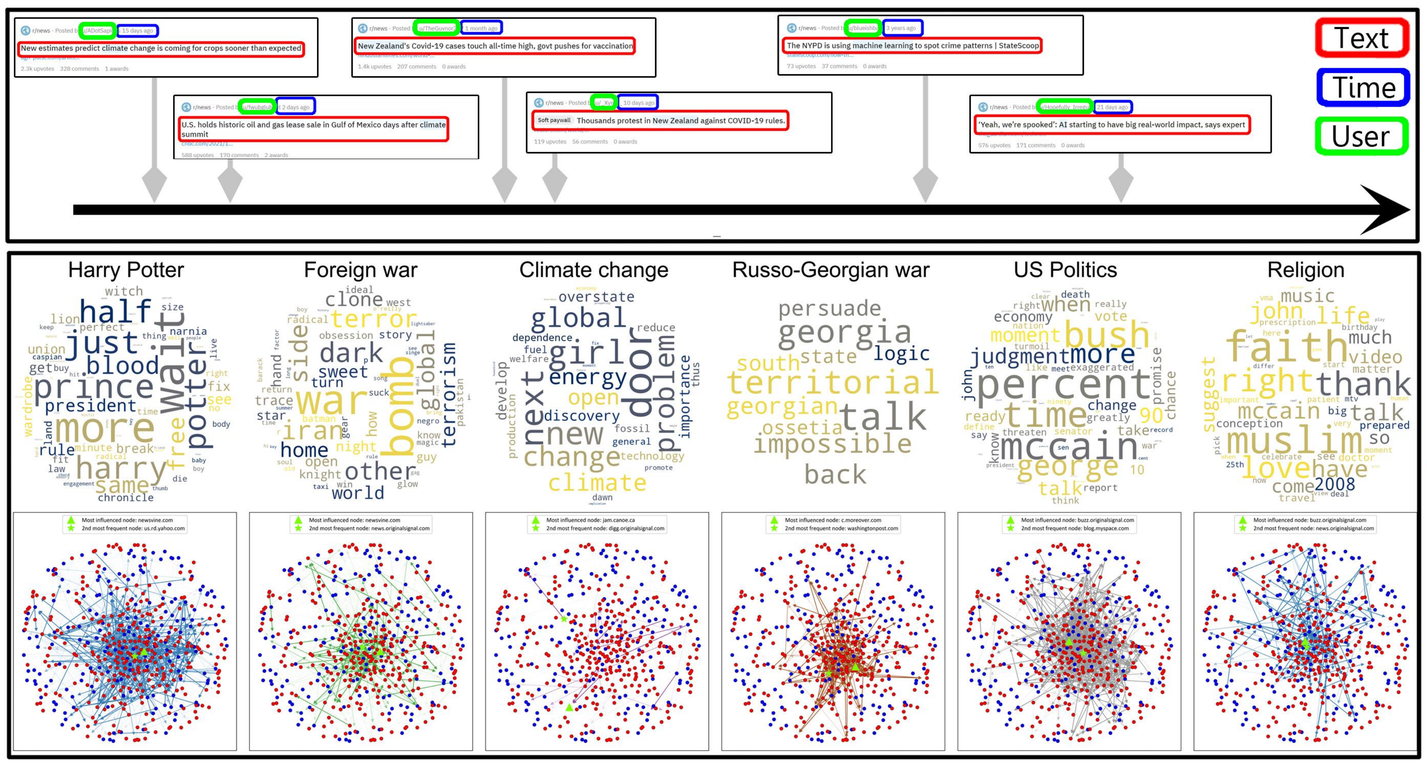
Dirichlet-Survival Process

Scalable Inference of Topic-Dependent Diffusion Networks

Gaël Poux-Médard[a], Julien Velcin[b], Sabine Loudcher

[a] https://gaelpouxmedard.github.io/
[b] https://eric.univ-lyon2.fr/jvelcin/

From a stream of textual documents spread by users at given times, our approach retrieves meaningful topics as well as the underlying topic dependent diffusion networks (here on the Memetracker dataset)
No information about the network's structure of cluster content is provided to the model. Our approach is nonparametric, meaning that the number of topics also inferred automatically.
The model is trained using a Sequential Monte Carlo algorithm.



Dirichlet-Survival Process + Language model

A dynamic clustering prior

We consider a **stream of user-generated timestamped documents** as input data. The idea is to group documents into clusters using the user who published them, their relative publication dates, and their textual content. We want to build a model that sequentially consider new documents as they appear in the data stream.
Using Bayes theorem, the posterior probability for a new document to belong to a cluster reads:

$$\underbrace{P(cluster|text,time)}_{Posterior\ probability} \propto \underbrace{P(text|cluster)}_{\substack{Likelihood \\ (language\ model)}} \underbrace{P(cluster|time)}_{\substack{Dynamic\ prior \\ (Dirichlet-Survival)}}$$

Traditionally in clustering models, the prior distribution is expressed as a Dirichlet process; the prior probability to belong to a cluster is proportional to the number of observations already in that cluster. It has later been extended to consider counts that vanish over time, which gave rise to a whole new class of processes: the Dirichlet-Point Processes[1].
Based on these advances, we design a prior that jointly consider the time and source of any publication. In previous works, Dirichlet Processes are coupled to a Hawkes point process. Here, we explore the junction between a Dirichlet Process and Survival Analysis. In particular, we join it to the NetRate model[2], which can be expressed as a point process. The resulting **Dirichlet-Survival prior**[3] allows to model dynamic topic-dependent underlying diffusion networks.

While most of existing diffusion models consider documents' textual **content, source and date** separately or sequentially, our approach considers all three of them jointly. Most importantly, it can run sequentially, meaning we can increasingly provide new data to the model as time passes.

Key points

- Infers topics based on content, time and source
- Infers one hidden diffusion network per topic
- Online optimization (SMC algorithm)
- Bayesian prior that can be used with other models
- Special case of yet unexplored Dirichlet-Point proc.
- Not the best state of the art method BUT novel approach to diffusion problems

Dirichlet-Survival process

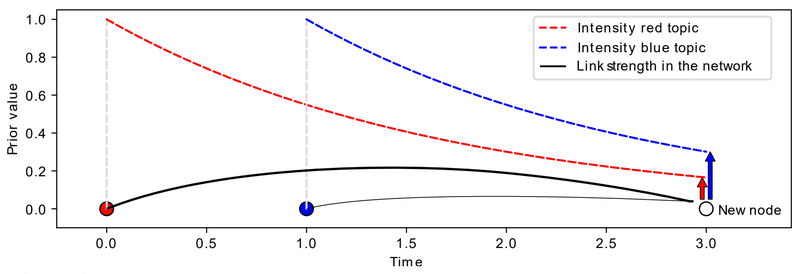


Figure 2 – A node strongly connected to the new node in the network has been contaminated by the red information at time t=0. Another node weakly connected to the new node in the network has been contaminated by the blue information at time t=1.
The interplay between these quantities give the prior probability that the new node is infected by either information.
Typically, the prior probability of getting infected equals the negative exponential of the product of time difference with link strength
→ $\lambda(t) \propto \sum_{j \to i} e^{-\alpha_{i,j}^{c}(t_j - t_i)}$, with $\alpha_{i,j}^{c}$ the link strength from j to i for topic c

Results

		Houston	NRxDM	DHP	NetRate
PL	NMI	**0.809**	0.669	0.449	-
	ARI	**0.688**	0.330	0.063	-
	AUC-ROC	**0.807**	0.719	-	0.731
	F1	**0.199**	0.106	-	0.005
	MAE	**0.267**	0.338	-	0.460
ER	NMI	**0.787**	0.711	0.638	-
	ARI	**0.631**	0.488	0.411	-
	AUC-ROC	**0.849**	0.800	-	0.659
	F1	**0.263**	0.176	-	0.005
	MAE	**0.229**	0.278	-	0.481
Blogs	NMI	**0.750**	0.668	0.372	-
	ARI	**0.609**	0.365	0.023	-
	AUC-ROC	0.701	0.613	-	**0.710**
	F1	**0.168**	0.087	-	0.005
	MAE	**0.374**	0.444	-	0.499

Figure 3 – Houston (our model) outperforms models that consider only time and content (DHP) or network structure (NetRate), or all three of them sequentially (NRxDM). The NMI and ARI evaluate how well topics are retrieved ; AUC-ROC, F1 and MAE evaluate how well the underlying network's edges have been retrieved.
Visuals of the reconstructed networks are presented in Fig.4

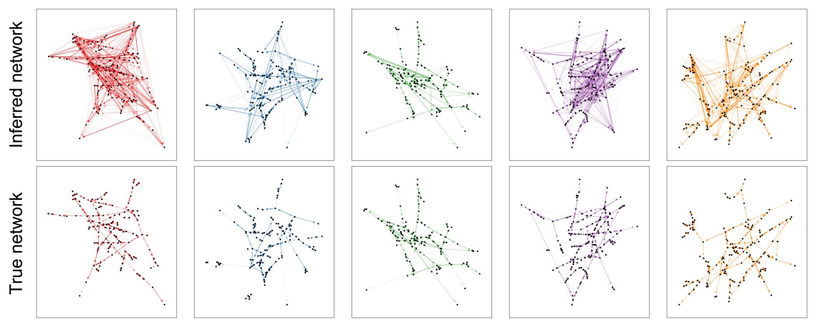


Figure 4 – True network and inferred networks on which topics spreads. Note that several topics can spread through a same node. The top networks have been inferred using only a sequence of textual content, date of publication and source user.

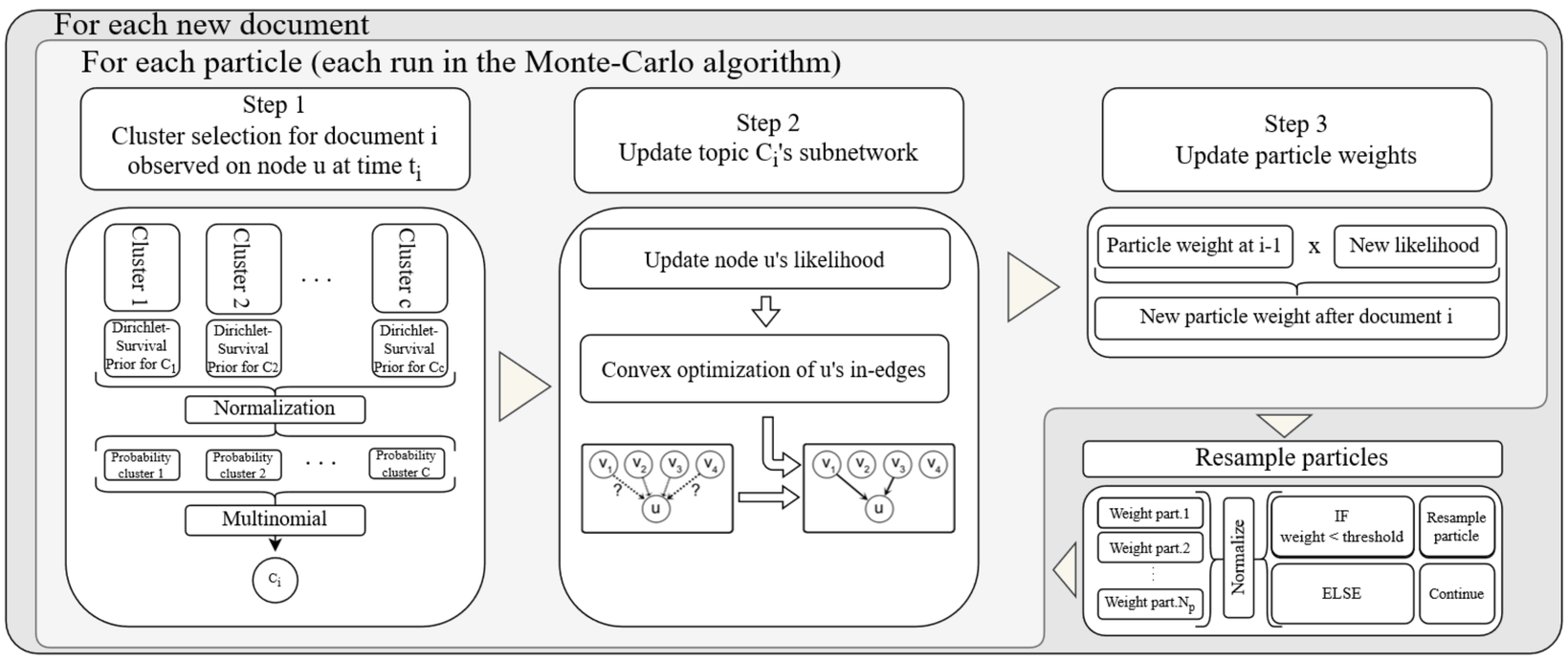


Figure 5 – Sequential Monte Carlo inference procedure. The Dirichlet-Survival prior intervenes in step 1, where it is multiplied by the likelihood of the associated model (here, a language model)

Conclusion

This approach to diffusion problems is novel and based on a yet unexplored class of Bayesian priors: the Dirichlet-Point processes. The Dirichlet-Survival process is a special case that merges Dirichlet Processes with the underlying diffusion networks inference literature, of which a significant part is based on Survival analysis (see NetInf, NetRate, InfoPath, KernelCascade, etc.).
By jointly considering time, structure and content of documents, we retrieve relevant topics and a good approximation of their diffusion subnetworks. We make a case that considering these three pieces of information jointly instead of sequentially improve our results. It is to our knowledge the first time this is possible using an online inference algorithm.

[1] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J. Smola, and Le Song. 2015. Dirichlet-Hawkes Processes with Applications to Clustering Continuous-Time Document Streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*
[2] Manuel Gomez-Rodriguez, David Balduzzi, Bernhard Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks. *The 28th International Conference on Machine Learning (ICML),* 2011.
[3] Gaël Poux-Médard, Julien Velcin, Sabine Loudcher. Dirichlet-Survival Process: Scalable Inference of Topic-Dependent Diffusion Networks. *ECIR*, 2023.