Properties of Reddit News Topical Interactions

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

[a] https://gaelpouxmedard.github.io/

Most models of information diffusion rely on the assumption that documents spread independently from each other. Besides, it has been pointed out that interactions between documents must be sparse and brief. We propose a Multivariate Powered Dirichlet-Hawkes process to model such interactions. In this case study, we propose to determine whether they play a significant role in the publication mechanisms of news headlines on Reddit. We consider a corpus of 100,000 news from 2019 and conclude that interactions play a minor role in this dataset.

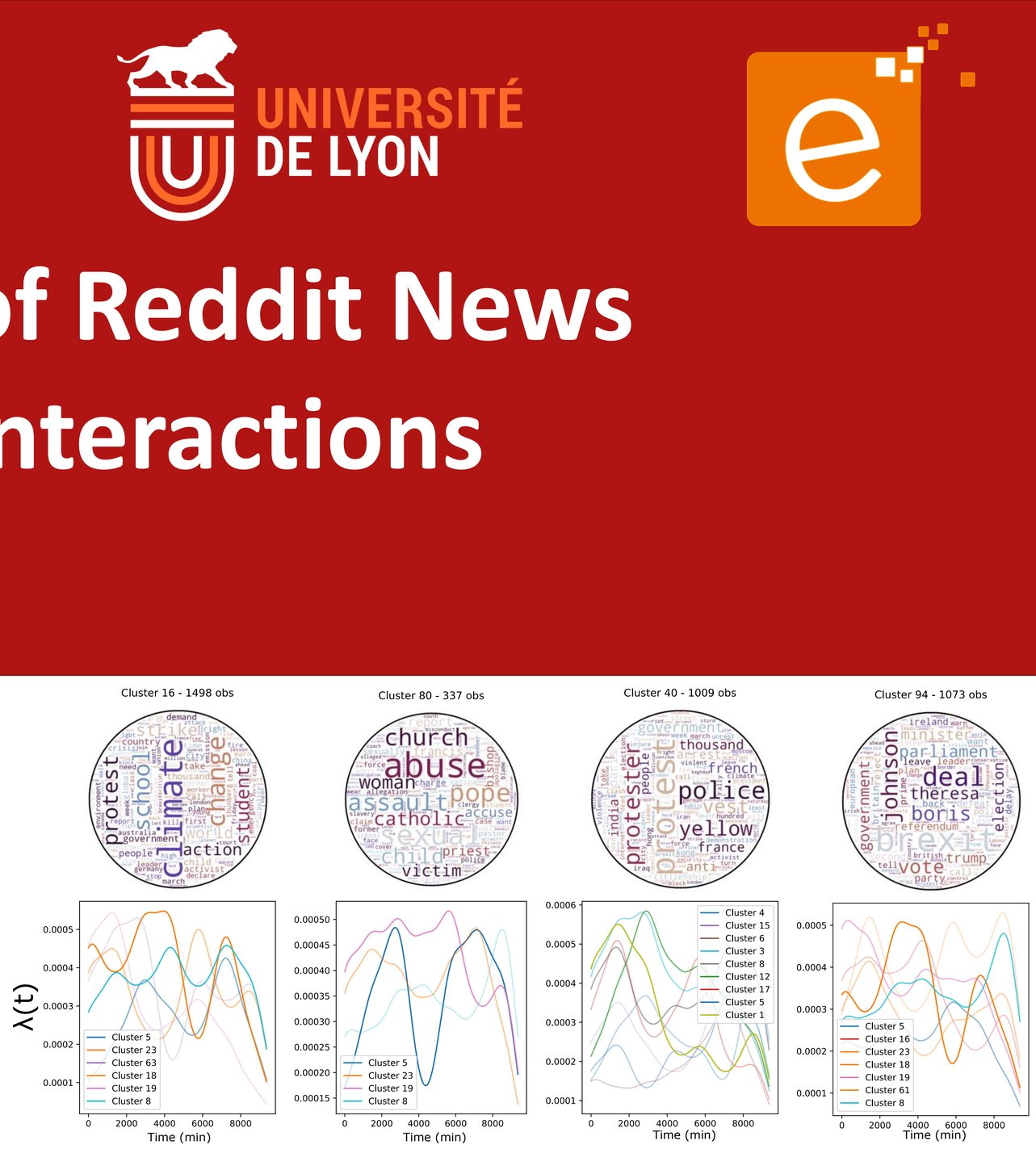

Figure 1 – An output of the Multivariate Powered Dirichlet-Hawkes Process - (Top) A set of inferred topics along with the vocabulary of their documents. (Bottom) Inferred instantaneous probability for one observation from a topic to trigger other observations in other topics at various ulterior times.

A dynamic clustering prior

We consider a **stream of timestamped documents** as input data. The idea is to group them into clusters using both their relative publication dates and their textual content. We want to build a model that sequentially model 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:

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

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. However, in 2015, N. Du *et al.* introduced the Dirichlet-Hawkes process as a way to make the number of observations already within a cluster vanish over time. These temporally weighted counts are expressed by the intensity $\lambda(t)$ of a Hawkes process, that is yet to be inferred. Each cluster is associated to its own intensity function. This is the (Powered[1]) Dirichlet-Hawkes Process[2]. We extend this process to the multivariate case, where the probability for a new document to belong to a cluster depends on the dynamic counts for documents within all clusters. The is the **Multivariate Powered Dirichlet Hawkes Process** (see Fig.2). We couple this prior to a Dirichlet-Multinomial language model (bag of words).

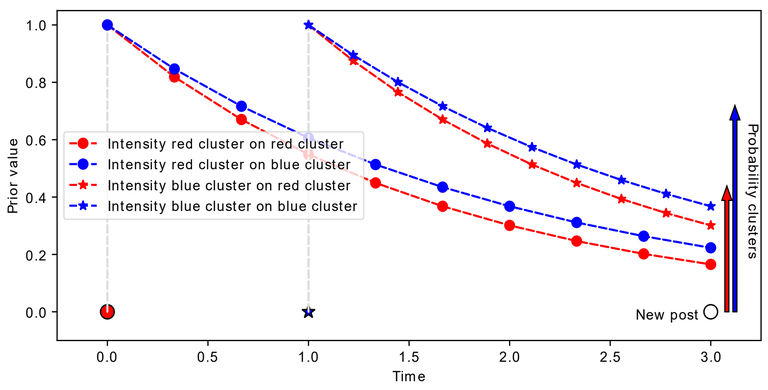

Figure 2 – Multivariate Hawkes intensity associated to two clusters. Arrows represent the prior probability for the new post to belong to either cluster at time t=3.

Reddit News dataset

Reddit News dataset : 102,045 headlines, 13,241 different tokens, 875,000 tokens in total.

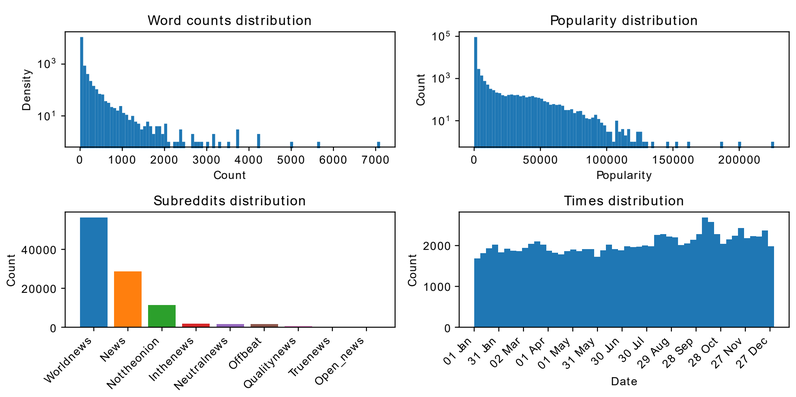

Figure 3 – Characteristics of the Reddit News dataset

Substantially

- Interactions between documents assumed to be rare and brief
- Grouping them into topics helps to spot them
- Multivariate Dirichlet-Hawkes process: see how topics influence each other over time
- Creation of Reddit News corpus (100k headlines)
- Visualize topical interactions as a temporal network
- Significant interactions are rare and brief
- Interactions do not play a significant role in news diffusion within the Reddit News corpus

Topical interactions network

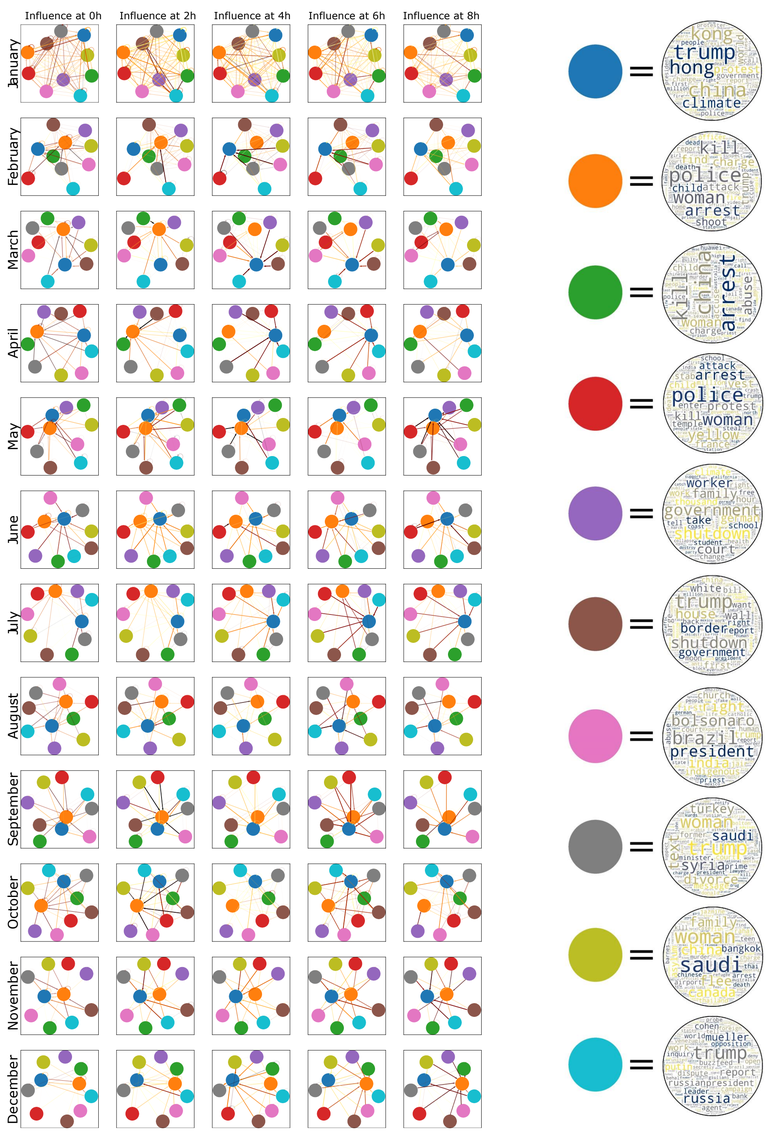

Figure 4 – Topical interaction network. Color: strength of the inferred edge (A); Transparency: confidence in this value, of effective interaction (W).

Quantitative results

In Figure 4, we present a sample of the inferred topical interaction network. The quantities represented are the edges strength (A) and the effective interaction (W). **Effective interaction** is the increase in instantaneous probability for an event to happen due to interactions with other events (see Fig.2).
We ran various experiments considering various timescales (day, hour, minute) and hyperparameters θ_0 (concentration parameter of the language model) and r (importance of the dynamic prior).

Interactions strength

$\vec{\kappa}(t)$	θ_0	r	$< A > (10^{-3})$	$< W > (10^{-5})$	$< A >_W (10^{-3})$	$\frac{<W^{intra}>}{<W^{extra}>}$
Minute	0.01	0.5	49(21)	342(889)	66(17)	1.8(62)
		1.0	48(20)	478(1124)	60(17)	1.4(43)
		1.5	48(20)	746(1901)	60(17)	1.0(33)
	0.001	0.5	50(22)	316(882)	66(17)	3.1(138)
		1.0	50(21)	279(752)	67(16)	2.6(105)
		1.5	50(22)	268(665)	67(16)	2.3(84)
Hour	0.01	0.5	49(18)	389(843)	56(17)	0.5(13)
		1.0	49(18)	478(1187)	56(17)	0.6(15)
		1.5	48(17)	471(789)	52(15)	0.7(13)
	0.001	0.5	50(21)	110(398)	61(17)	1.7(67)
		1.0	50(18)	133(506)	57(17)	1.4(60)
		1.5	49(17)	183(554)	55(17)	1.1(37)
Day	0.01	0.5	49(18)	41(97)	55(17)	1.2(34)
		1.0	49(19)	63(131)	54(17)	1.2(31)
		1.5	49(19)	91(187)	53(18)	1.2(31)
	0.001	0.5	50(20)	18(90)	60(19)	1.1(59)
		1.0	50(19)	23(101)	58(17)	1.0(50)
		1.5	50(19)	37(111)	56(18)	1.0(36)

Table 1 – Interaction strength - Overall, interactions between clusters are weak. The large standard deviations suggest that there is a variety of interacting behaviours. Interactions tend to happen within a cluster (self-interactions).

Interactions range

$\vec{\kappa}(t)$	θ_0	r	κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	κ_7	κ_8	κ_9
			0m	10m	20m	30m	40m	50m	60m	70m	80m
Minute (m)	0.01	0.5	133	421	407	451	428	403	395	345	121
		1	198	591	580	607	532	575	521	507	224
		1.5	308	937	893	914	955	840	808	810	304
	0.001	0.5	218	509	457	424	371	340	313	178	52
		1	142	435	396	388	343	327	272	187	45
		1.5	104	388	366	353	326	333	290	215	61
			0h	2h	4h	6h	8h				
Hour (h)	0.01	0.5	247	430	502	456	324				
		1	329	538	549	542	451				
		1.5	229	615	532	526	411				
	0.001	0.5	62	149	119	137	92				
		1	77	164	172	149	111				
		1.5	104	244	223	197	156				
			0d	1d	2d	3d	4d	5d	6d		
Day (d)	0.01	0.5	22	45	47	46	47	47	37		
		1	35	71	72	68	68	70	61		
		1.5	51	100	101	105	98	105	82		
	0.001	0.5	9	20	21	21	21	21	17		
		1	12	26	24	27	28	26	22		
		1.5	11	41	42	41	41	41	35		

Table 2 – Effective interaction range - All the values are given in ten-thousandth (10^{-5}). There is a decreasing trend over time for all kernels.

Conclusion

We recover previous conclusions on interactions between Reddit documents: they are rare and do not last in time. We find that interactions do not increase significantly the probability of a document getting published in the Reddit News dataset.

[1] Gaël Poux-Médard, Julien Velcin and Sabine Loudcher. 2021. Powered Hawkes-Dirichlet Process: Challenging Textual Clustering using a Flexible Temporal Prior. In *Proceedings of IEEE International Conference on Data Mining, ICDM'21*
[2] 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*