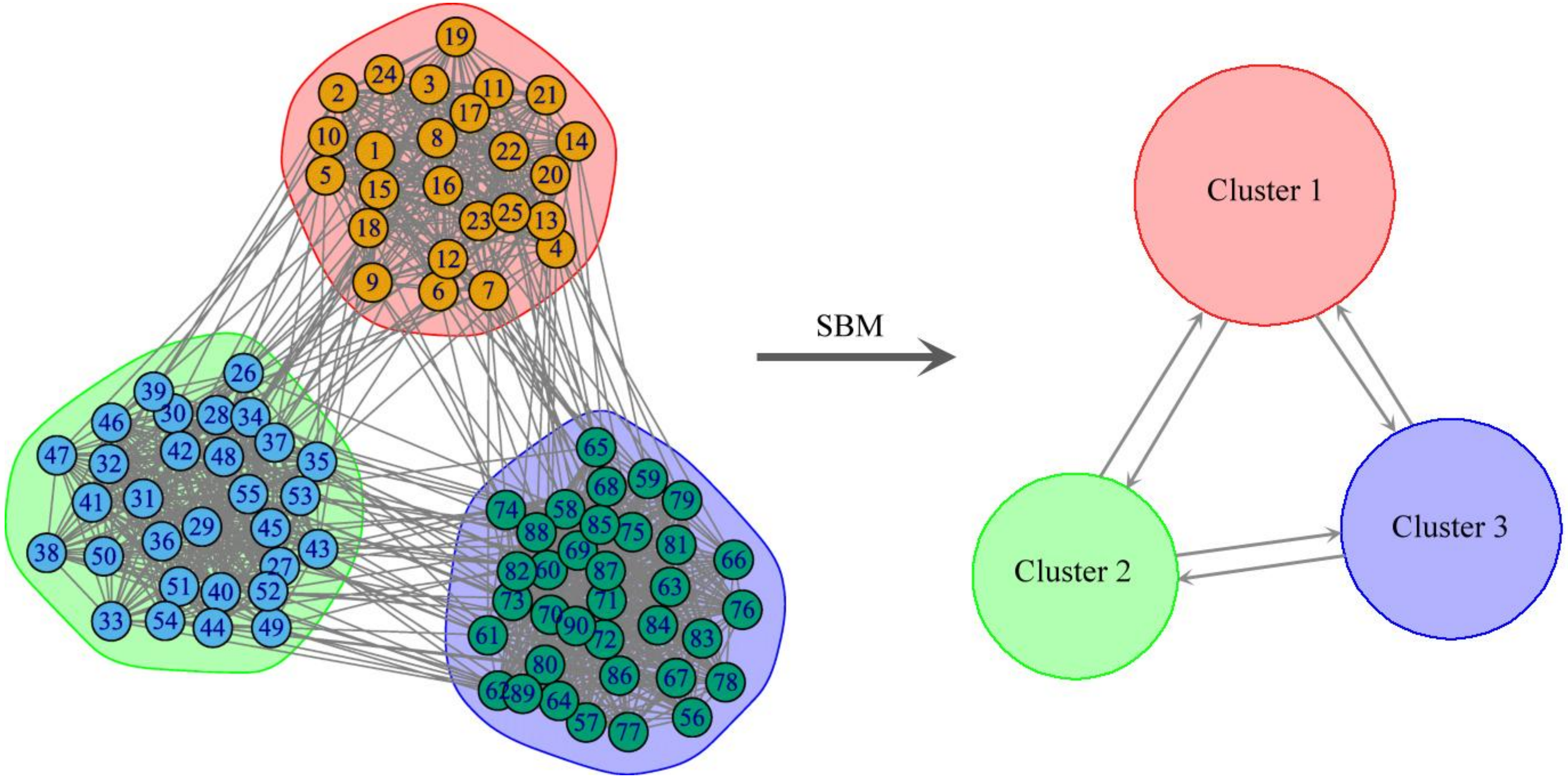
Dynamic block models

Inferring clusters that can vary in time

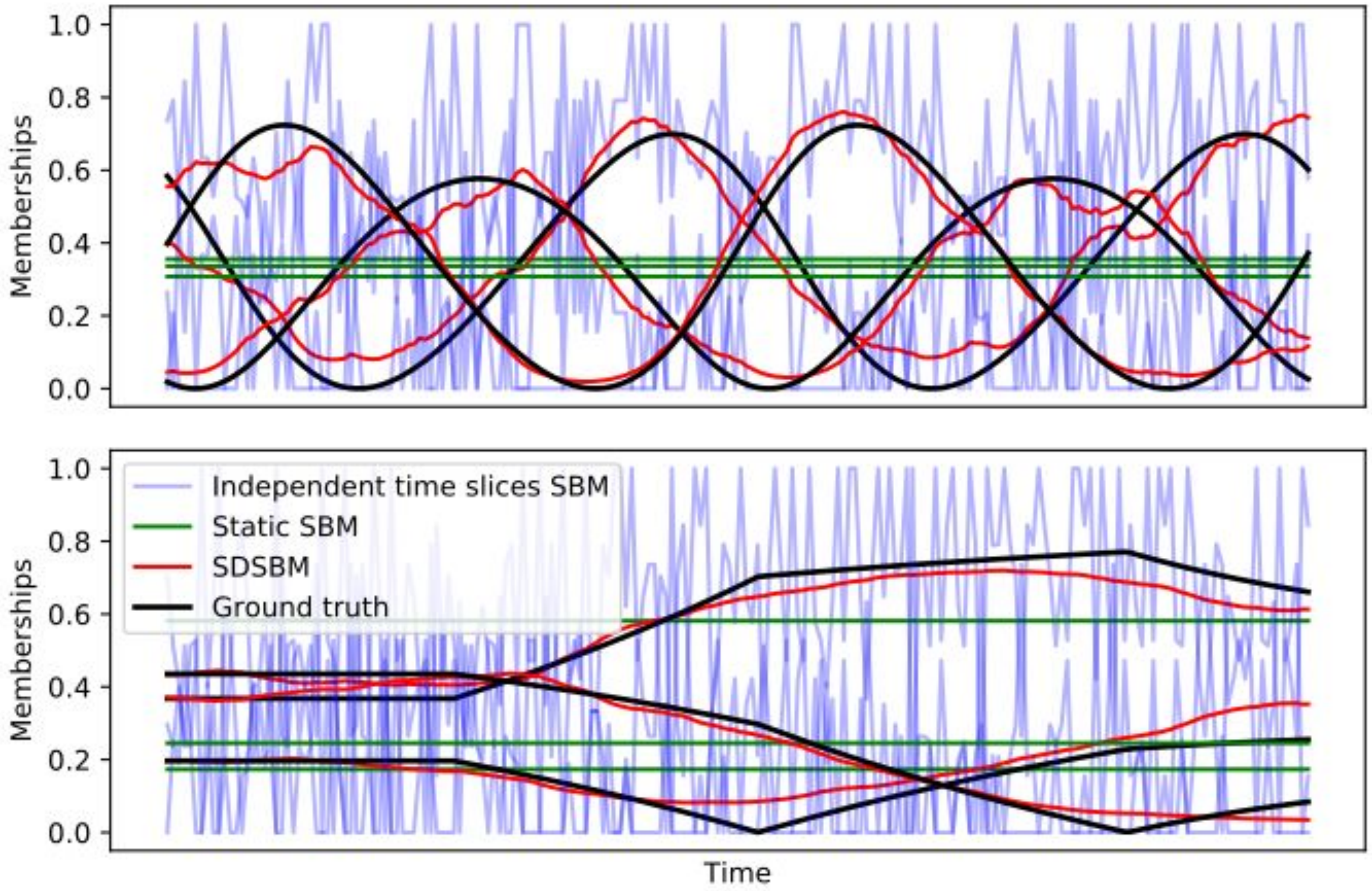
5

A block modelling approach

- Stochastic Block Models
  - One node = one document ; links can be citations, similarities, etc.



Dynamic block modeling



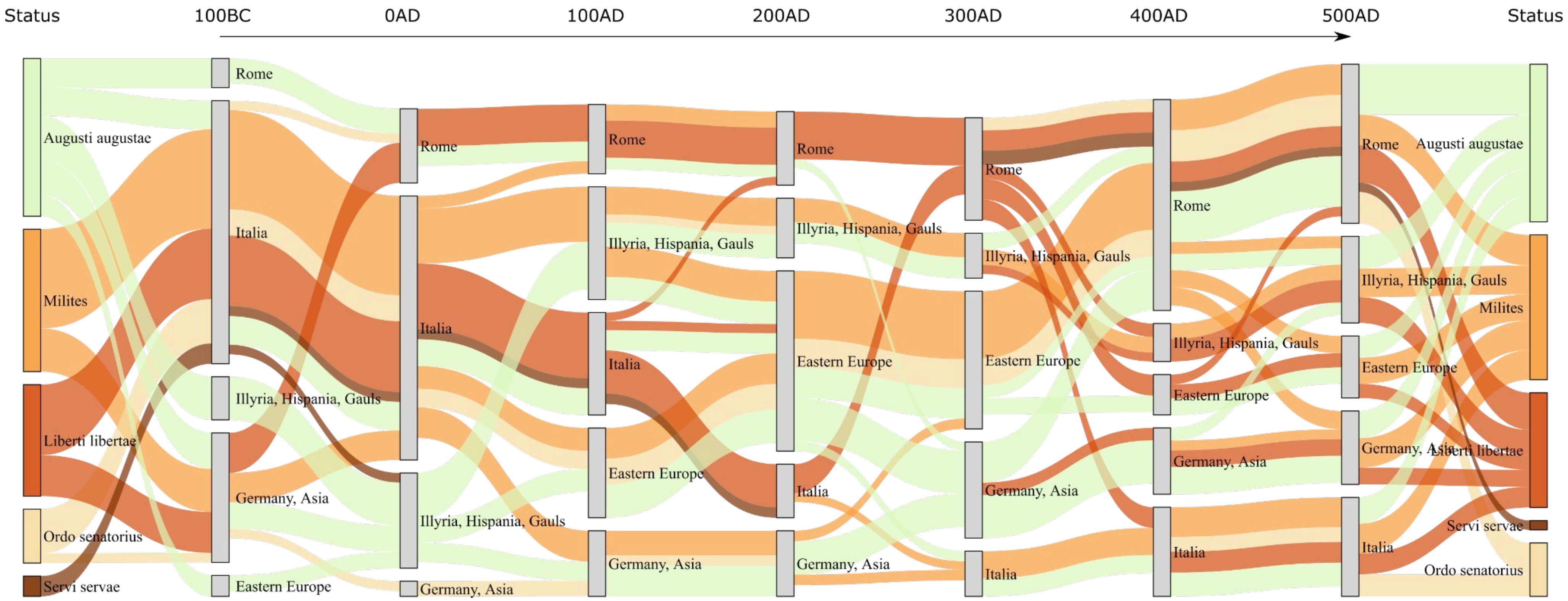
Features and evaluation

- Few observations needed in each time slice
- Linear complexity
- Minimal assumptions
  - Block structure
  - No abrupt variations

		ROC	AP	NCE
Epi	SDSBM	**0.9025(11)**	**0.3700(17)**	**0.1151(11)**
	NC	0.8420(22)	0.3435(36)	0.1582(19)
	MMSBM	0.8597(12)	0.2141(16)	0.1451(13)
Lastfm	SDSBM	**0.8942(8)**	**0.0168(1)**	**0.1284(11)**
	NC	0.8393(5)	0.0157(2)	0.1785(7)
	MMSBM	0.8647(5)	0.0115(2)	0.1493(4)
Wiki	SDSBM	**0.9759(2)**	**0.0659(9)**	**0.0459(3)**
	NC	0.9092(7)	0.0608(10)	0.1195(8)
	MMSBM	0.9576(7)	0.0622(4)	0.0565(8)
Reddit	SDSBM	**0.9803(3)**	**0.4295(54)**	**0.0312(3)**
	NC	0.8508(5)	0.3598(17)	0.1846(7)
	MMSBM	0.9798(2)	**0.4269(40)**	0.0322(3)

Possible output

- Geographic distribution of roman social status over time
  - Document : latin inscription mentioning a status
  - Links : between a document and a region of the empire



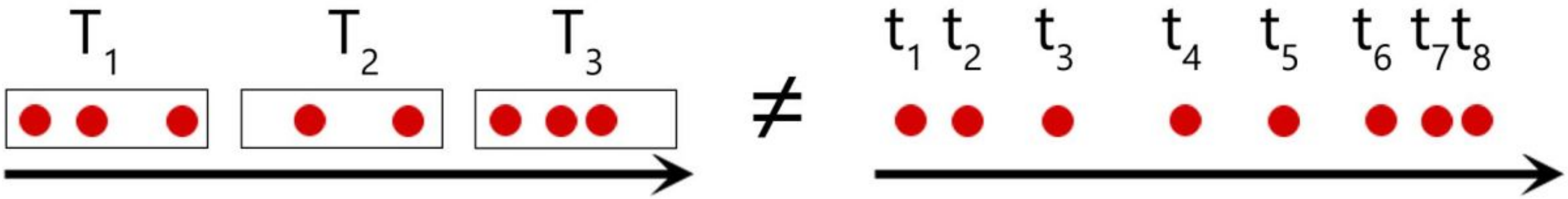
Dirichlet-point processes

Explicitly modeling time

6

About dataset generation

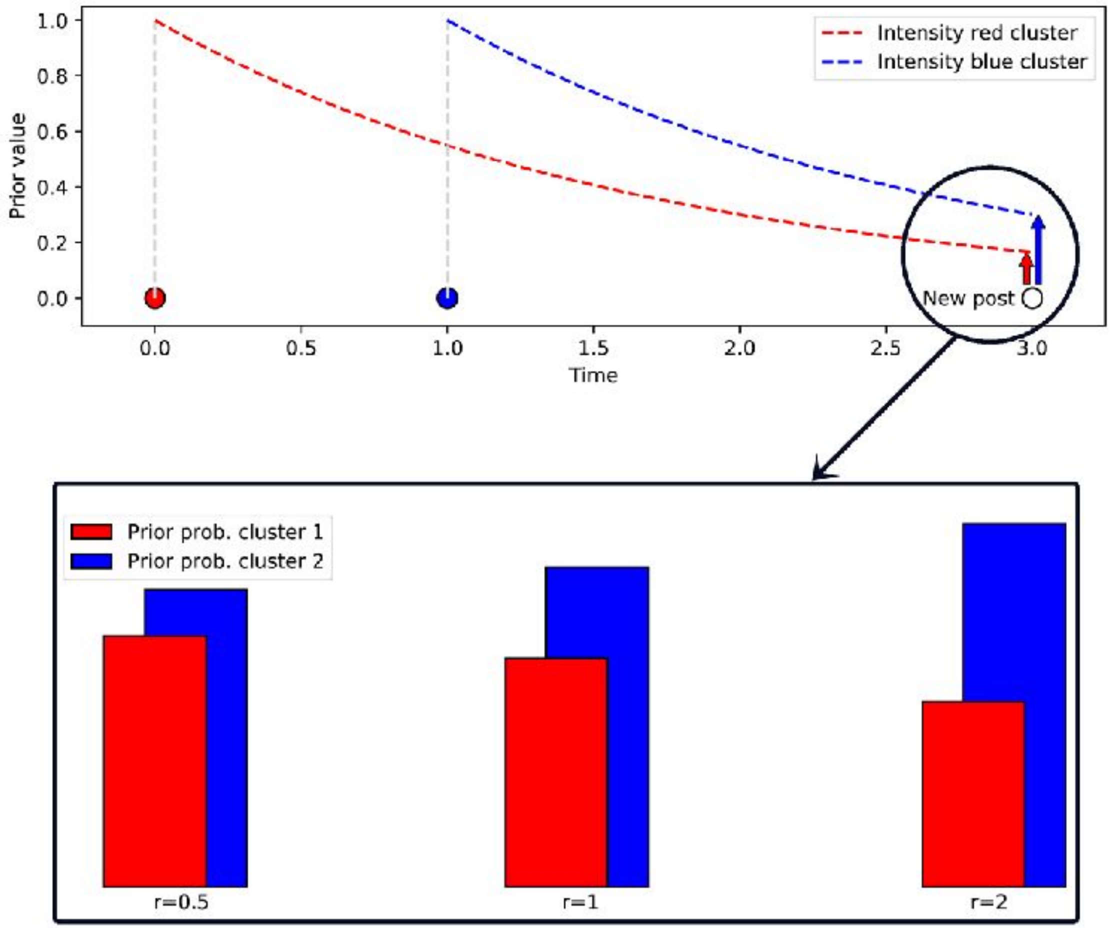
- Slicing time into episodes can be biased
- Most models do it
  - The ones previously introduced in this presentation
  - DTM (Blei06), ToT (Wang06), RCRP (Amr08), DDCRP (Blei10), etc.



(Powered) Dirichlet-Hawkes process

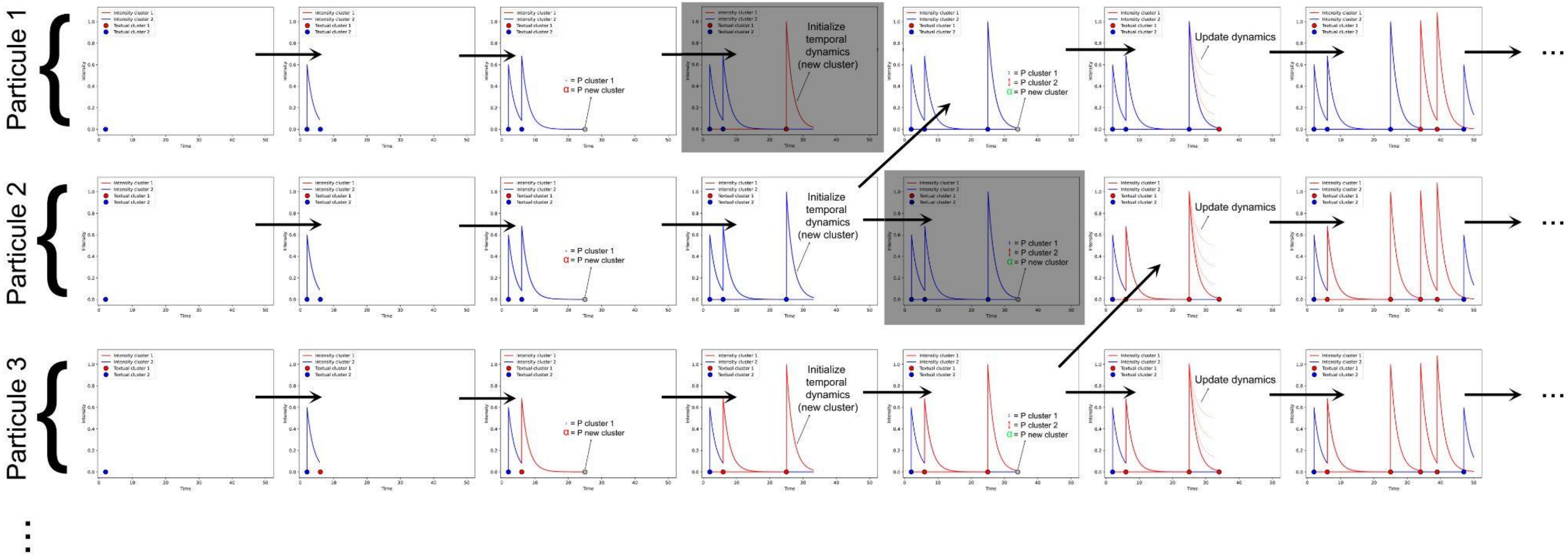
(Poux-Médard et al., ICDM 2021), (Du et al., KDD, 2015)

- Clusters self-replicate
- Data:
  - Textual content
  - Publication date
- Output:
  - Documents' cluster
  - Clusters temporal intensity
- Powered version (ours)
  - Handle challenging cases
  - Relax perfect correlation hyp.



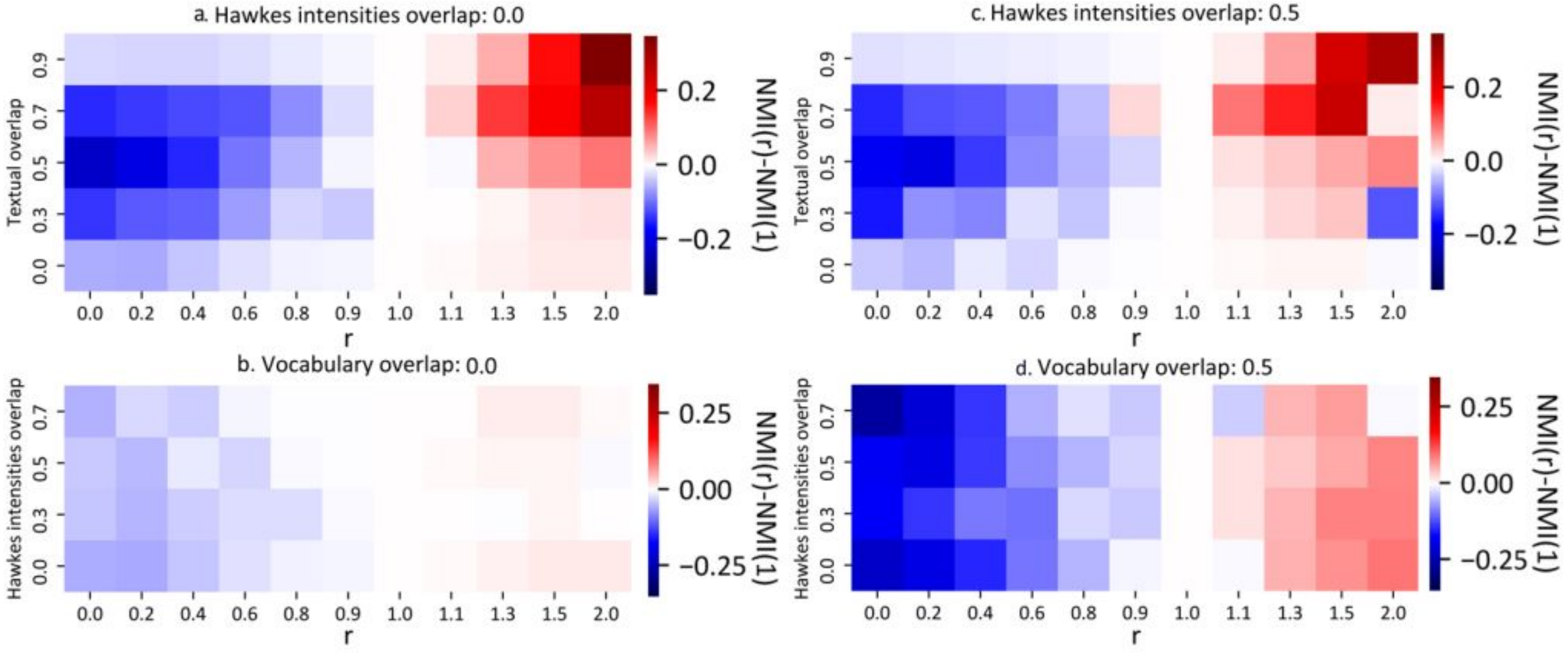
Sequential Monte-Carlo inference

- Data is treated sequentially
- New clusters can be opened
- The algorithm explores the space of possible clustering

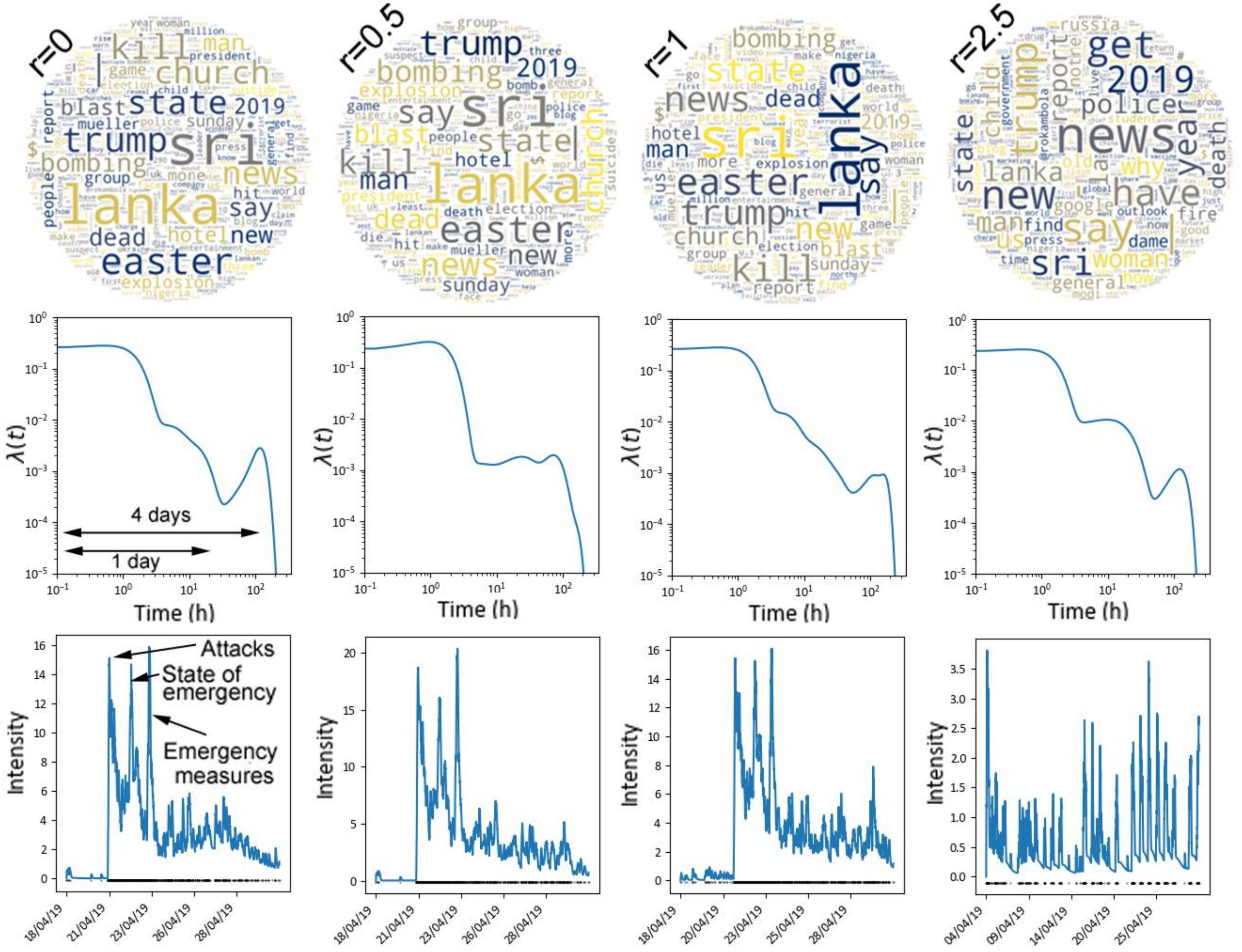


Evaluation

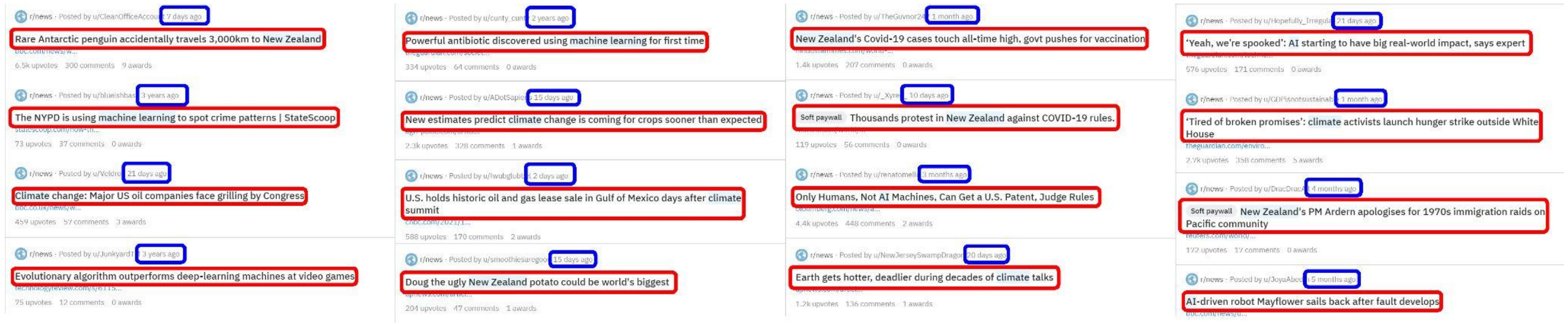
- Evaluation w.r.t (Du2015)
- Large overlaps = challenging situations
- r is a hyperparameter that allows to account for challenging cases



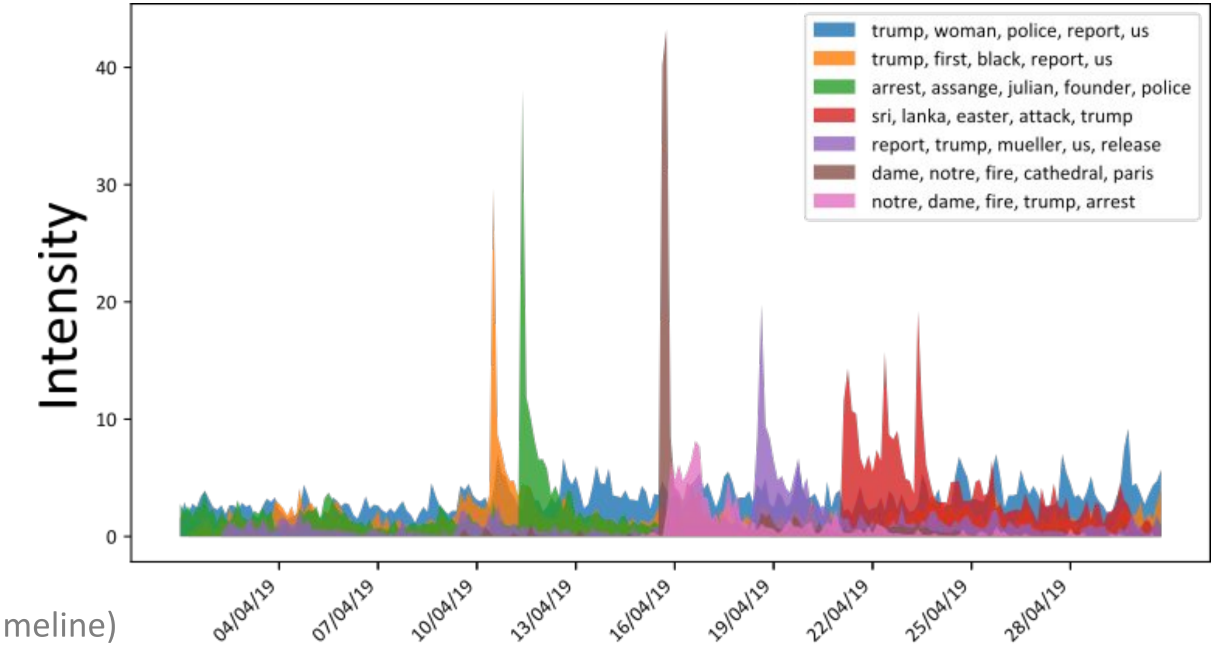
Output example: Sri Lanka 2019 bombings



Generating summaries

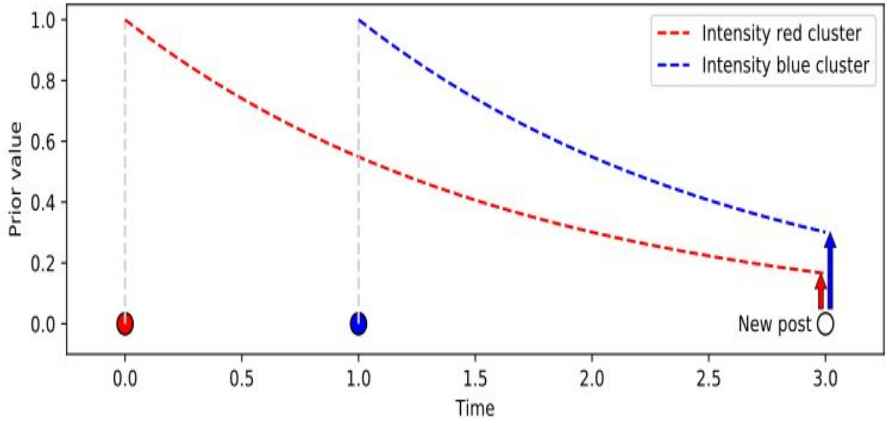


(Raw Reddit posts)

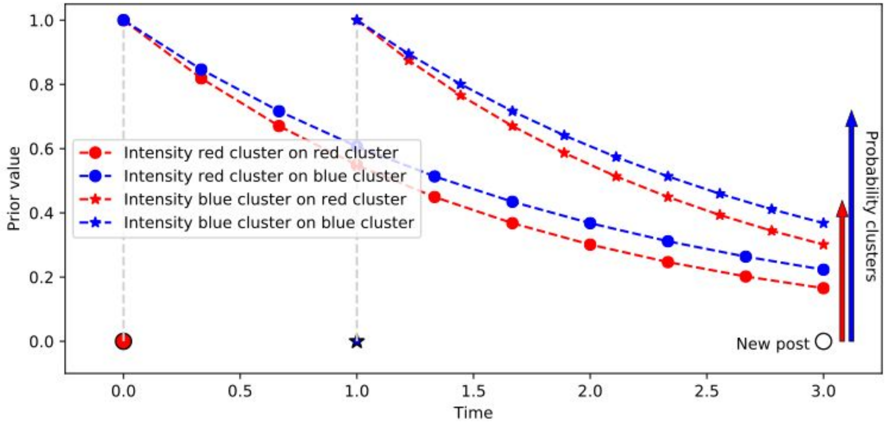


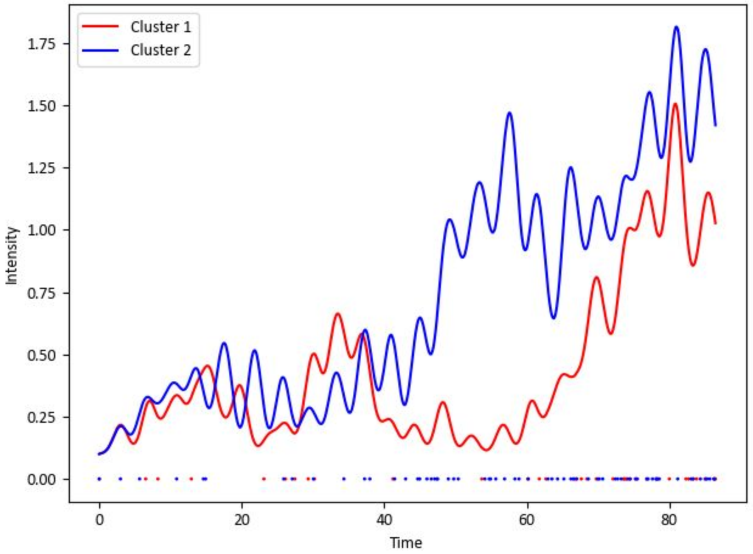
(Inferred timeline)

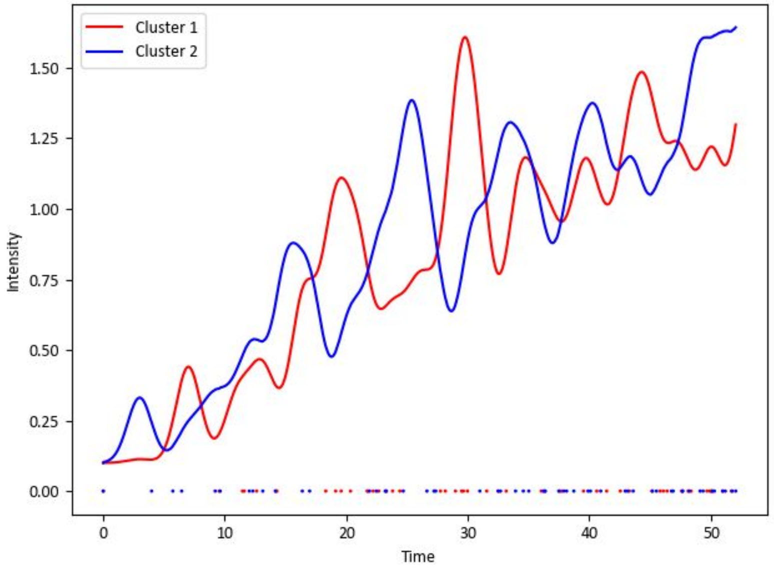
Going further - Multivariate case



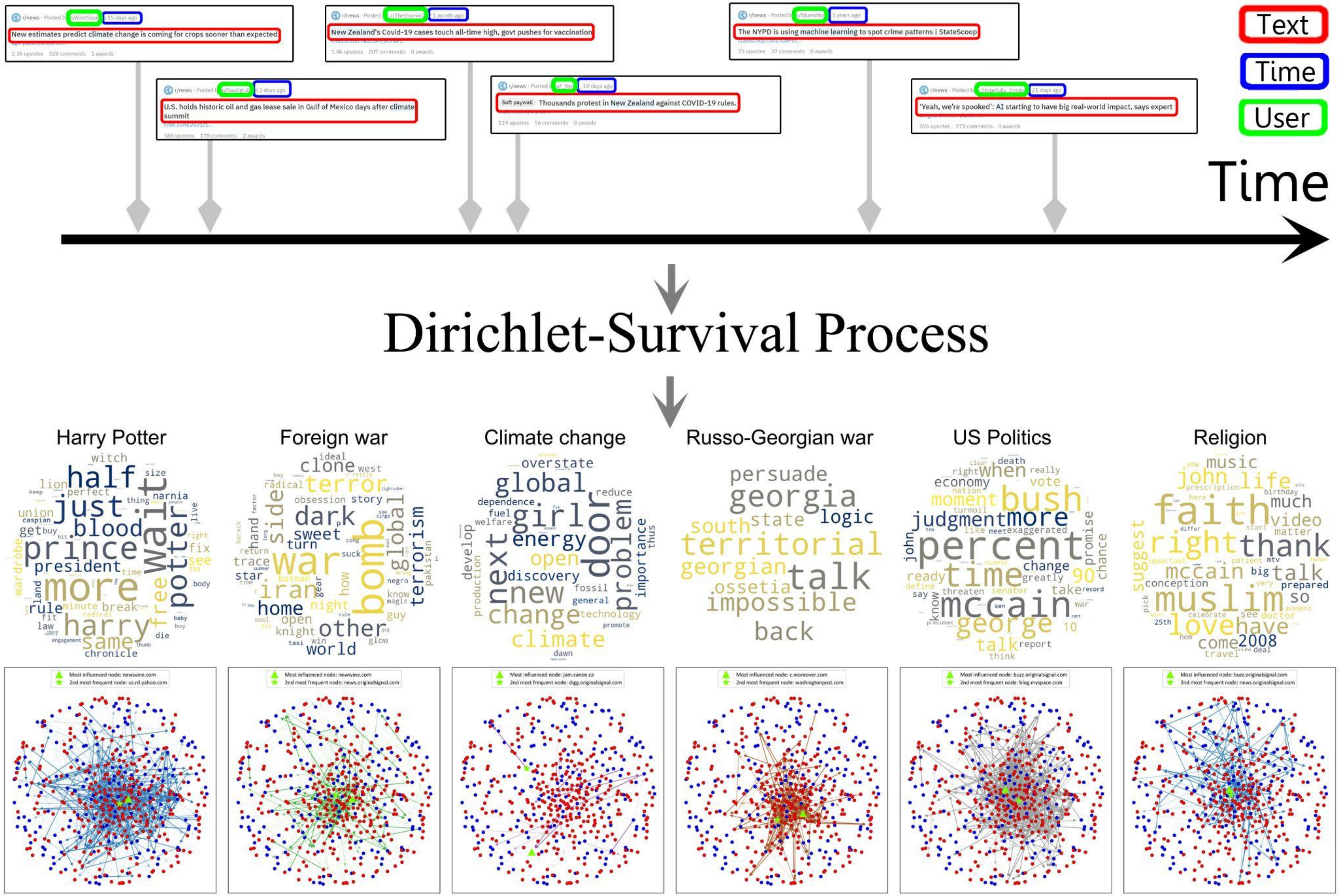
Multivariate







Going further - Dirichlet-point processes



Conclusion and future works

Conclusion

- Different ways to model time:
  - discrete time slices
  - time as a continuous variable
- Different kinds of models:
  - retrospective models
  - adaptive models over time
- Influence of textual information

Current projects

- Project **LIFRANUM** (ERIC Lab, MARGE, BnF): Identify and structure the corpus of digital French literatures
- Projet **POIVRE** (ERIC Lab - EDF): Viewpoint detection on energy issues through Twitter

Many thanks to…

- Mohamed Dermouche
- Marian-Andrei Rizoiu
- Young-Min Kim
- Antoine Gourru

References

M. Dermouche, J. Velcin, S. Loudcher, L. Khouas: A Joint Model for Topic-Sentiment Evolution over Time (ICDM 2014)

Y. M. Kim, J. Velcin, S. Bonnevay, M. A. Rizoiu: Temporal Multinomial Mixture for Instance-oriented Evolutionary Clustering (ECIR 2015)

M.A. Rizoiu, J. Velcin, S. Bonnevay, S. Lallich: ClusPath: A Temporal-driven Clustering to Infer Typical Evolution Paths (DMKD 2016)

A. Gourru, J. Velcin, C. Gravier, J. Jacques: Dynamic Gaussian Embedding of Authors (WWW 2022).

G. Poux-Médard, J. Velcin, S. Loudcher: Powered Dirichlet-Hawkes Process (ICDM 2021)