Interactions entre informations dans les processus de diffusion

Sous la direction de J. Velcin et S. Loudcher

Gaël Poux-Médard

Université de Lyon, France
Lyon 2, ERIC UR 3083

13 septembre 2022

Introduction

- Grandes quantités de données :
  - 1M+ de posts Reddit/jour, 500M de tweets/jour, etc.
- Comprendre les mécanismes de diffusion est fondamental
  - Limiter la diffusion de fake news, de virus, etc.
  - Promouvoir des contenus sur la santé, l'énergie, etc.
  - Compréhension de nos biais humains : influence, polarisation, etc.
  - Applications directes : recommandation, publicité, résumés, etc.

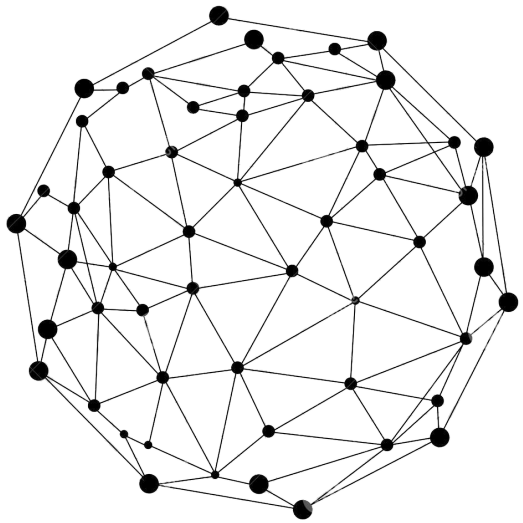


Figure 1 – Les informations en ligne se diffusent sur un réseau

Flux de documents

- Modéliser interactions entre documents affine cette compréhension

→ Influence des précédents documents sur les suivants ?

- Réaction tweets climatiques $\overset{?}{\propto}$ Articles sur des feux de forêt
- Réaction tweets politiques $\overset{?}{\propto}$ Articles sur l'achat d'un chiot

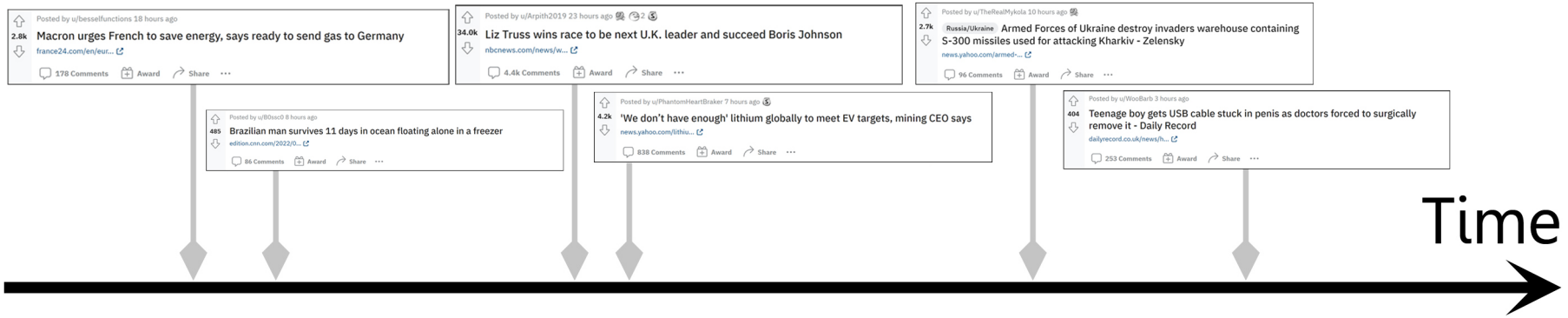


Figure 2 – Un flux de documents

Définitions

- **Information** : entité susceptible d'être diffusée (tweet, news, ...)
- **Action** : réaction à des informations (retweet, publication, ...)
- **Interaction** : lorsque l'effet joint de plusieurs informations est différent de la somme de leurs effets individuels.

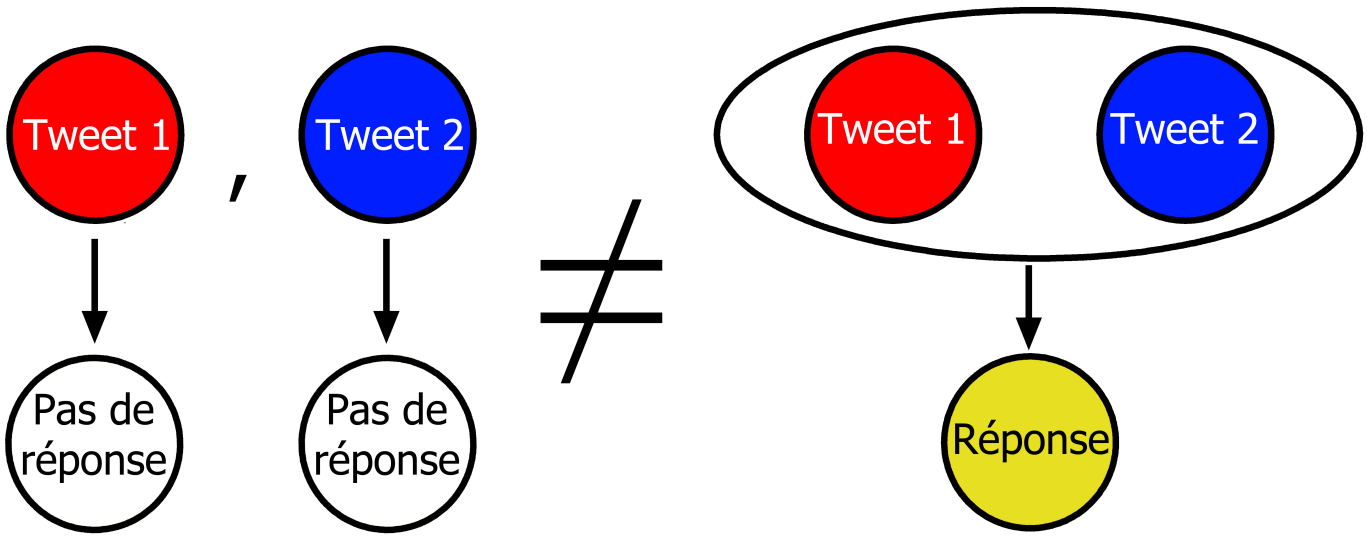


Figure 3 – Exemple schématique d'une interaction

Travaux similaires

- La plupart des modèles existants ne considèrent pas les interactions
- Plusieurs travaux récents proposant des processus de diffusion avec interactions (Wang et al., 2019; Zhu et al., 2020)
  - Définition de règles microscopiques, simulation, puis comparaison

  $\rightarrow$ Pas d'apprentissage des données
- Peu de travaux d'apprentissage automatique sur le sujet
  - Clash of the Contagions (Myers and Leskovec, 2012)
  - Correlated cascades (Zarezade et al., 2017)

Problématiques

1. Les interactions sont-t-elles fréquentes ?
2. Les interactions perdurent-t-elles dans le temps ?
3. Comment modéliser les interactions à grande échelle ?
4. Quel est le rôle des interactions dans les processus de diffusion ?

Contributions de la thèse

Table 1 – Contributions détaillées dans le manuscrit de thèse

	SIMSBM	IMMSBM	SDSBM	InterRate	PDP	PDHP	MPDHP
Publication	ICDM'22	RecSys'21	-	ECML-PKDD'21	-	ICDM'21	CNA'22
Auto-interactions	x	x	x	x	x	x	x
Interactions paires	x	x	x	x		x	x
Interactions n-plet	x		x			x	x
Clustering	x	x	x		x	x	x
Temps discrets			x	x		x	x
Temps continus				x		x	x
Inférence séquentielle					x	x	x

Stochastic Block Models

Les interactions sont-elles fréquentes ?

- Modèles :
  - IMMSBM (Poux-Médard *et al.*, RecSys 2021)
  - SIMSBM (Poux-Médard *et al.*, ICDM 2022)
  - SDSBM

Stochastic Block Models

- Réduction de dimensionnalité via clustering
  - Noeuds : documents (ex. tweets)
  - Liens : résultat d'une relation entre documents (ex. retweet ou non)

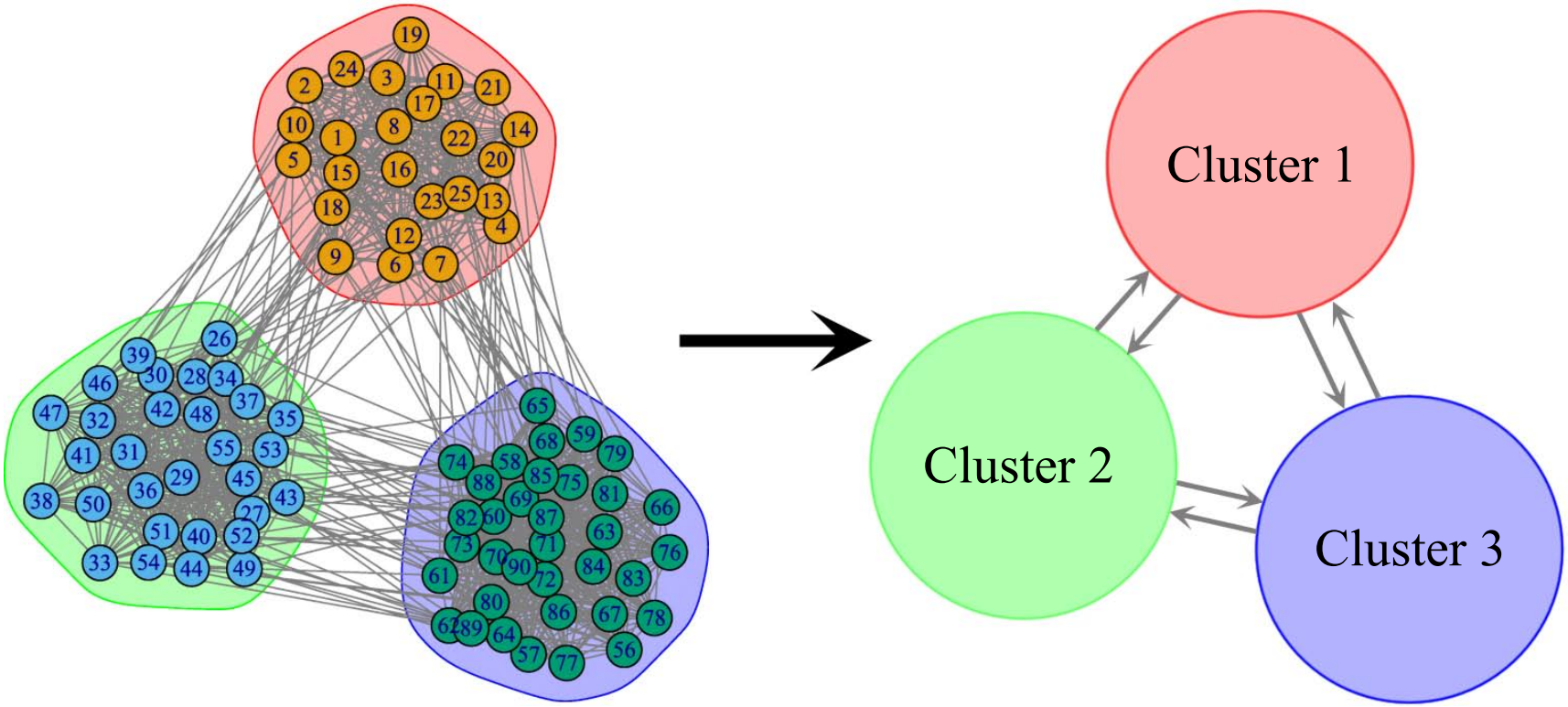


Figure 4 – SBM - Supposer l'existence de groupes simplifie le système

IMMSBM

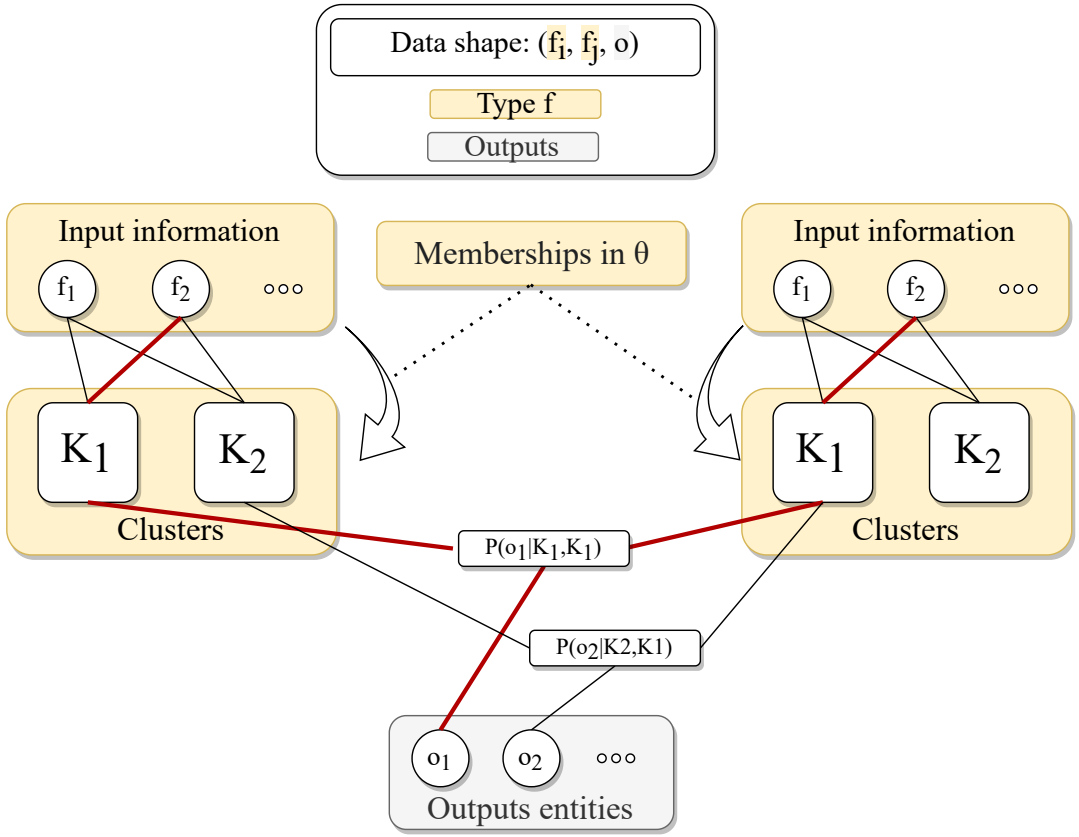


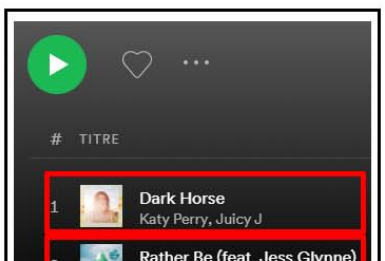
Figure 5 – Modèle IMMSBM proposé - (Poux-Médard et al., 2021a)

- Inférence via algorithme EM

Tâche

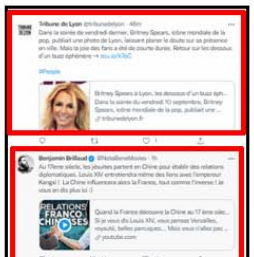
- Quatre jeux de données et interactions de paires :
  - Spotify : (chanson, chanson) → chanson suivante
  - Reddit : (mot, mot) → mot réponse
  - Twitter : (tweet, tweet) → retweet
  - PubMed : (symptôme, symptôme) → maladie

Songs ▶ Next song?



Message ▶ Words in answer?

Tweets ▶ Retweet?



Symptoms ▶ Diseases?

Résultats

- Prédiction des relations non-observées ; scores P@10, F1, AUC-ROC

		P@10	Max-F1	AUC (ROC)
PubMed	Naive	0.212	0.160	0.863
	MMSBM	0.627 ± 0.002	0.393 ± 0.002	**0.911 ± 0.000**
	IMMSBM	**0.656 ± 0.001**	**0.411 ± 0.001**	**0.911 ± 0.002**
	Up.lim.	0.668	0.450	0.936
Twitter	Naive	0.462	0.147	0.554
	MMSBM	0.529 ± 0.005	0.254 ± 0.005	0.741 ± 0.004
	IMMSBM	**0.610 ± 0.004**	**0.349 ± 0.006**	**0.800 ± 0.001**
	Up.lim.	0.737	0.748	0.959
Reddit	Naive	0.488	0.164	0.660
	MMSBM	0.495 ± 0.000	0.177 ± 0.000	0.686 ± 0.000
	IMMSBM	**0.499 ± 0.000**	**0.181 ± 0.000**	**0.687 ± 0.000**
	Up.lim.	0.558	0.582	0.933
Spotify	Naive	0.355	0.088	0.573
	MMSBM	0.426 ± 0.006	0.167 ± 0.003	0.707 ± 0.002
	IMMSBM	**0.502 ± 0.006**	**0.228 ± 0.005**	**0.723 ± 0.002**
	Up.lim.	0.570	0.607	0.944

Figure 6 – Considérer les interactions améliore nos résultats

Résultats

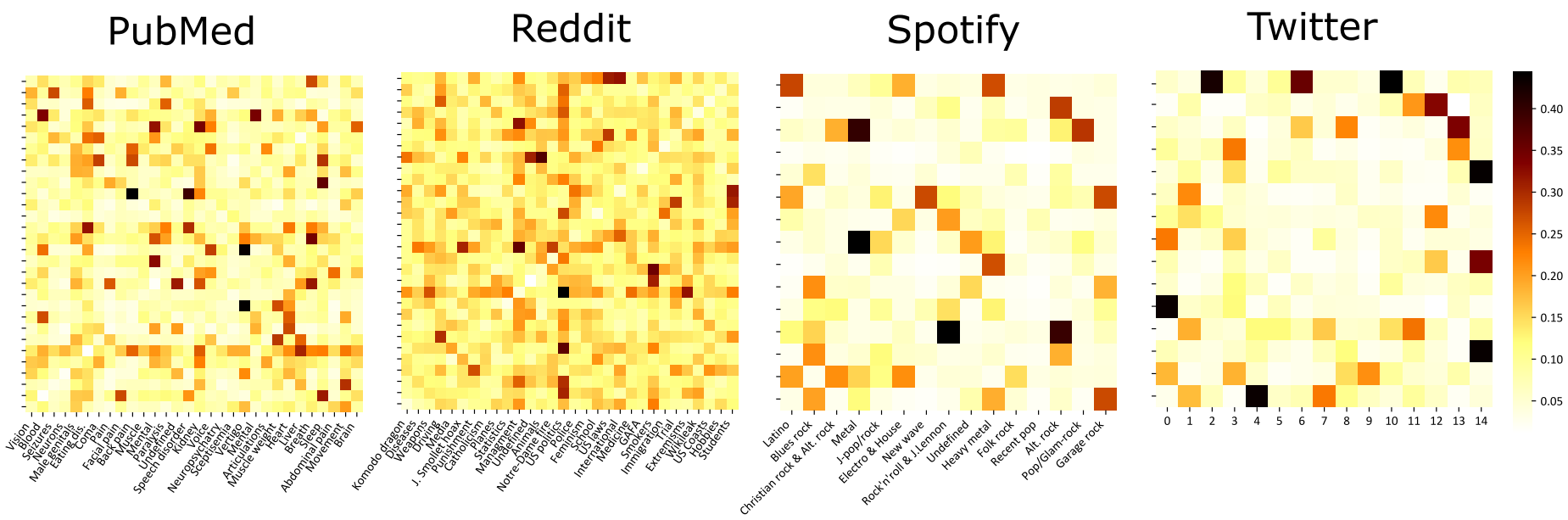


Figure 7 – **Influence des interactions** – Peu d'interactions modifient grandement la probabilité d'une réaction.

La majorité des interactions ont peu d'effet

- La plupart des interactions n'a que peu d'impact sur le système

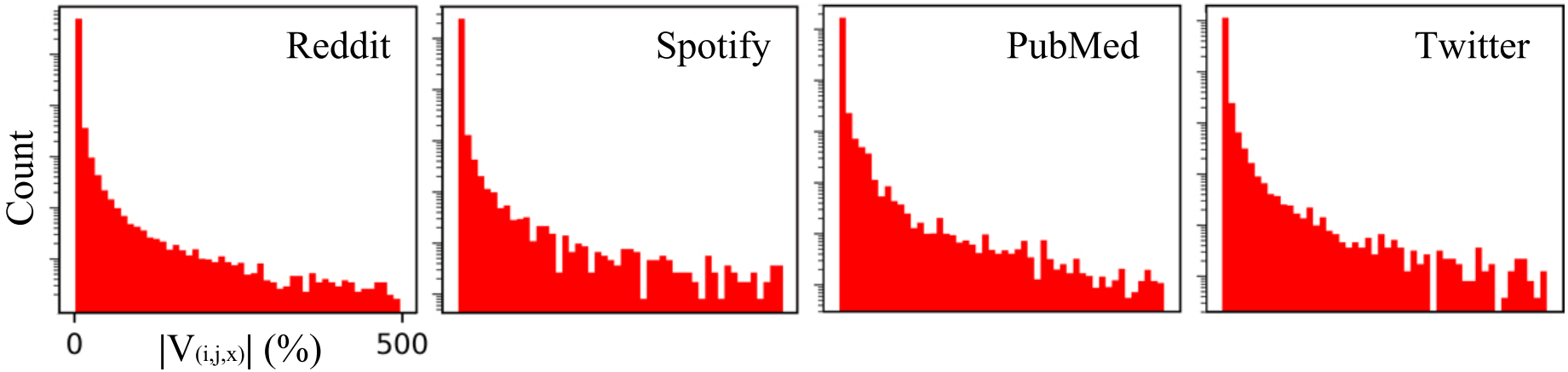


Figure 8 – Histogramme de la variation relative des probabilités d'un événement imputable aux interactions.

Les interactions majeures sont rares

Un problème de temps ?

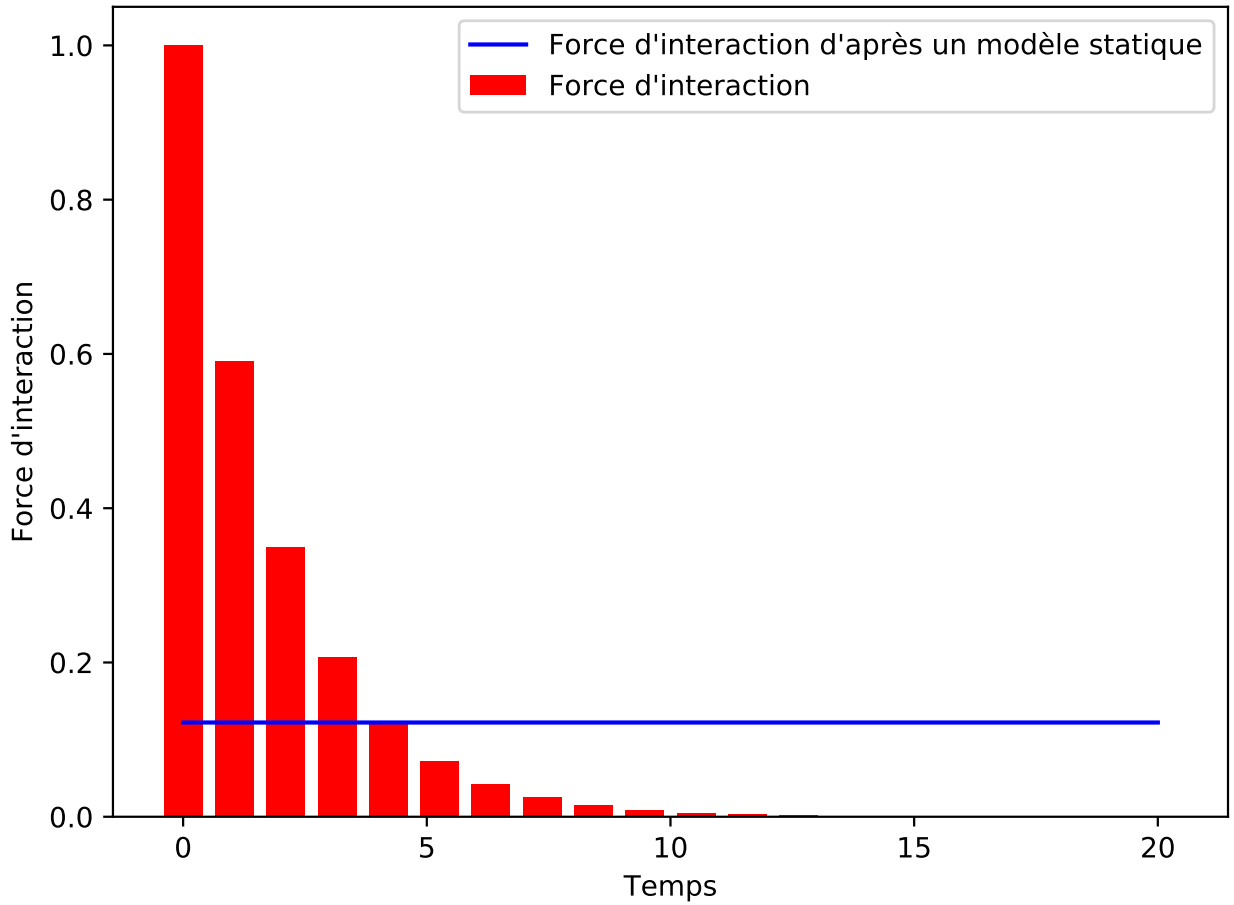


Figure 9 – Les modèles statiques peuvent omettre des interactions.

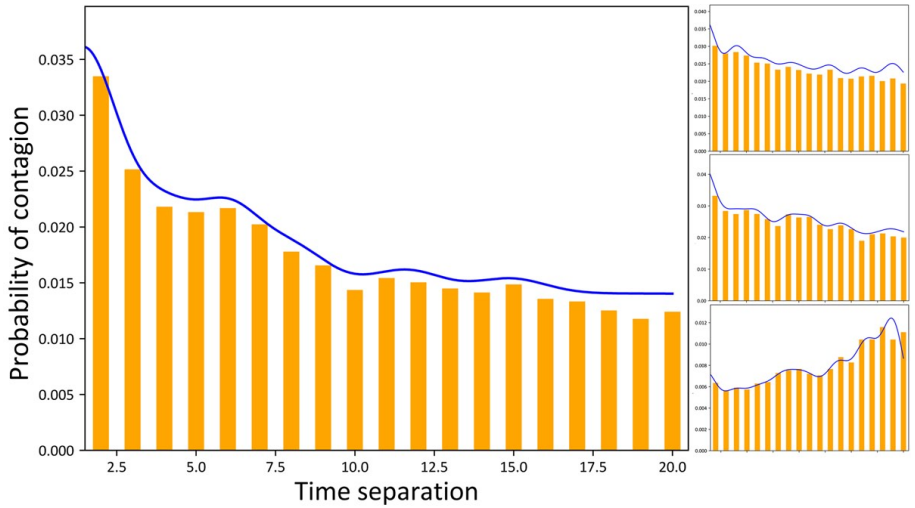
Modèle dynamique

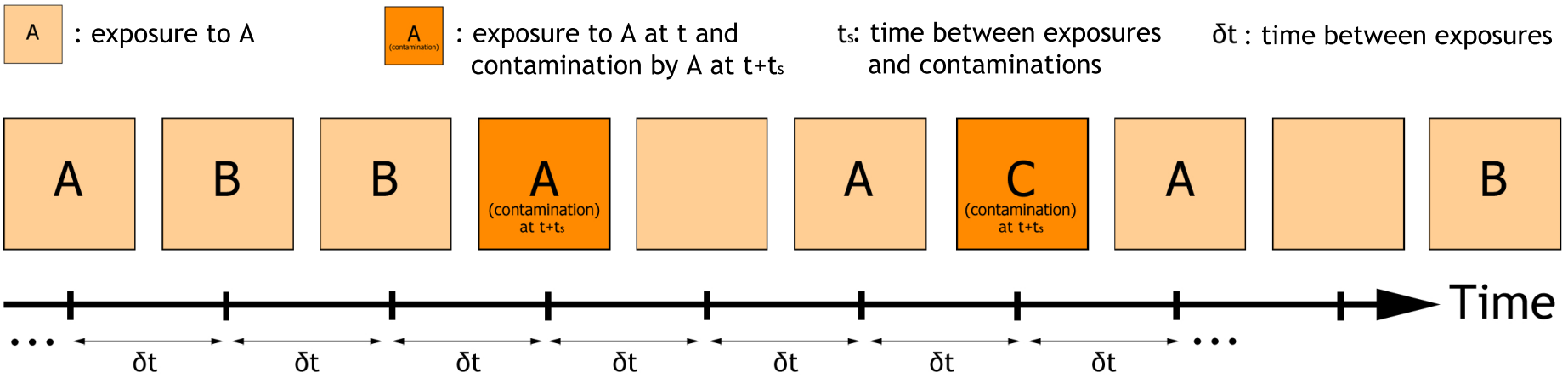
Les interactions évoluent-t-elles dans le temps ?

- Modèle :
  - InterRate (Poux-Médard *et al.*, ECML-PKDD 2021)

InterRate

- Les interactions de paires persistent-t-elles dans le temps ?
  - Information A au temps t_A et B au temps $t_B > t_A$: comment A influe B après un temps $\Delta t = t_B - t_A$?
- Modèle convexe





Résultats (quantitatifs)

- Prédire la probabilité d'une réaction dans le temps
  - Scores : erreur quadratique résiduelle, divergence de JS, best case F1

		RSS	JS div.	BCF1
Twitter	IR-RBF	0.0015	0.000 06	0.983
	IR-EXP	0.0011	0.000 05	0.986
	ICIR	0.0137	0.000 63	0.961
	Naive	0.0161	0.000 73	0.938
	CoC	0.0017	0.000 07	0.957
	IMMSBM	0.0147	0.000 68	0.954
PD	IR-RBF	1.1268	0.007 58	0.979
	IR-EXP	1.5526	0.008 67	0.966
	ICIR	3.5359	0.018 23	0.938
	Naive	3.6527	0.019 15	0.945
	CoC	1.2409	0.008 09	0.974
	IMMSBM	20.3773	0.087 01	0.767
Ads	IR-RBF	0.0043	0.000 04	0.981
	IR-EXP	0.0030	0.000 03	0.985
	ICIR	0.0983	0.000 85	0.966
	Naive	0.1453	0.001 26	0.913
	CoC	0.0045	0.000 05	0.974
	IMMSBM	0.0155	0.000 15	0.954

Figure 10 – Considérer les interactions temporelles améliore nos résultats (Poux-Médard *et al.*, ECML-PKDD 2021)

Résultats (qualitatifs)

- Trois jeux de données :
  - Twitter : anciens tweets influencent nouveaux retweets ?
  - Ads : exposition à des pubs influencent les probabilités de click ?
  - Dilemme du prisonnier : coopérations/trahisons passées influencent le comportement présent ?

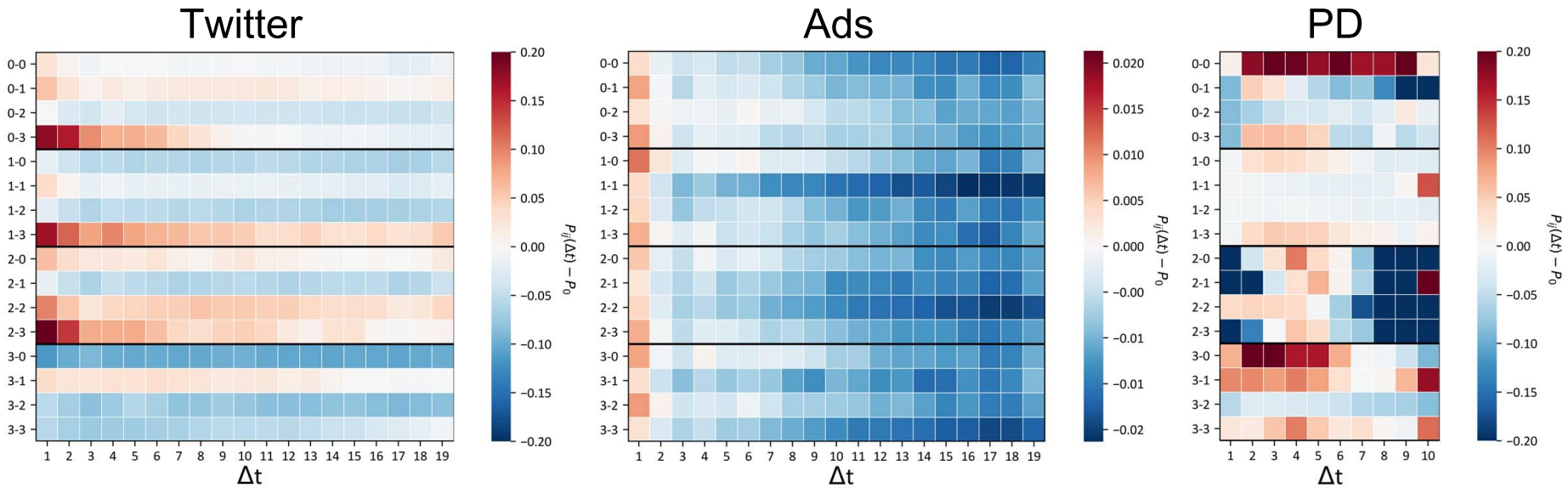
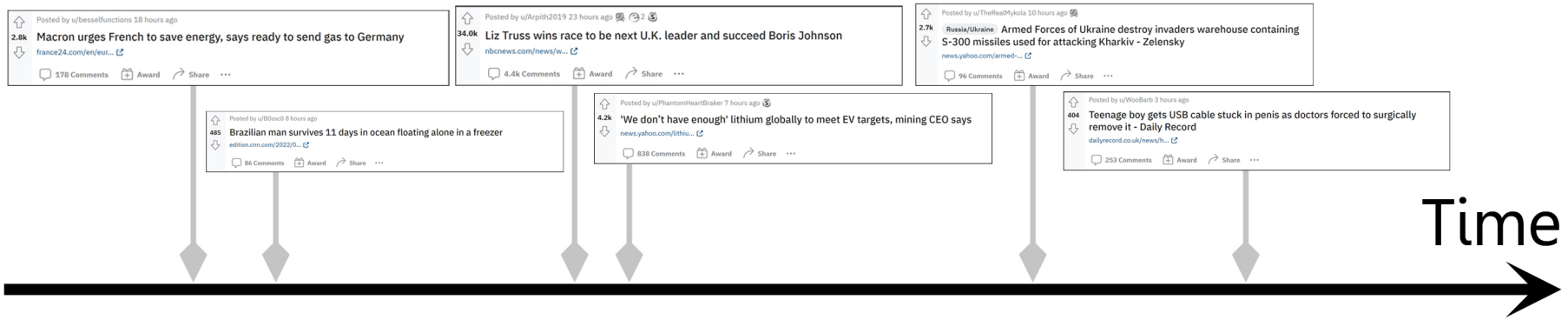


Figure 11 – Conclusion : les interactions entre informations ont un effet bref (Poux-Médard *et al.*, ECML-PKDD 2021)

Synthèse

- Les interactions sont peu fréquentes
  - → Le clustering est nécessaire
- Les interactions sont brèves
  - → Modéliser le temps est nécessaire
- Les interactions améliorent nos résultats
- Solution : modéliser conjointement des paires de clusters et leur interaction dynamique
  - Direction choisie : Dirichlet-Hawkes Processes (Du *et al.*, KDD 2015)
  - Petits plus : modèles séquentiels & modèles de langue



Dirichlet-Hawkes Processes

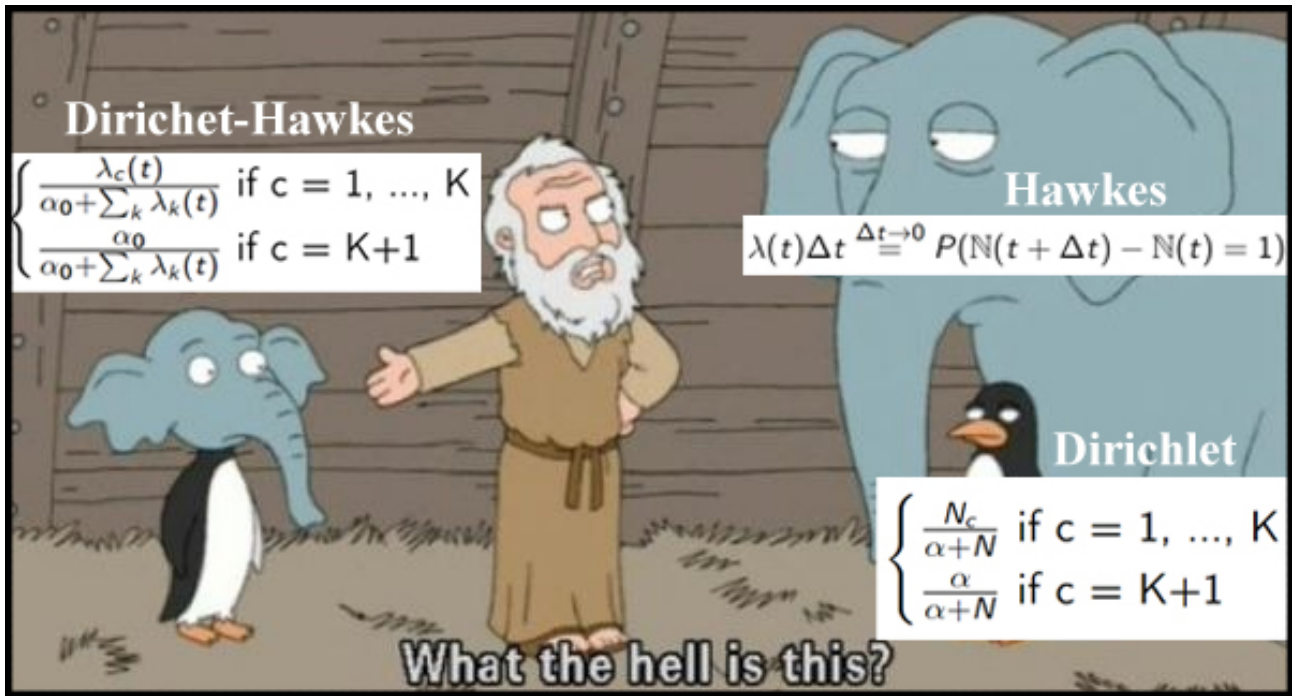
Modéliser des interactions brèves et rares ?

- Modèles :
  - Powered Dirichlet Process
  - Powered Dirichlet-Hawkes Process (Poux-Médard *et al.*, ICDM 2021)
  - Multivariate Powered DHP (Poux-Médard *et al.*, CNA 2022)

Processus de Dirichlet-Hawkes

- (Du et al., 2015) : Prior de Dirichlet-Hawkes (Bayesian inference)

$$P(\text{cluster}|\text{texte, temps, } \mathcal{H}) \propto \underbrace{P(\text{texte}|\text{cluster})}_{\text{Vraisemblance textuelle (Dirichlet-Multinomiale)}} \times \underbrace{P(\text{cluster}|\text{temps, } \mathcal{H})}_{\text{Prior temporel (Dirichlet-Hawkes)}}$$



Processus de Dirichlet - Modèles séquentiels

- Processus de Dirichlet :

$$DP(C_i = c | C_1, C_2, ..., C_{i-1}, \alpha) = \begin{cases} \frac{N_c}{\alpha+N} & \text{if c = 1, ..., K} \\ \frac{\alpha}{\alpha+N} & \text{if c = K+1} \end{cases}$$

- Un a priori utile pour modéliser des données séquentielles

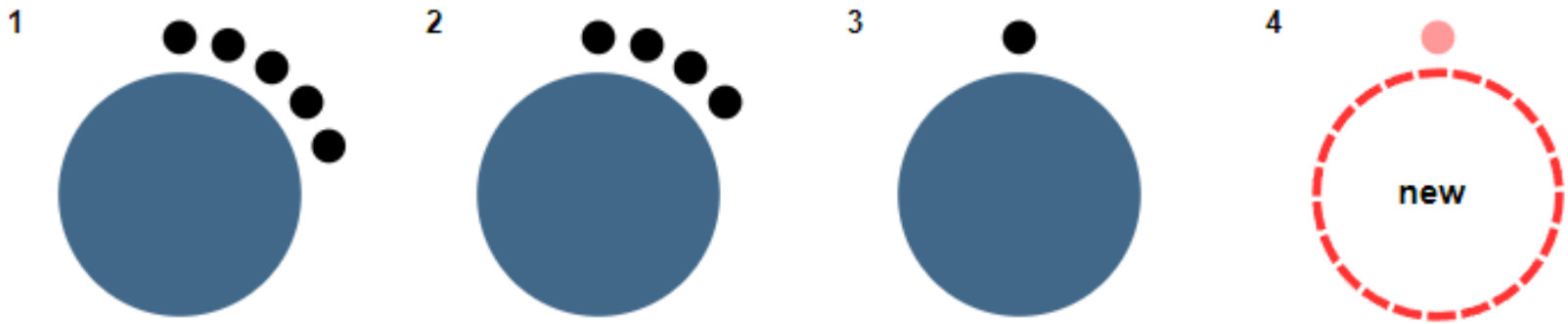


Figure 12 – Réalisation d'un processus de Dirichlet (10 étapes)

Processus de Dirichlet - Un choix arbitraire

- Le Proc.Dir. a une propriété "les riches s'enrichissent" linéaire
  - Pourquoi cette linéarité ? (Lee and Sang, 2022)
  - Pourquoi même cette hypothèse ? (Wallach et al., 2010)
- Le processus de Dirichlet est un choix de prior arbitraire.
  - D'autres priors sont possibles (Welling, 2006; Lee and Sang, 2022)
  - Le choix de l'a priori importe (Wallach et al., 2009)
  - Peu de variations (Wallach et al., 2010; Pitman and Yor, 1997)

Processus de Dirichlet - Powered Dirichlet Process

- Proposition : Powered Dirichlet Process :

$$PDP(C_i = c | C_1, ..., C_{i-1}, \alpha, r) = \begin{cases} \frac{N_c^r}{\alpha + \sum_k N_k^r} & \text{if c = 1, ..., K} \\ \frac{\alpha}{\alpha + \sum_k N_k^r} & \text{if c = K+1} \end{cases}$$

  - Généralisation de plusieurs modèles :
    - $r < 0$: "les riches s'appauvrissent"
    - $r = 0$: "les riches ne s'enrichissent pas" (Uniform Process)
    - $0 < r < 1$: "les riches s'enrichissent moins"
    - $r = 1$: "les riches s'enrichissent" (Dirichlet Process)
    - $r = \frac{\log(N_k - \beta)}{\log(N_k)}$: "les riches s'enrichissent" (Pitman-Yor Process)
    - $r > 1$: "les riches s'enrichissent plus"

Processus de Hawkes

- $\lambda(t) \propto$ probabilité instantanée d'un nouvel événement
- Processus de Hawkes : $\lambda(t)$ dépend des événements passés $\mathcal{H}_t$
  $\rightarrow$ "Processus auto-stimulé"
- Typiquement : $\lambda(t|\mathcal{H}_t) = \lambda_0 + \sum_{t_i \in \mathcal{H}_t} \underbrace{\phi(t - t_i)}_{\text{Kernel function(s)}}$

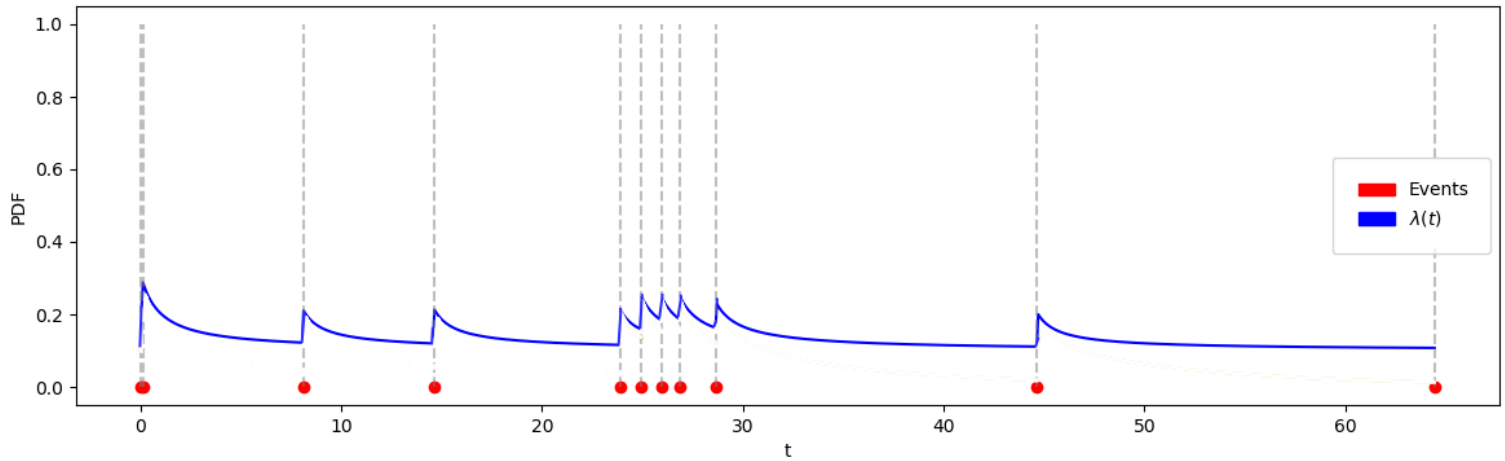


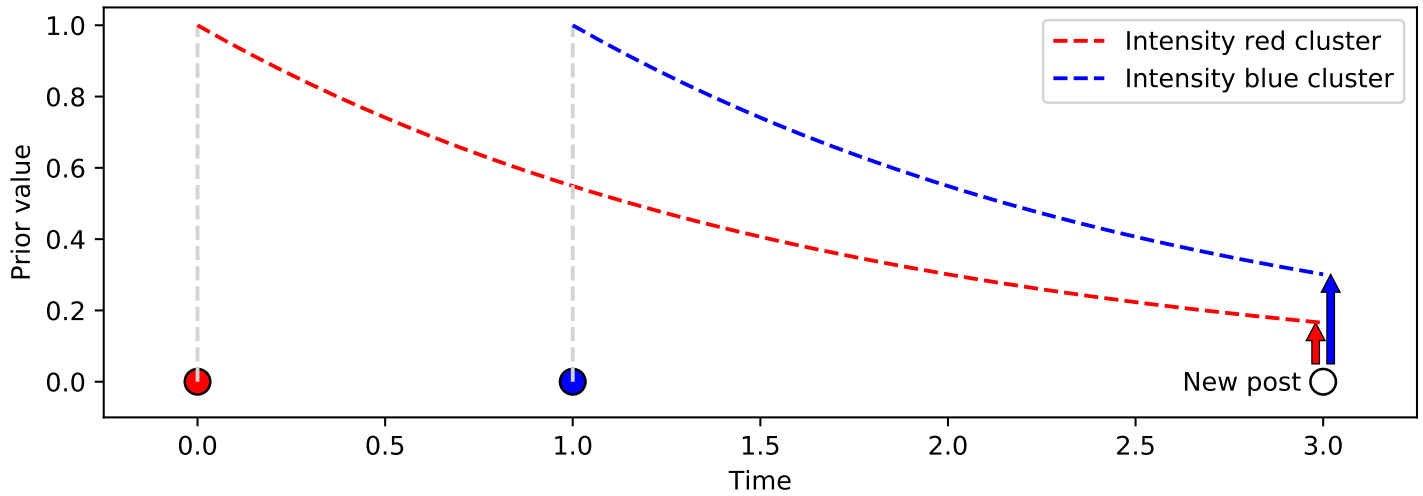
Figure 13 – Pourrait modéliser des dynamiques de publication en ligne

Processus de Dirichlet-Hawkes - Mathématiquement

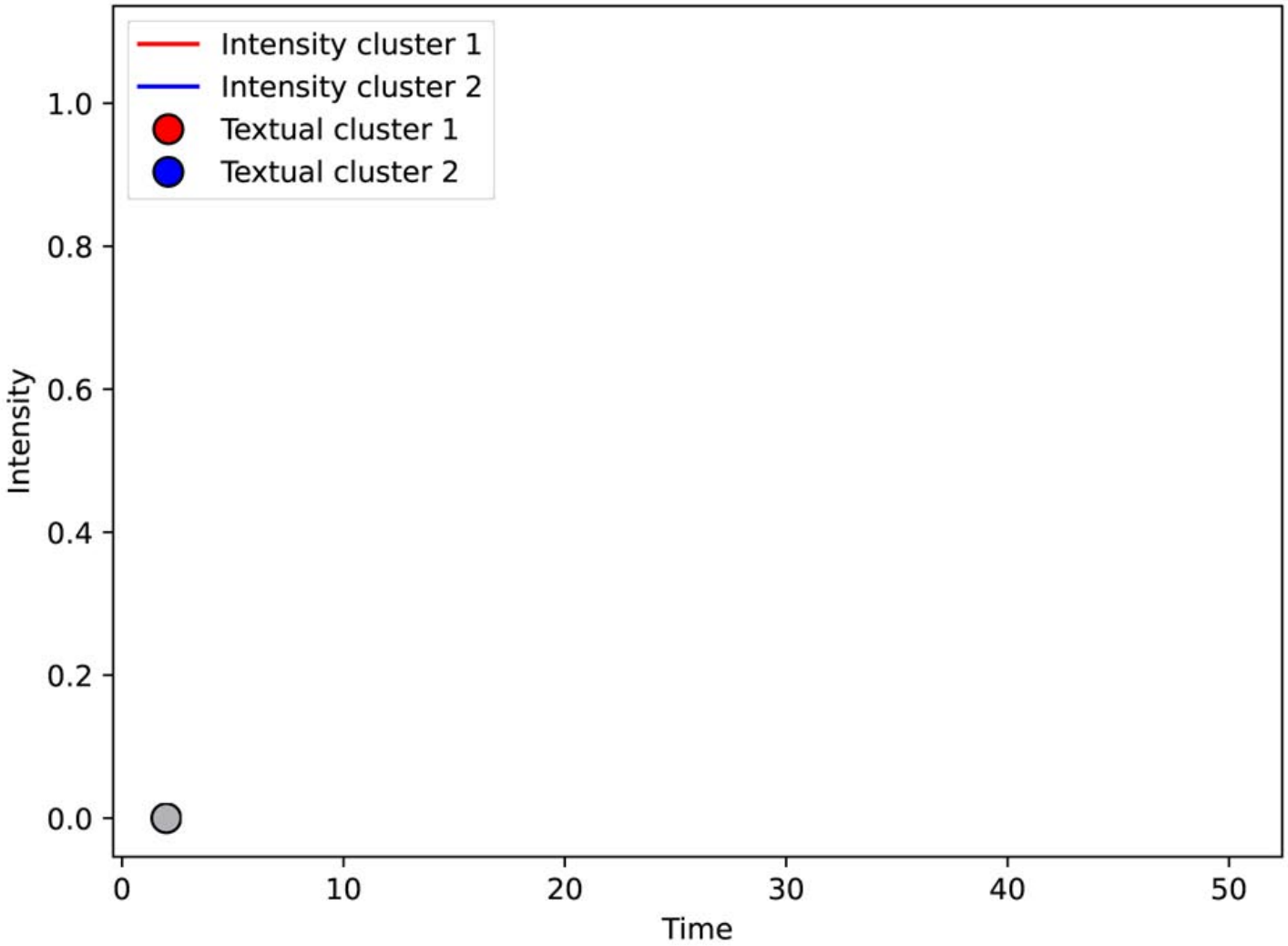
- Dirichlet-Hawkes Process (Du et al., 2015)

→ Comptes N_c de Dirichlet remplacés par intensités $\lambda_c(t)$:

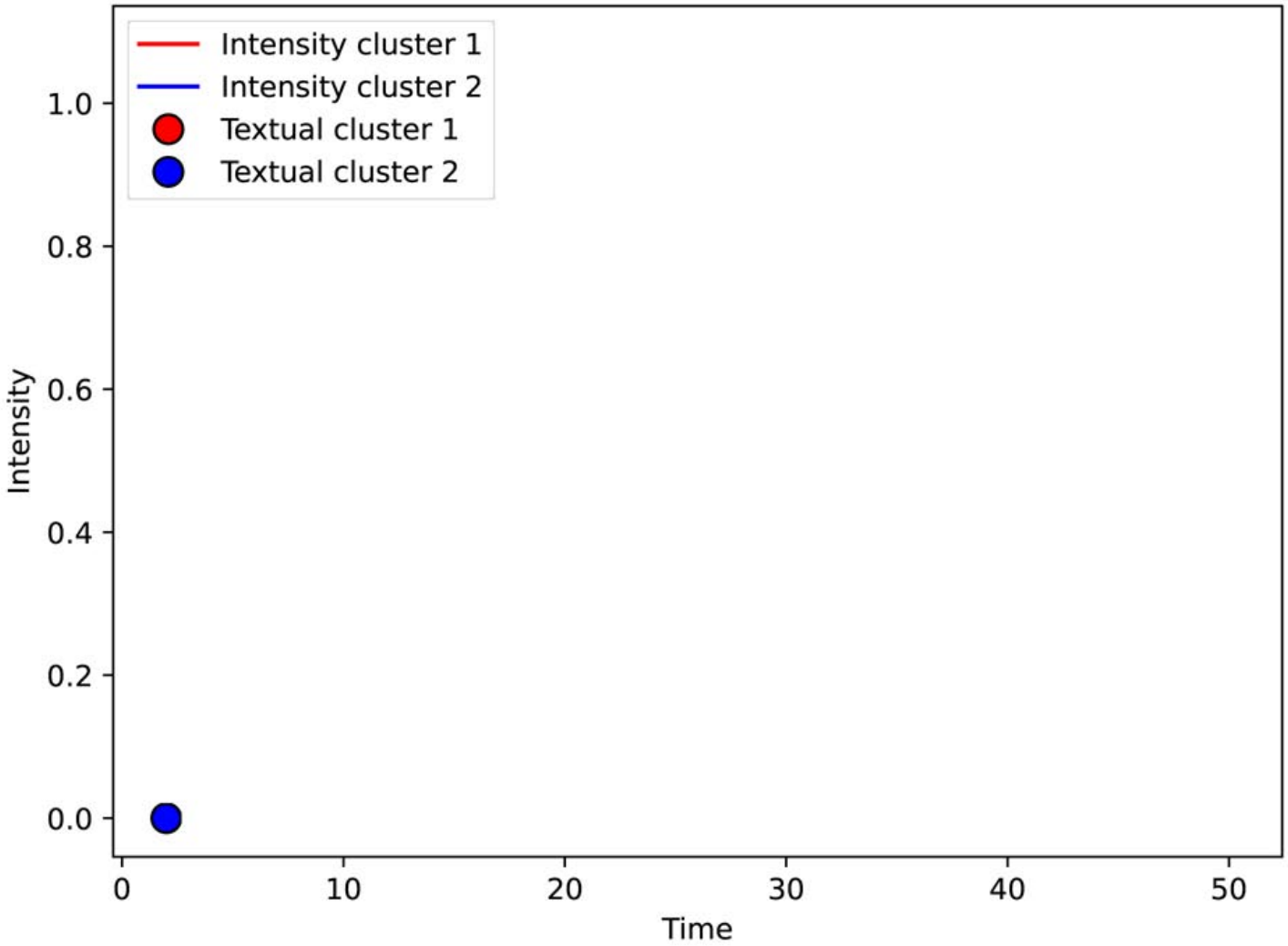
$$\underbrace{P(c|t,\mathcal{H})}_{\substack{\text{Prior temporel} \\ \text{(Dirichlet-Hawkes)} \\ \downarrow}} = \begin{cases} \frac{\lambda_c(t)}{\alpha_0+\sum_k \lambda_k(t)} & \text{if c = 1, ..., K} \\ \frac{\alpha_0}{\alpha_0+\sum_k \lambda_k(t)} & \text{if c = K+1} \end{cases}$$



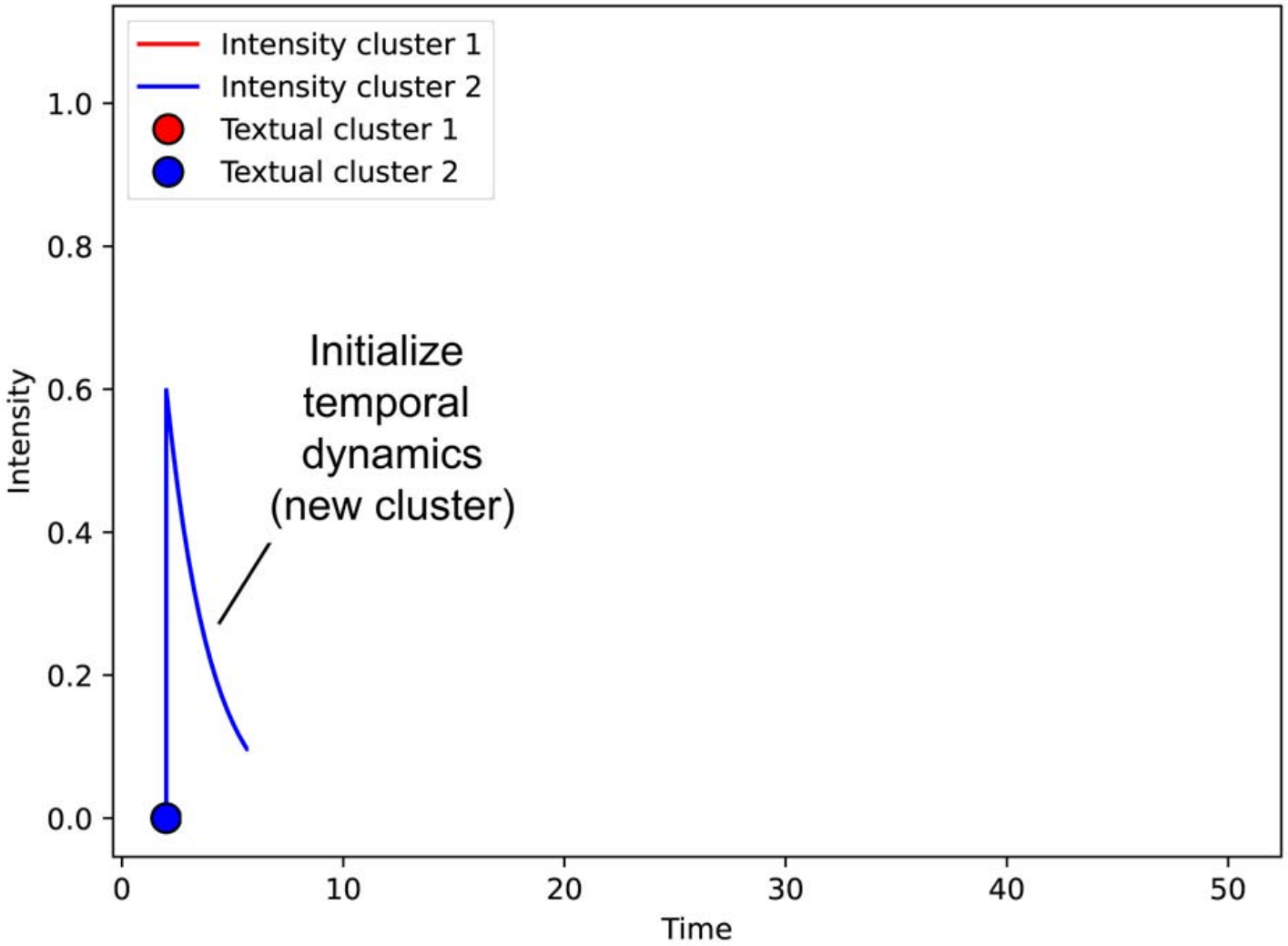
Processus de Dirichlet-Hawkes - Inférence (1 instance)



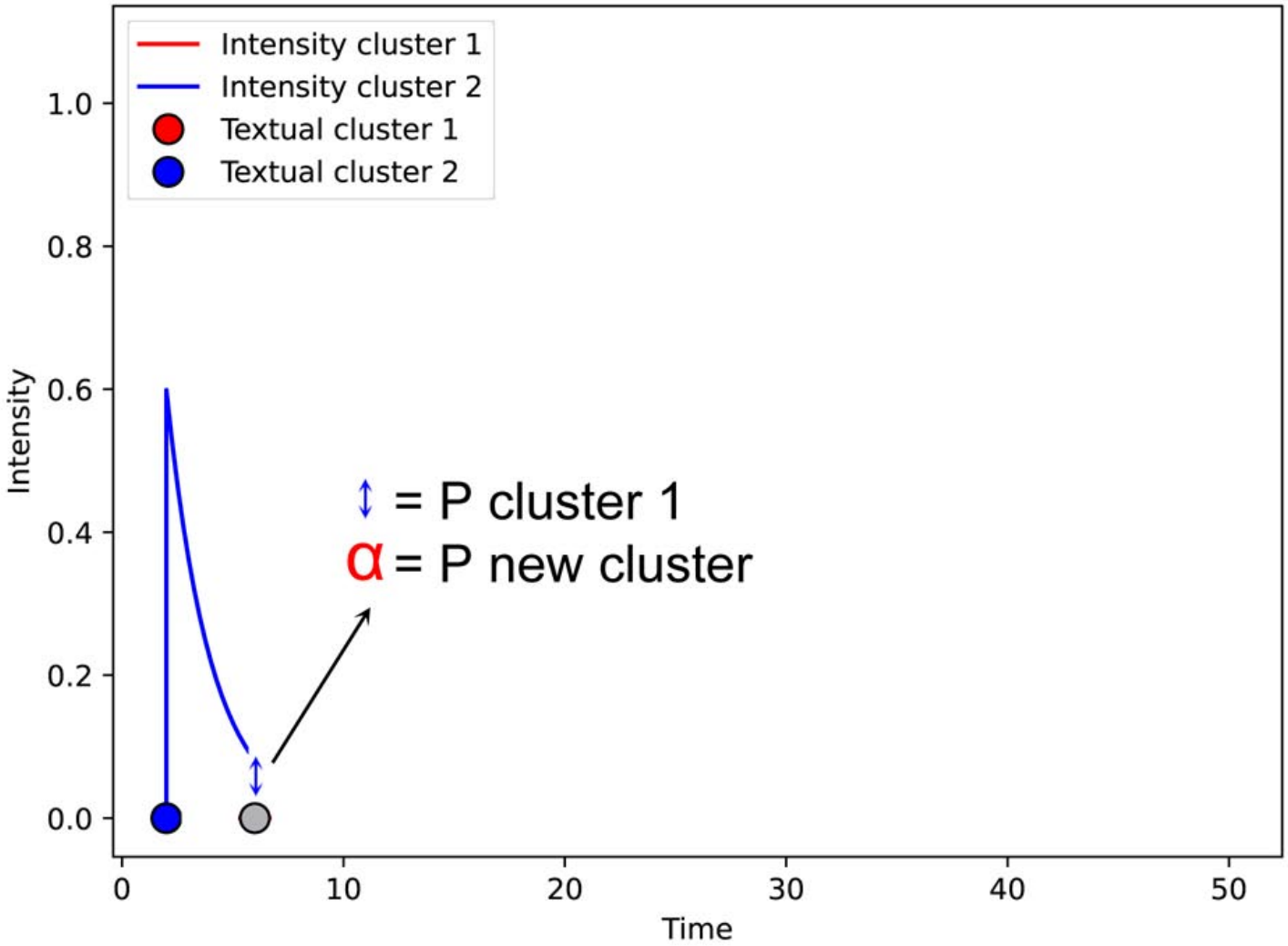
Processus de Dirichlet-Hawkes - Inférence (1 instance)



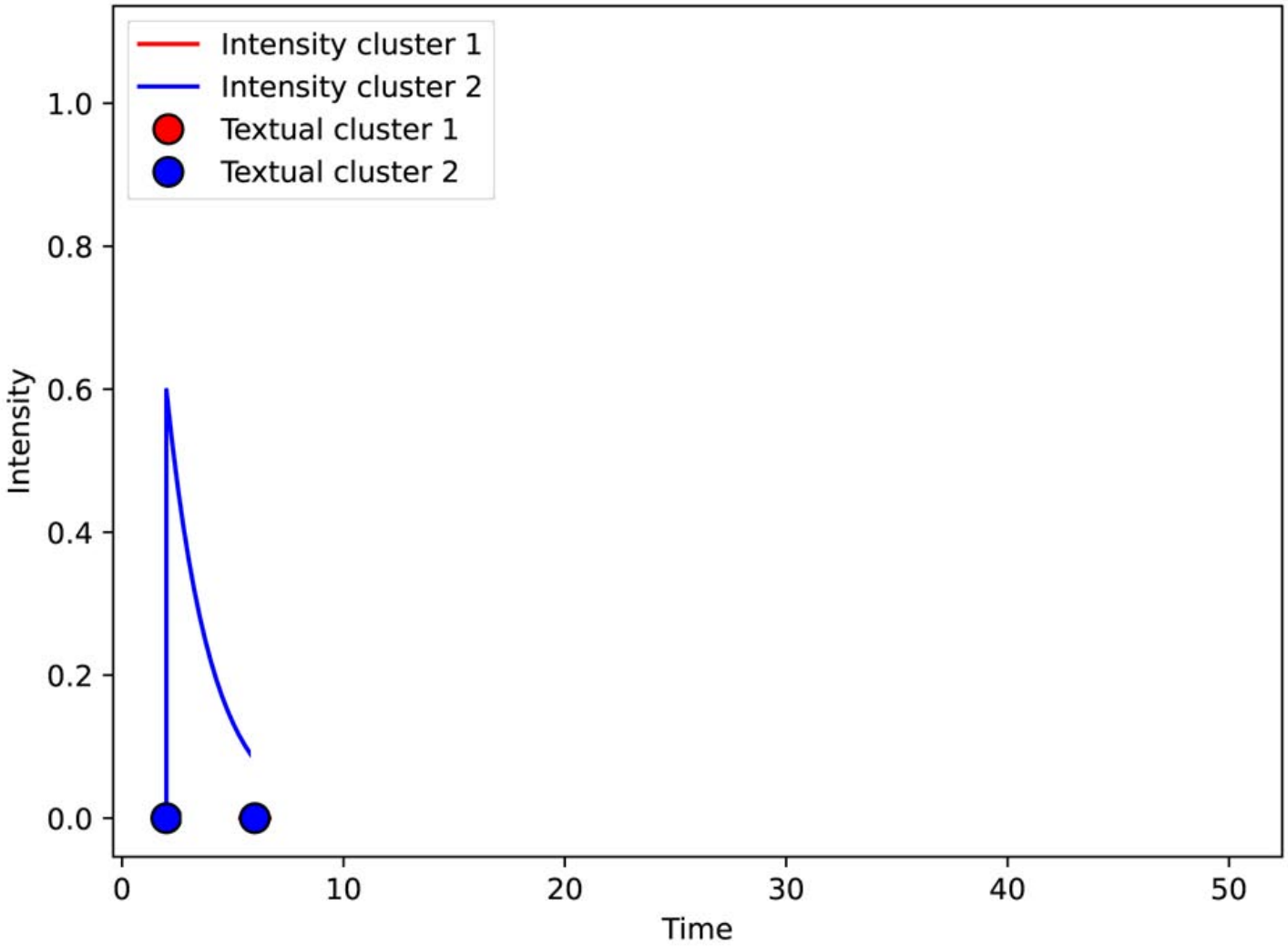
Processus de Dirichlet-Hawkes - Inférence (1 instance)



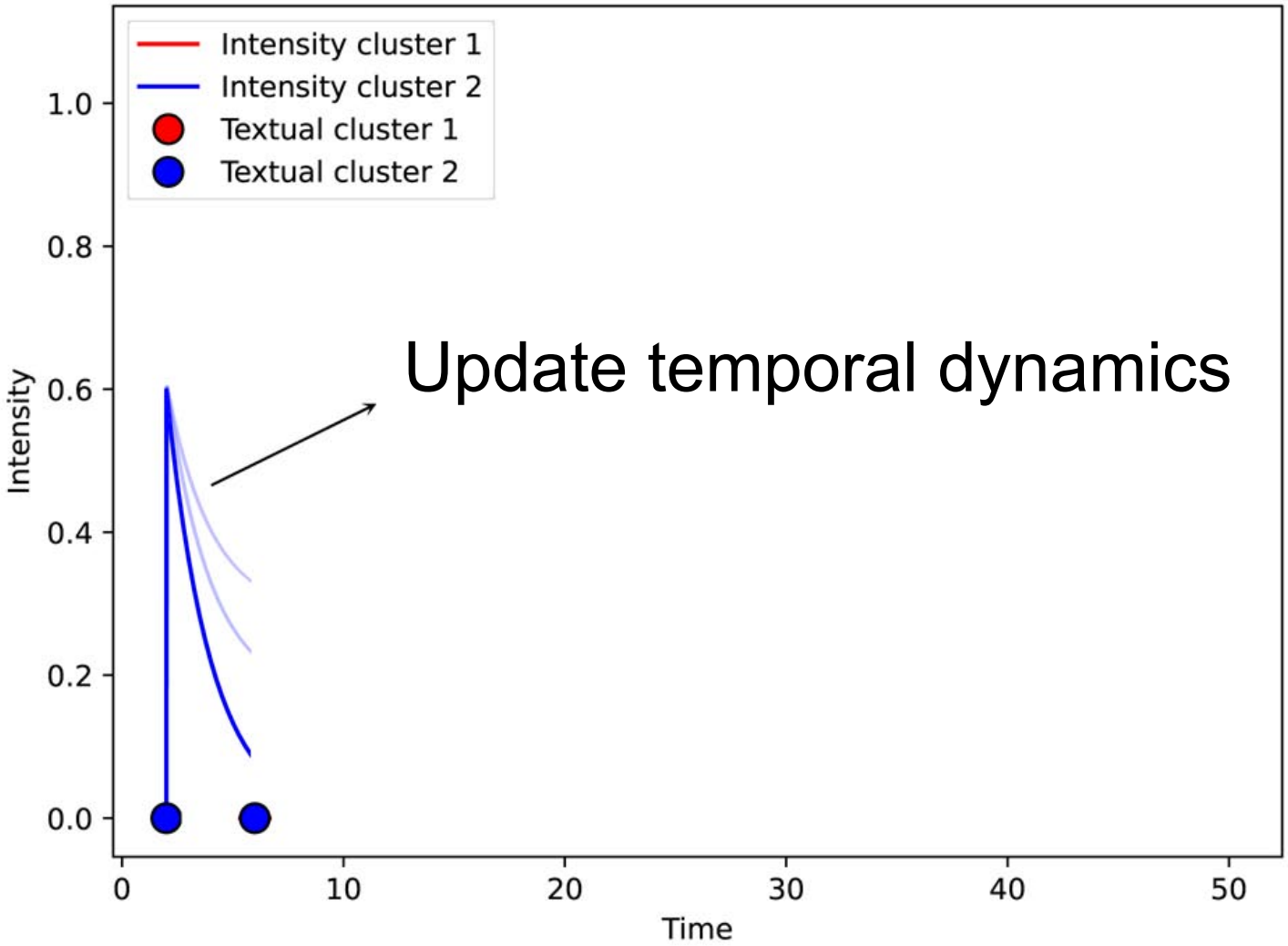
Processus de Dirichlet-Hawkes - Inférence (1 instance)



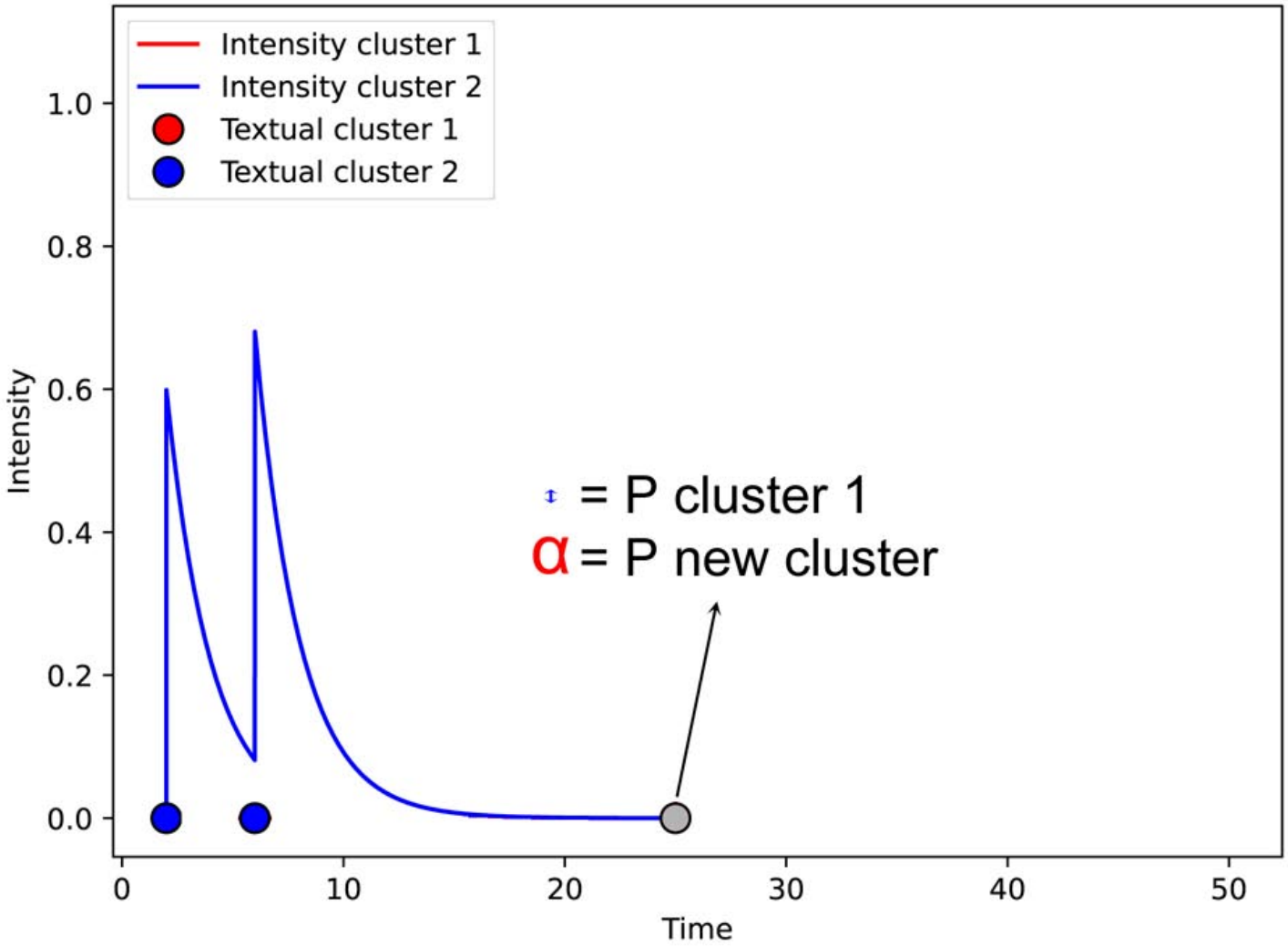
Processus de Dirichlet-Hawkes - Inférence (1 instance)



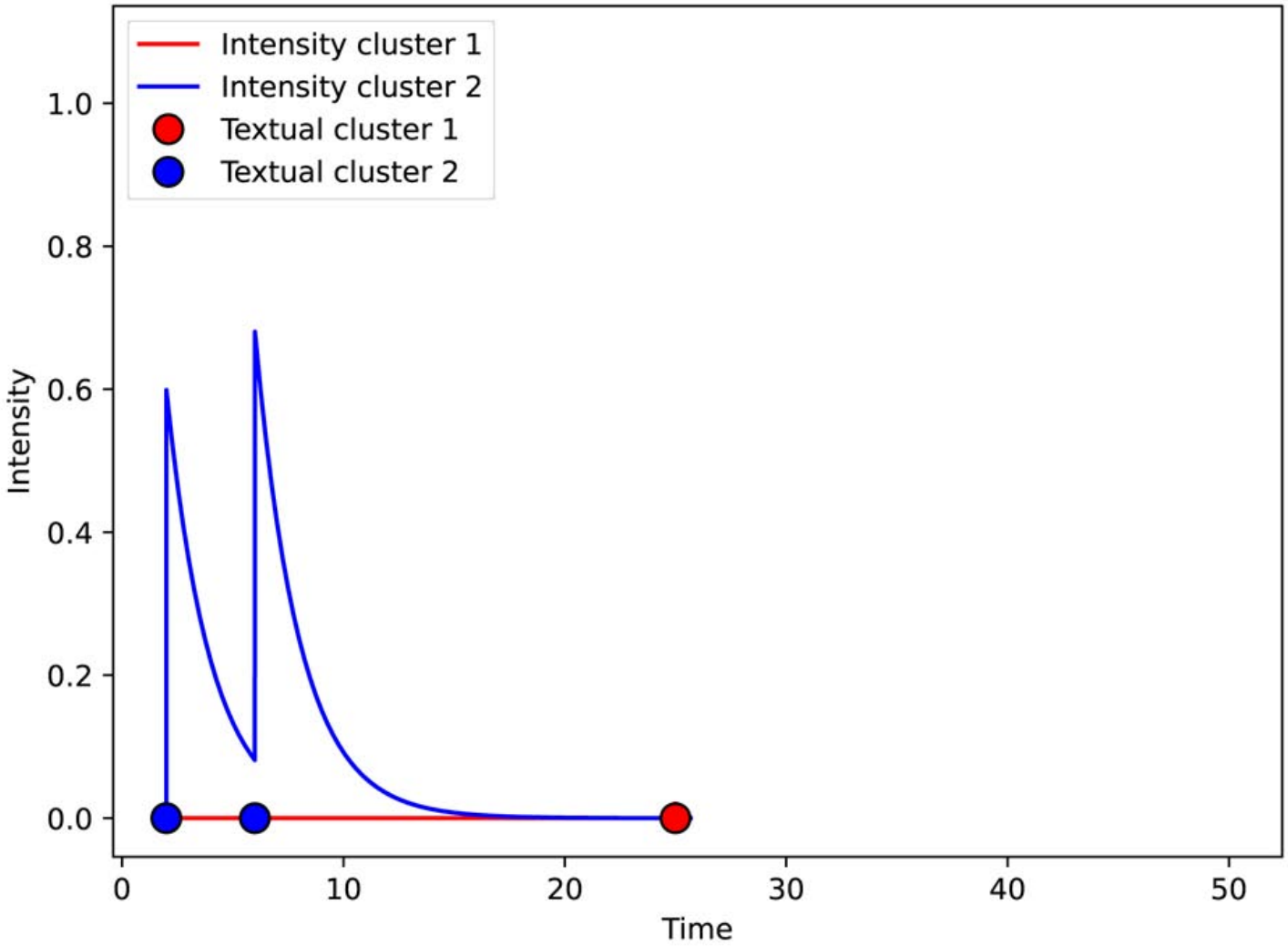
Processus de Dirichlet-Hawkes - Inférence (1 instance)



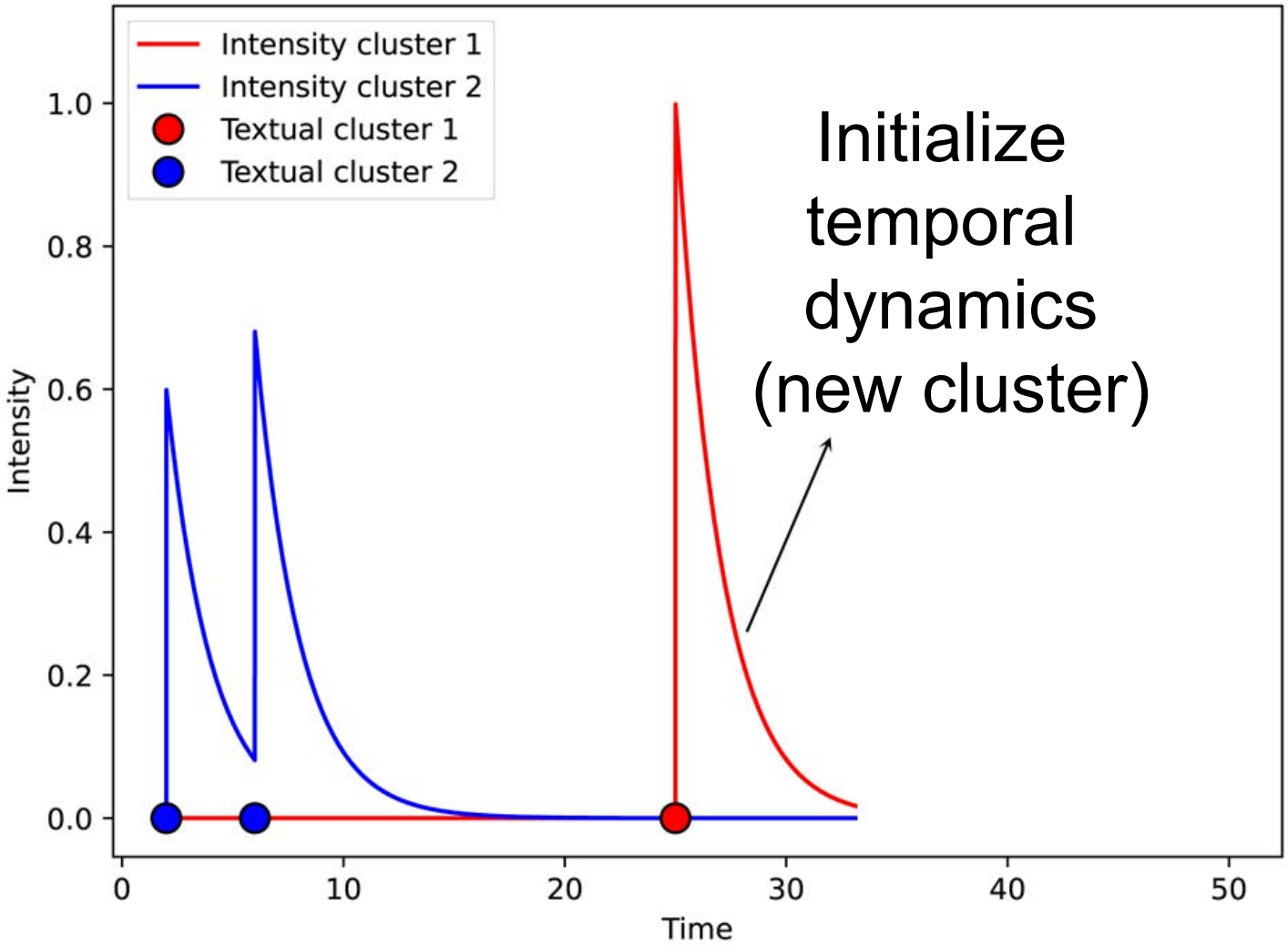
Processus de Dirichlet-Hawkes - Inférence (1 instance)



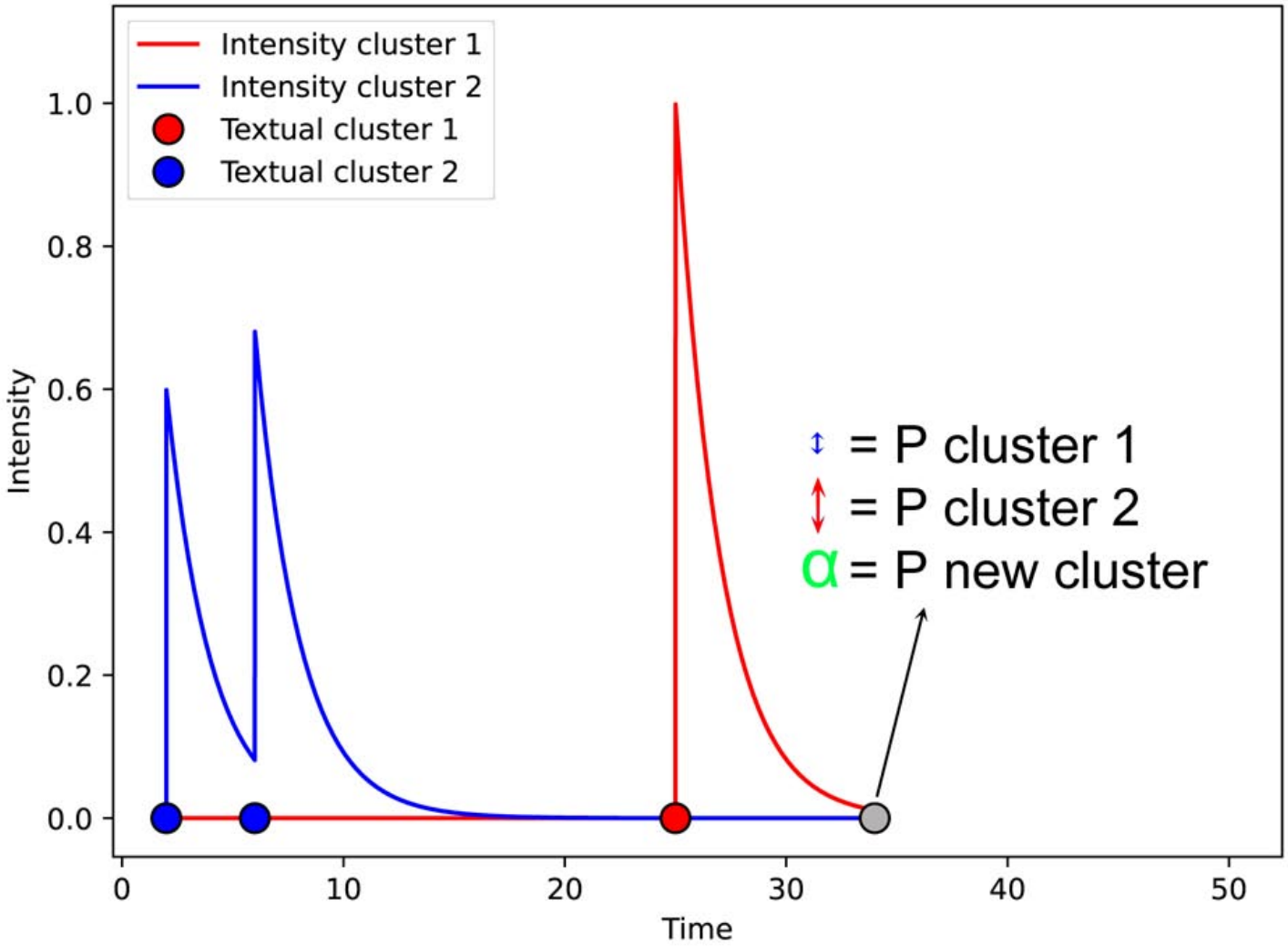
Processus de Dirichlet-Hawkes - Inférence (1 instance)



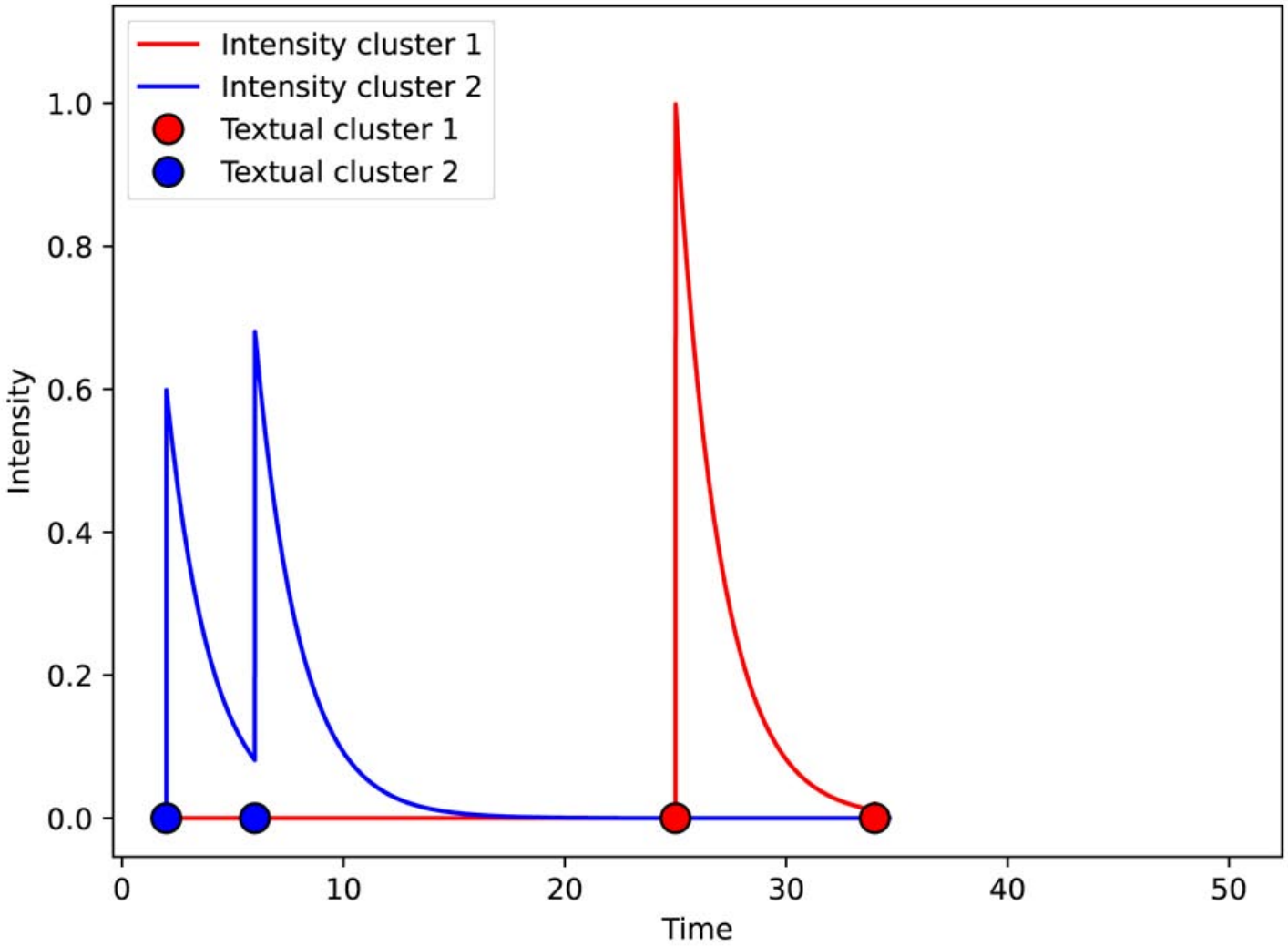
Processus de Dirichlet-Hawkes - Inférence (1 instance)



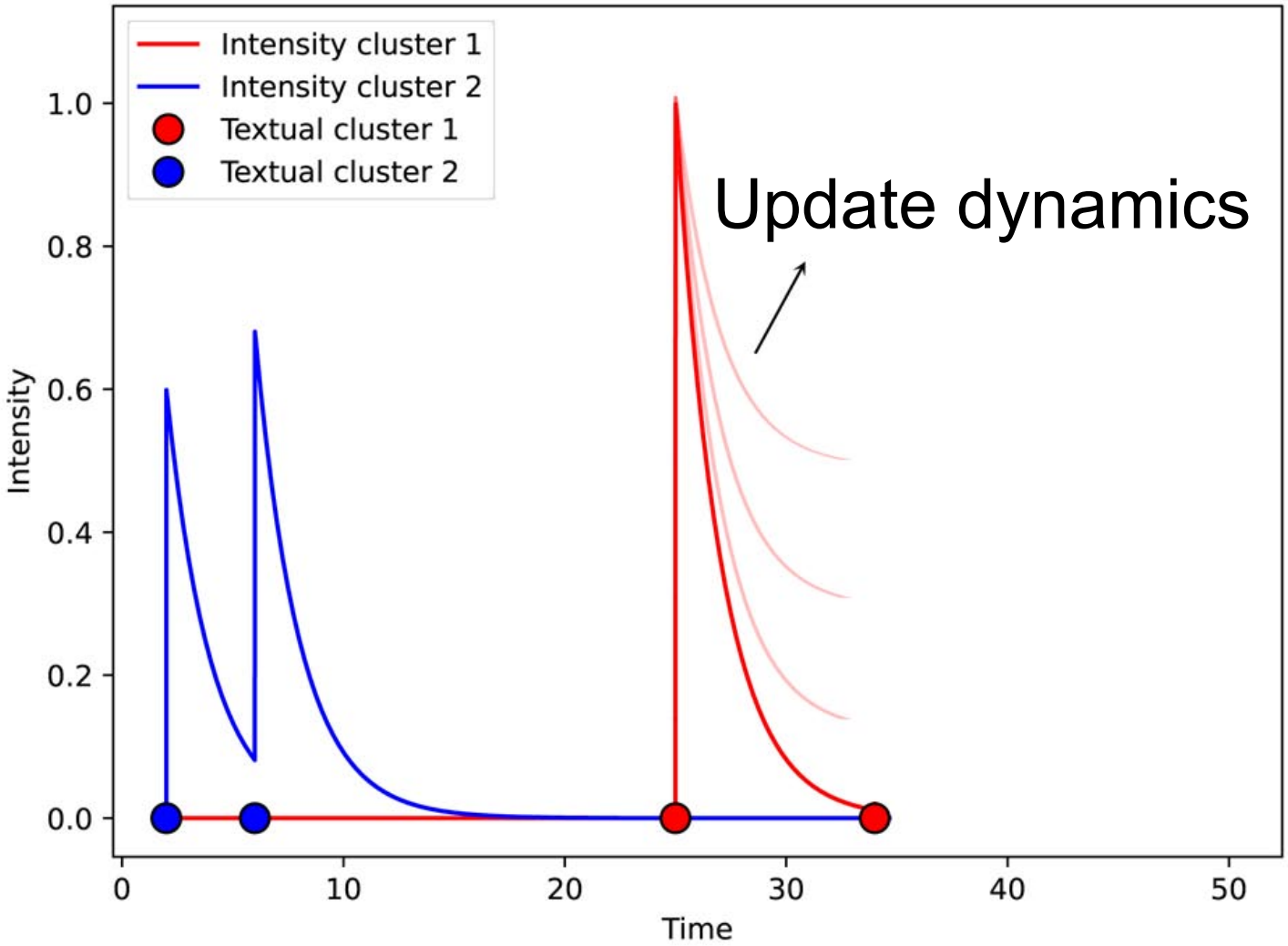
Processus de Dirichlet-Hawkes - Inférence (1 instance)



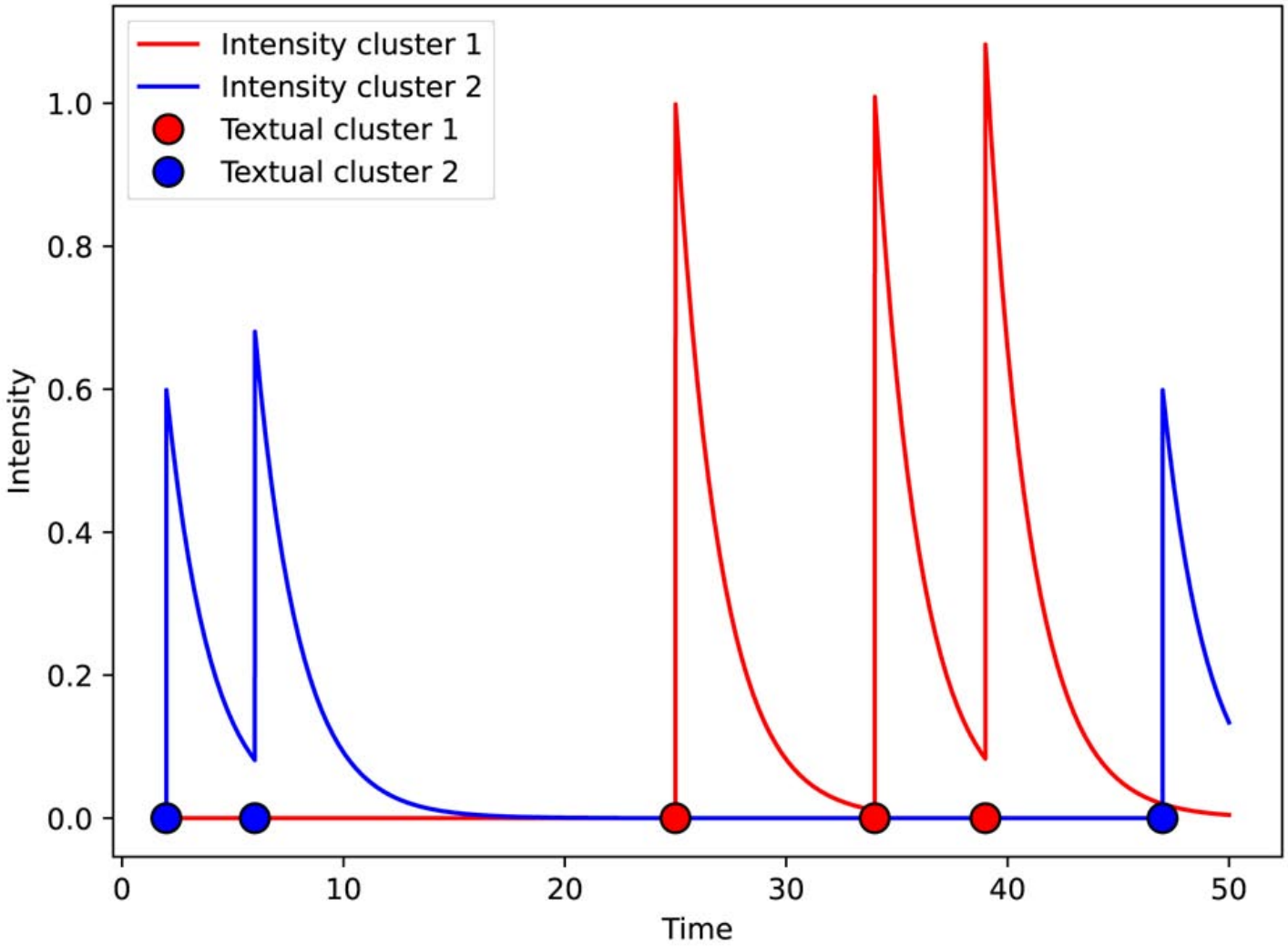
Processus de Dirichlet-Hawkes - Inférence (1 instance)



Processus de Dirichlet-Hawkes - Inférence (1 instance)

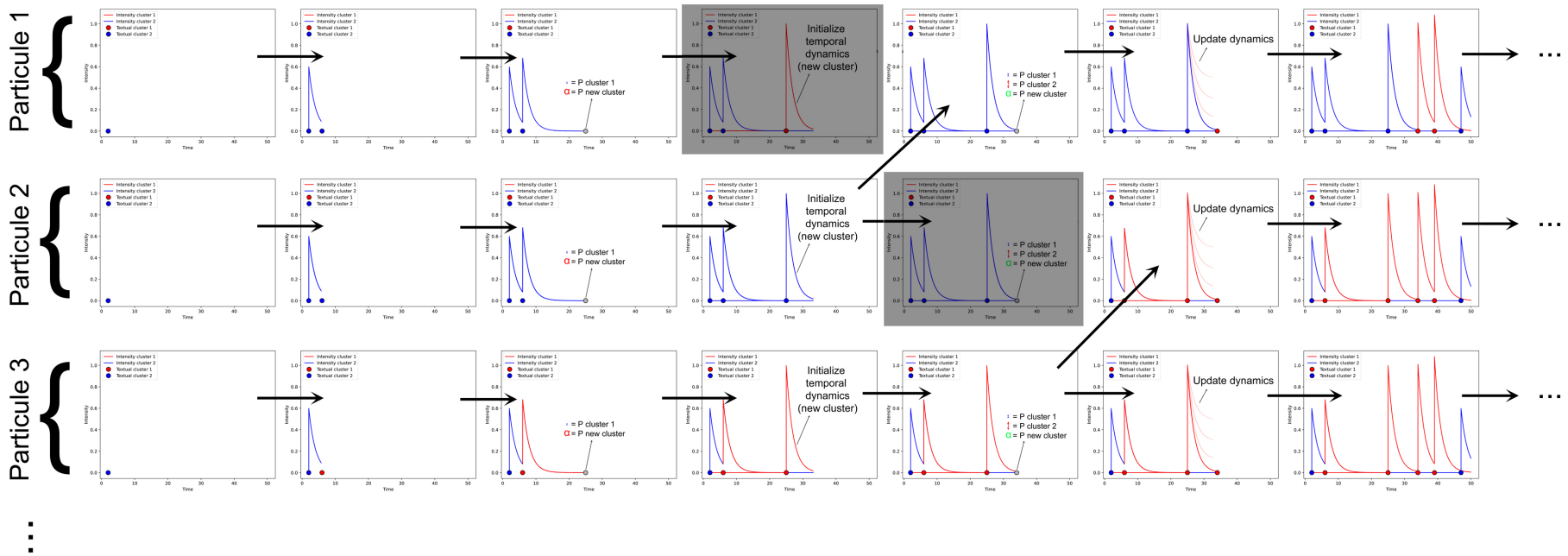


Processus de Dirichlet-Hawkes - Inférence (1 instance)



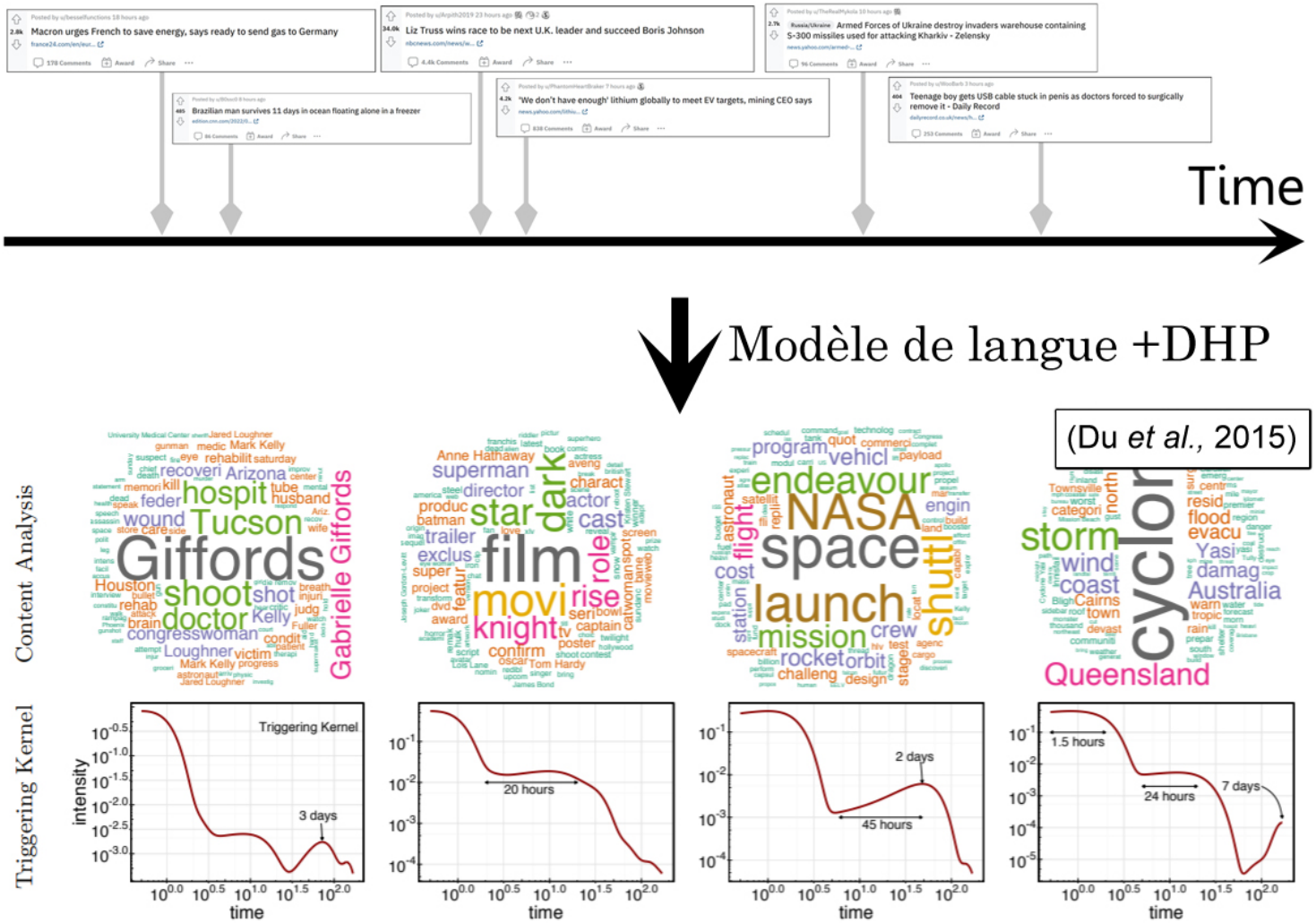
Processus de Dirichlet-Hawkes - Inférence (SMC)

- Algorithme de Monte Carlo séquentiel



Processus de Dirichlet-Hawkes - Résultats

→ Modélise des interactions intra-cluster



Powered Dirichlet-Hawkes Process - Fusion de PDP et DHP

- Le Dirichlet-Hawkes Process fait des hypothèses :
  - Situations simples (dynamique de chaque cluster distincte)
  - Dépendance linéaire à l'information temporelle
- Powered Dirichlet-Hawkes Prior (Poux-Médard et al., 2021b) :

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

- Généralisation :
  - Uniform process : $r = 0$ (info textuelle uniquement)
  - Dirichlet-Hawkes process : $r = 1$ (info textuelle et temporelle)
  - Deterministic Hawkes process : $r \to \infty$ (info temporelle uniquement)

Powered Dirichlet-Hawkes Process - Effet de r

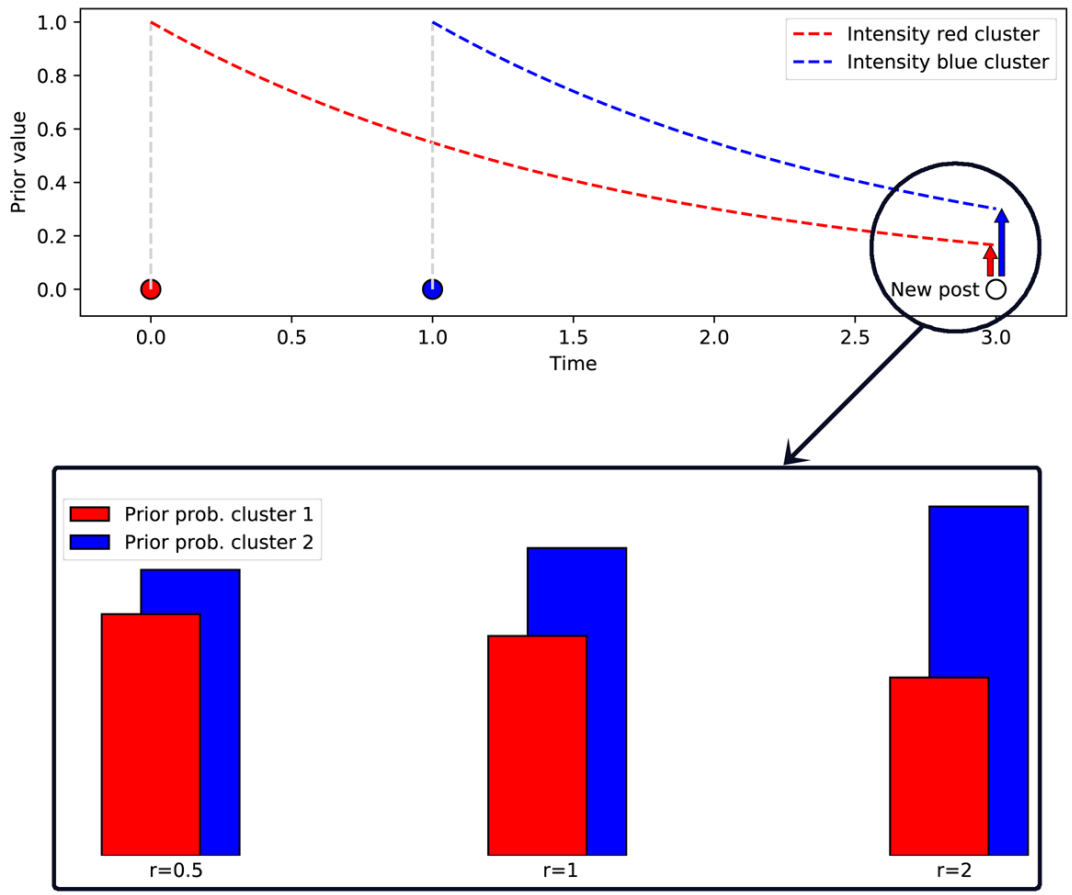


Figure 14 – Effet de r sur les probabilités a priori des clusters

Powered Dirichlet-Hawkes Process - Implications

$$P(\text{cluster}|\text{texte, temps}) \propto \underbrace{P(\text{texte}|\text{cluster})}_{\text{Vraisemblance textuelle}} \times \underbrace{P(\text{cluster}|\text{temps, r})}_{\text{PDHP}}$$

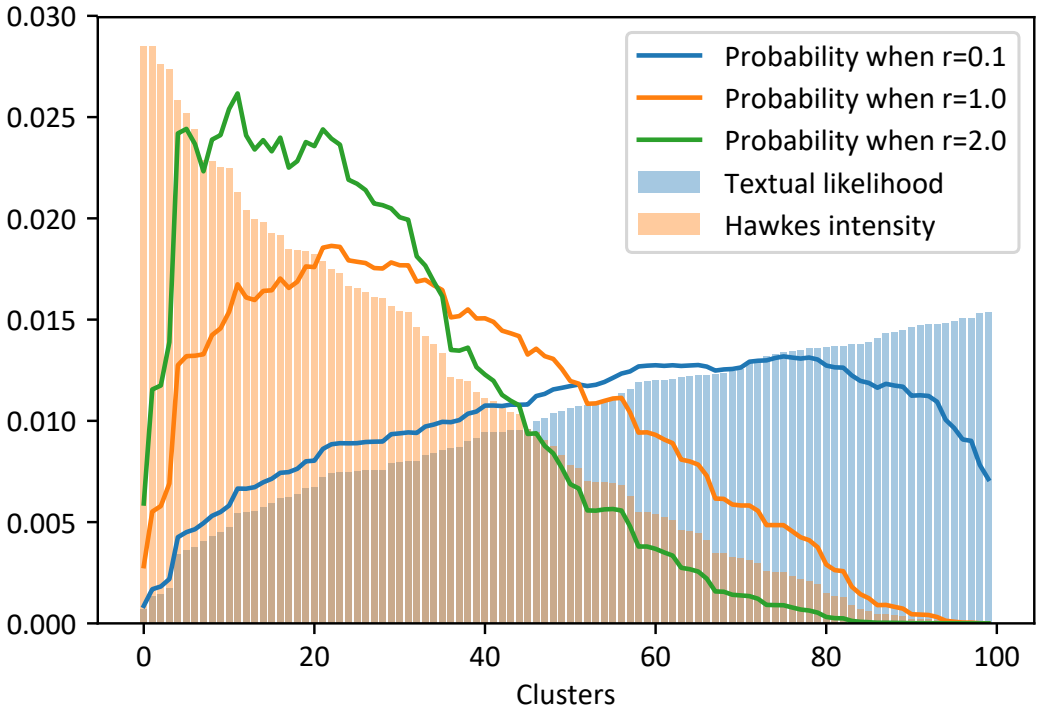


Figure 15 – Effet de r sur la probabilité a posteriori des clusters

Powered Dirichlet-Hawkes Process - Situations complexes

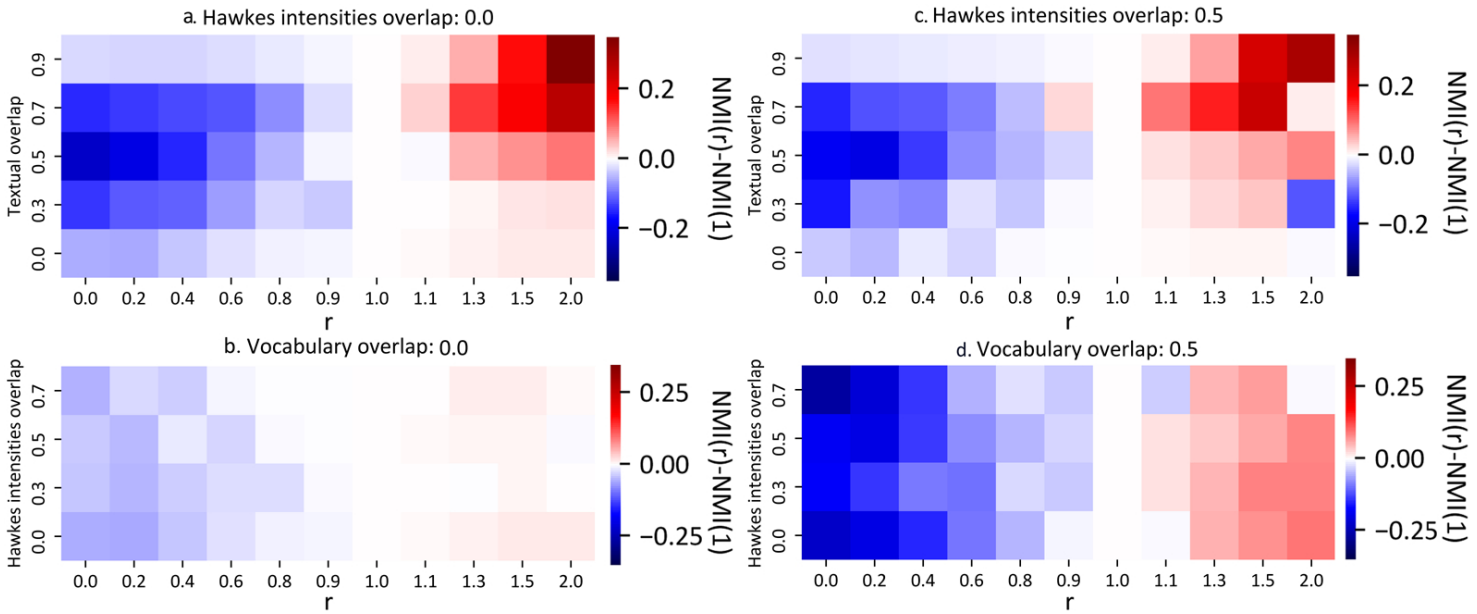
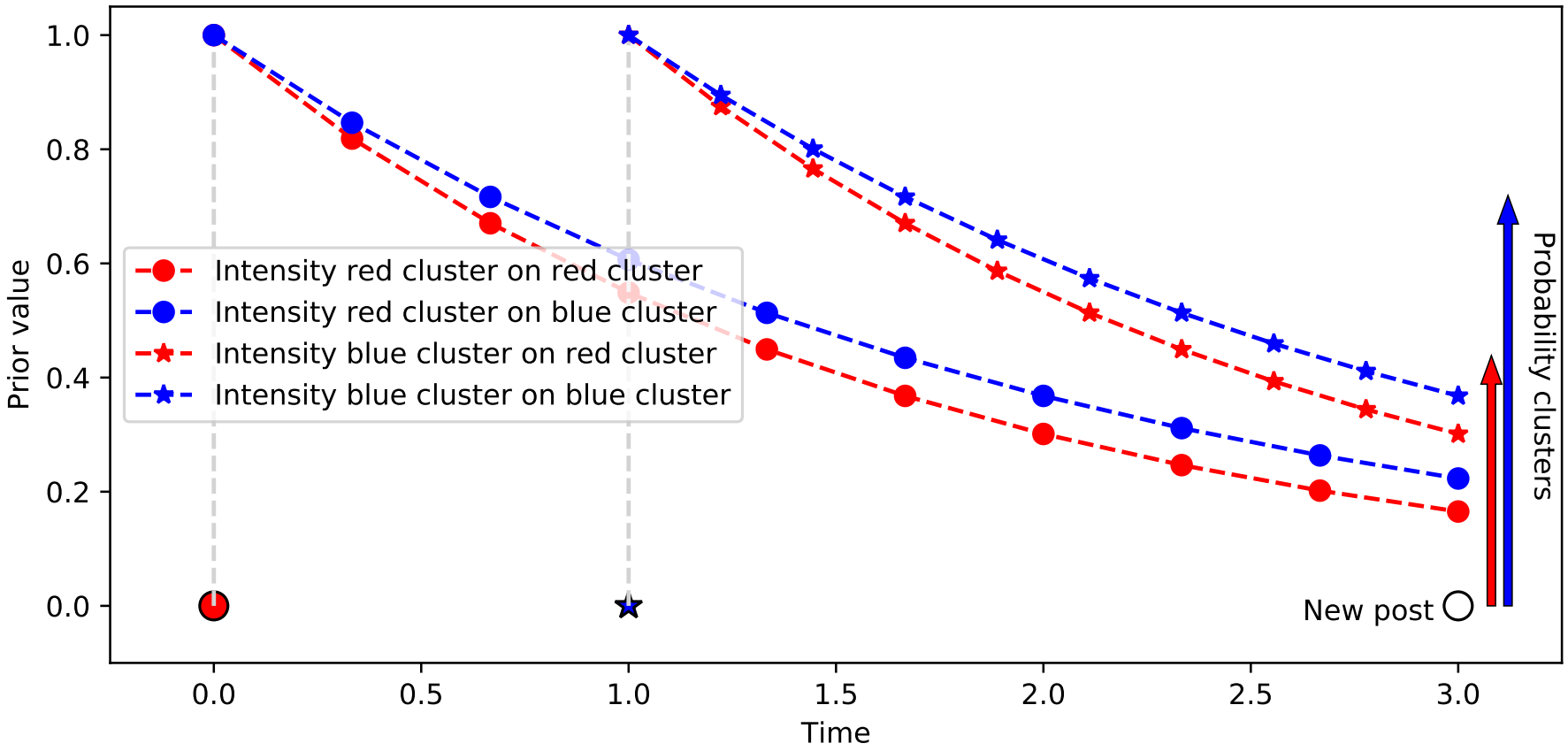


Figure 16 – PDHP permet de raffiner les résultats de DHP

- Score : normalized mutual information
- PDHP s'adapte à plusieurs situations mieux que DHP (+0.3 NMI) :
  - Fort recouvrement de vocabulaire entre clusters
  - Fort recouvrement des dynamiques entre clusters

Multivariate Powered Dirichlet-Hawkes Process

- Multivariate Powered Dirichlet-Hawkes prior (Poux-Médard et al., 2022)
  - Modélise des interactions intra-cluster et extra-cluster



Multivariate Powered Dirichlet-Hawkes Process - Données synthétiques

- Génération de données synthétiques
  - 2 clusters, 20 mots/doc., 5000 obs., 100 runs
- Récupération des clusters sous-jacents ?

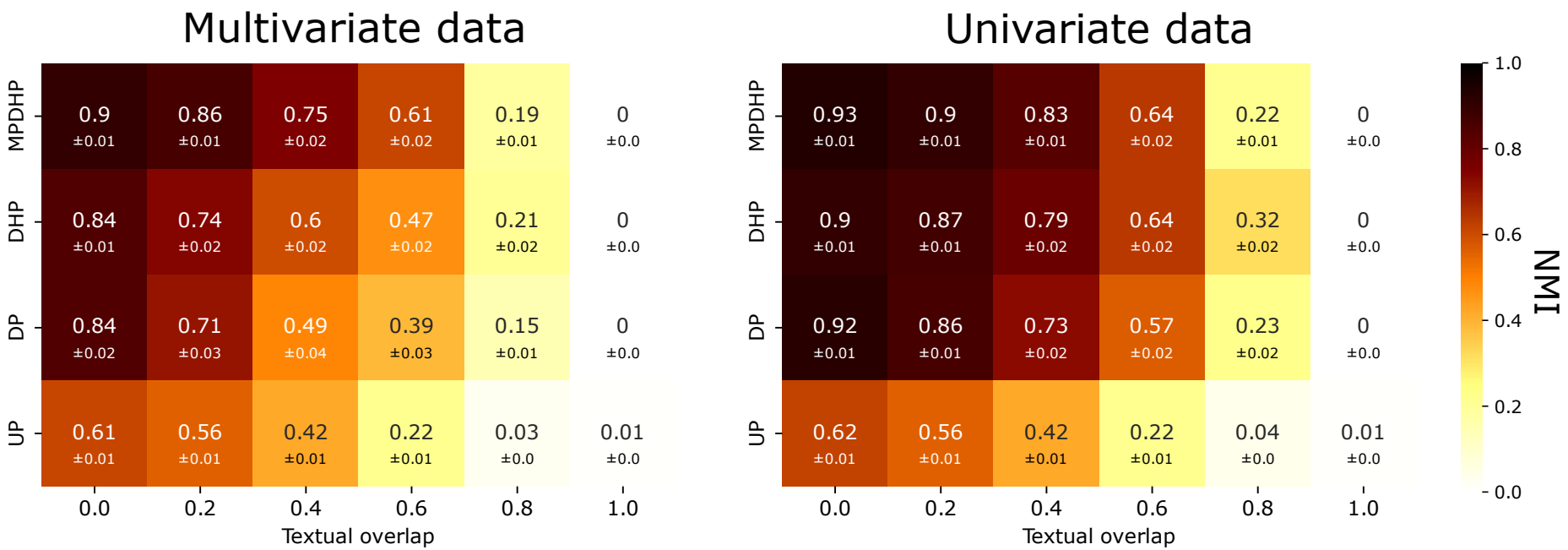
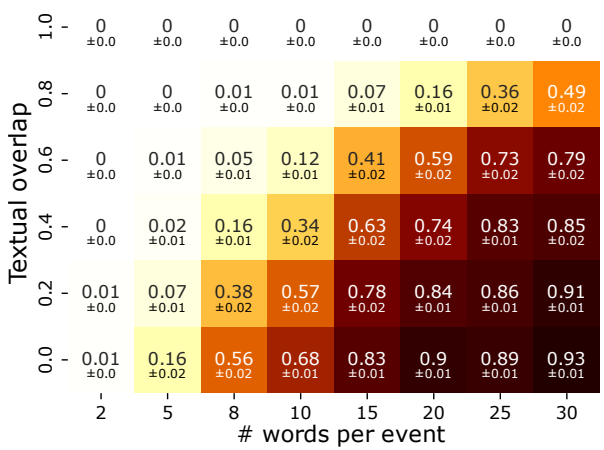


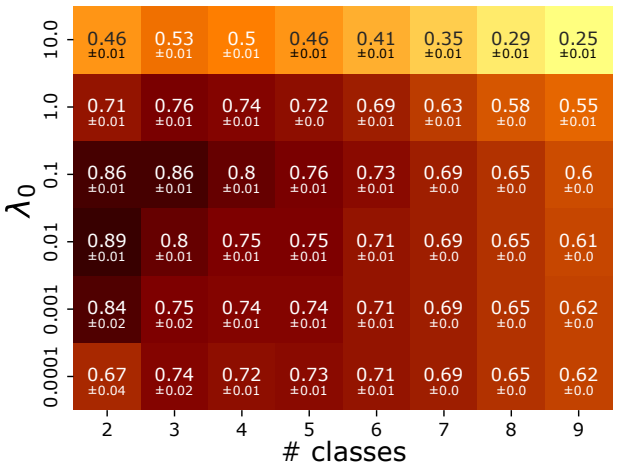
Figure 17 – Résultats de MPDHP sur des données synthétiques

Multivariate Powered Dirichlet-Hawkes Process - Cas limites

MPDHP fonctionne avec peu de données



MPDHP fonctionne lorsque plusieurs clusters existent simultanément



MPDHP requiert peu de ressources informatiques

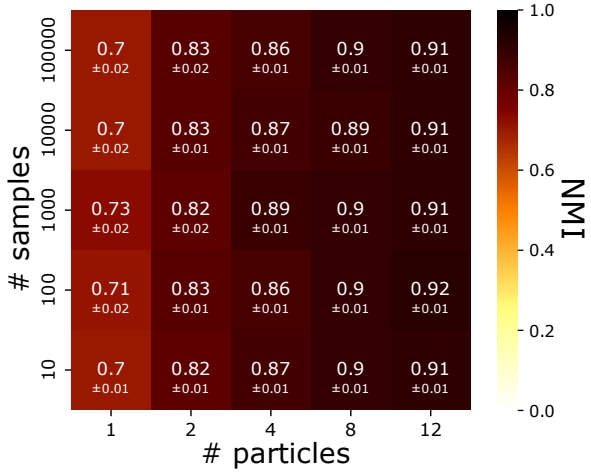


Figure 18 – Résultats complémentaires de MPDHP sur des données synthétiques

Jeu de données Reddit

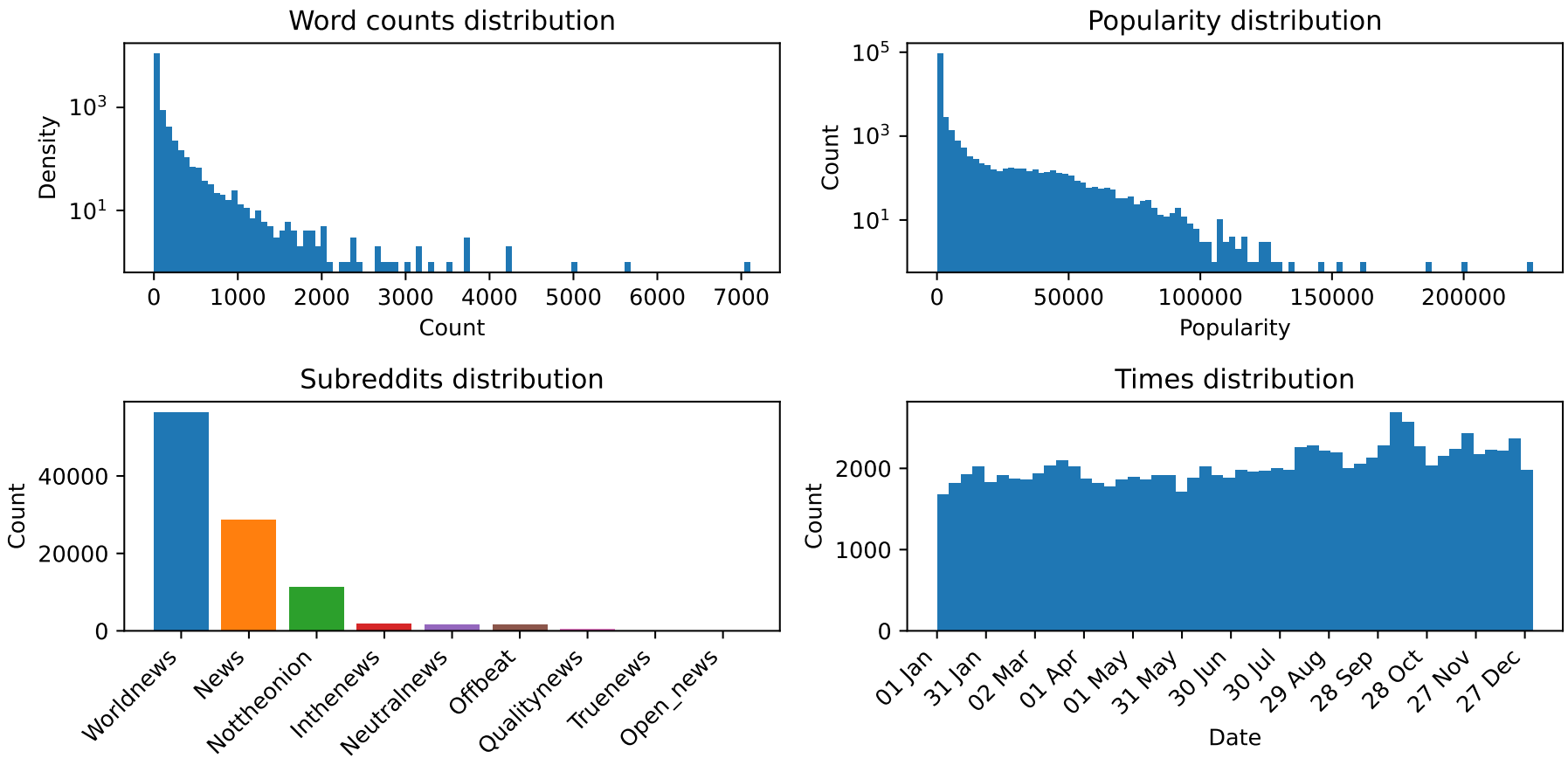


Figure 19 – Caractéristiques du jeu de données Reddit News (∼100 000 titres)

Clusters inférés

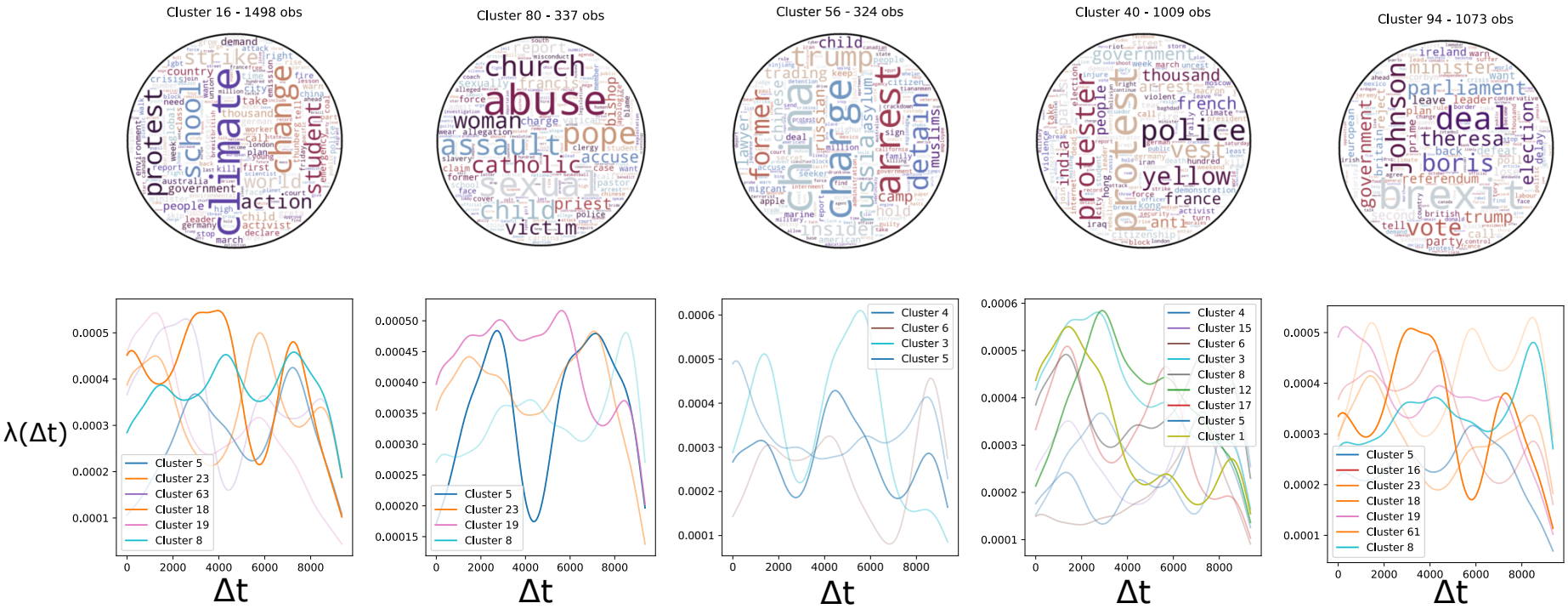


Figure 20 – Quelques clusters inférés par MPDHP (haut), et leur dynamique d'interaction avec les autres clusters (bas). Ex. : un document du cluster 16 influe fortement la probabilité de publication d'un document du cluster 23 environ 4000s après sa publication.

Réseau d'interaction entre clusters

- Réseau d'interaction entre clusters

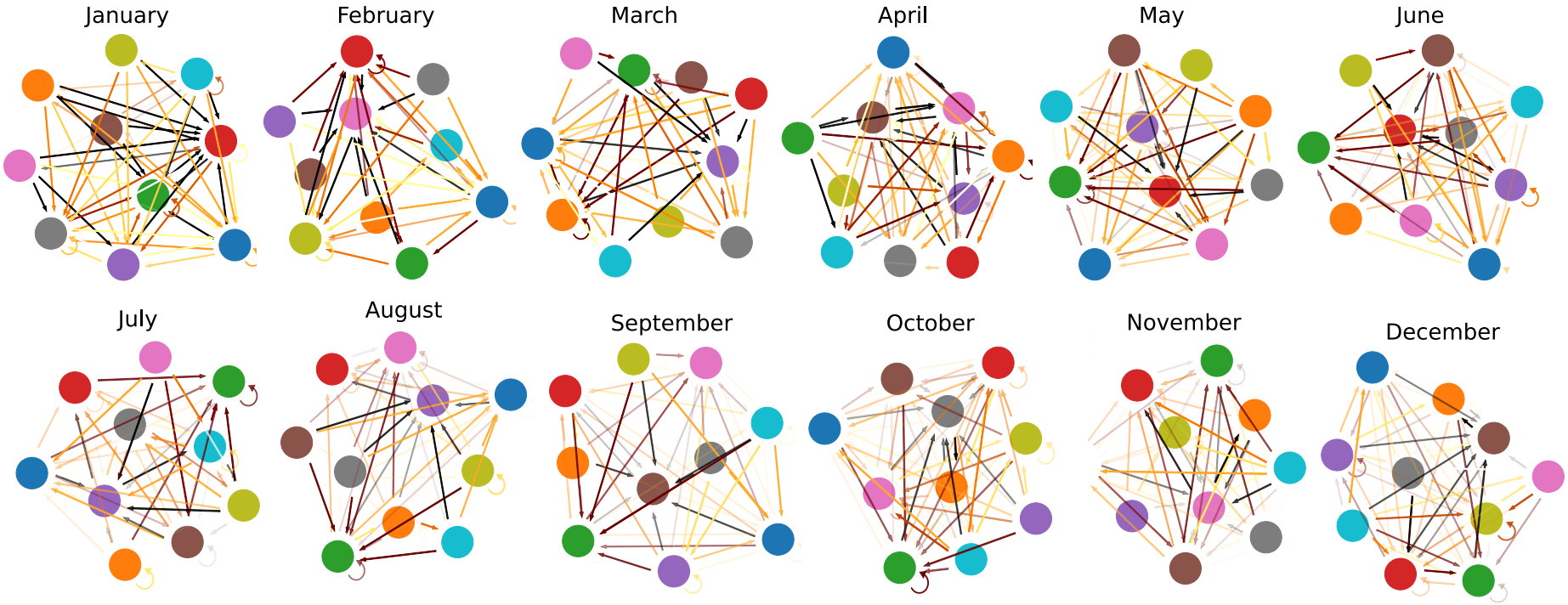


Figure 21 – Évolution du réseau d'interaction à courte portée entre les clusters sur toute l'année. Noeuds = clusters ; Liens = force d'interaction

Force des interactions - Résultats quantitatifs

Table 2 – **Force des interactions** – Les interactions sont faibles en moyenne ; A = Matrice d'adjacence ; W = Interactions effectives

$\vec{\kappa}(t)$	r	<A>(10^{-3})	<W>(10^{-5})	<A>$_W(10^{-3})$	$\frac{<W^{intra}>}{<W^{extra}>}$
Minute	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	0.5	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	0.5	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)

Force des interactions - Illustration

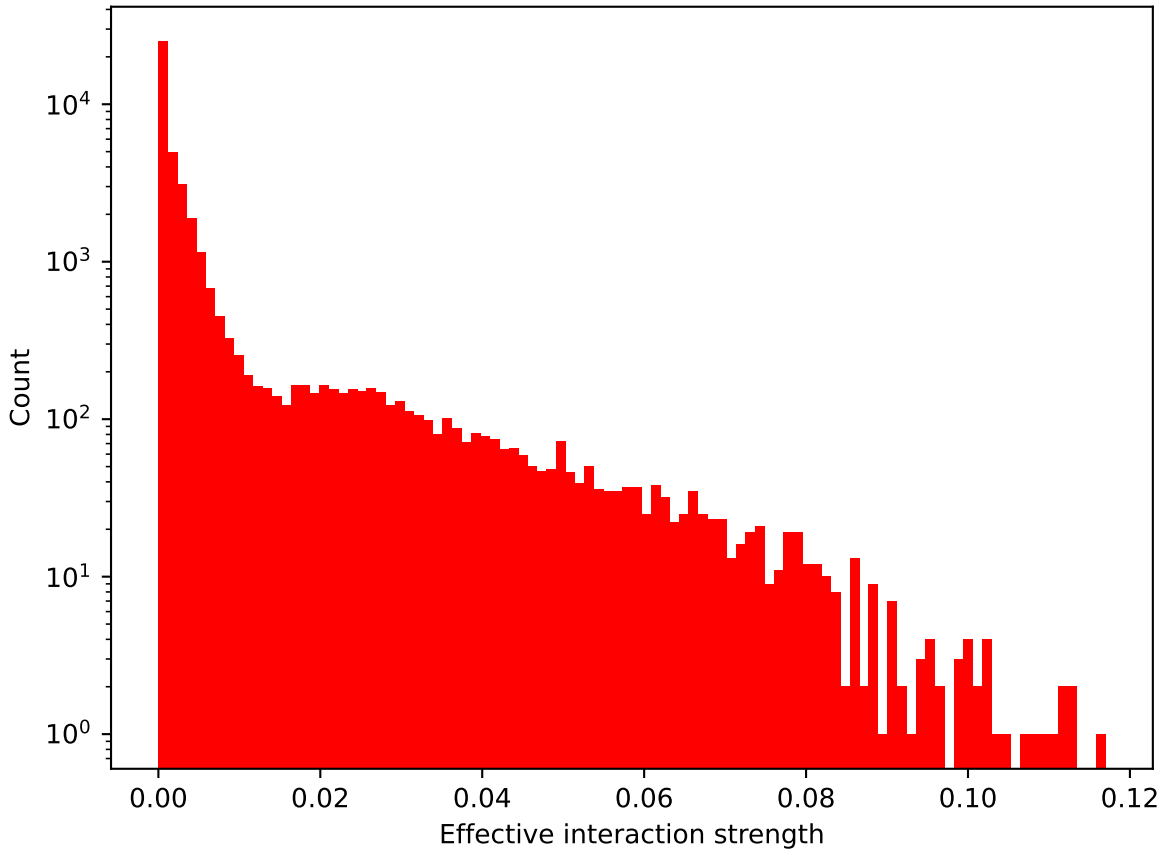


Figure 22 – Distribution typique de la force des interactions dans Reddit News (heure, r=1).

Portée des interactions

Table 3 – La force des interactions effectives diminue dans le temps

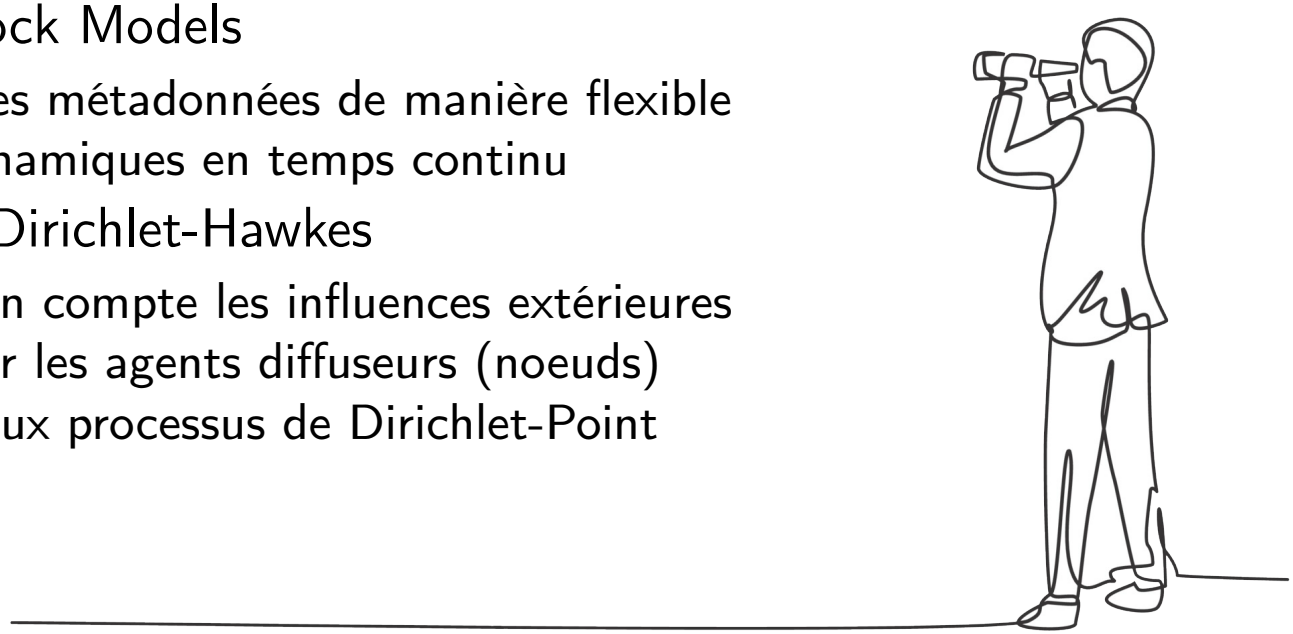
$\vec{\kappa}(t)$	r	κ_1	κ_2	κ_3	κ_4	κ_5	κ_6	κ_7	κ_8	κ_9
		0m	10m	20m	30m	40m	50m	60m	70m	80m
Minute	0.5	218	509	457	424	371	340	313	178	52
	1	142	435	396	388	343	327	272	187	45
	1.5	104	388	366	353	326	333	290	215	61
		0h	2h	4h	6h	8h	-	-	-	-
Hour	0.5	62	149	119	137	92				
	1	77	164	172	149	111				
	1.5	104	244	223	197	156				
		0d	1d	2d	3d	4d	5d	6d	-	-
Day	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		

Conclusion

- Dans la plupart des jeux de données étudiés :
  - Les interactions majeures sont peu fréquentes
  - Les interactions sont brèves
  - Les interactions semblent jouer un rôle mineur dans les diffusions
- Cependant :
  - Considérer les interactions améliore sensiblement nos résultats
  - Leur importance fluctue en fonction des jeux de données
- Ainsi :
  - Nos conclusions n'ont pas vocation à être universelles
  - Nos modèles permettent de les étudier dans des cas spécifiques

Perspectives - Aperçu

- Pistes possibles :
  - Modèles non limités aux interactions
    - SBMs : recommandation, archéologie, ...
    - DHPs : résumés, topic modelling, modération, ...
  - Stochastic Block Models
    - Utiliser des métadonnées de manière flexible
    - SBMs dynamiques en temps continu
  - Processus de Dirichlet-Hawkes
    - Prendre en compte les influences extérieures
    - Considérer les agents diffuseurs (noeuds)
    - Étendre aux processus de Dirichlet-Point



Perspectives - Dirichlet-Point processes

- Les processus de Dirichlet-Hawkes sont un cas particulier des processus de Dirichlet-Point.
  - (Du et al., 2015) a le premier exploré cette connexion avec DHP
  - D'autres modèles ont suivi : HDHP, IBHP, PDHP, MPDHP
- Autant de nouvelles perspectives que de combinaisons entre :
  - Les processus de Dirichlet et leurs variantes
  - Les processus ponctuels et leurs variantes

(**DP**, HDP, nHDP, **PDP**, IBP, PIBP, PnHDP, PPY, PnPY, PHPY, ...)

$\times$

(**Hawkes**, **Multi Hawkes**, Survival, Cox, Poisson, Determinantal, ...)

$=$

(**DHP**, HDHP, IBHP, **PDHP**, **MPDHP**, ... ?)

Merci de votre attention !

Table 4 – Contributions détaillées dans le manuscrit de thèse

	SIMSBM	IMMSBM	SDSBM	InterRate	PDP	PDHP	MPDHP
Publication	ICDM'22	RecSys'21	-	ECML-PKDD'21	-	ICDM'21	CNA'22
Auto-interactions	x	x	x	x	x	x	x
Interactions paires	x	x	x	x		x	x
Interactions n-plet	x		x			x	x
Clustering	x	x	x		x	x	x
Temps discrets			x	x		x	x
Temps continus				x		x	x
Inférence séquentielle					x	x	x

Bibliographie I

Ahmed, A. and Xing, E. (2008). Dynamic non-parametric mixture models and the recurrent chinese restaurant process : with applications to evolutionary clustering. *SIAM International Conference on Data Mining*, pages 219–230.

Blei, D. M. and Frazier, P. (2010). Distance dependent chinese restaurant processes. *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 87–94.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, pages 113–120.

Diao, Q. and Jiang, J. (2014). Recurrent chinese restaurant process with a duration-based discount for event identification from twitter. *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, pages 388–397.

Du, N., Farajtabar, M., Ahmed, A., Smola, A., and Song, L. (2015). Dirichlet-hawkes processes with applications to clustering continuous-time document streams. *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Du, N., Song, L., Smola, A., and Yuan, M. (2012). Learning networks of heterogeneous influence. *NIPS*, 4 :2780–2788.

Bibliographie II

Du, N., Song, L., Woo, H., and Zha, H. (2013). Uncover topic-sensitive information diffusion networks. *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, 31 :229–237.

Gomez-Rodriguez, M., Balduzzi, D., and Schölkopf, B. (2011). Uncovering the temporal dynamics of diffusion networks. *ICML*, pages 561–568.

Gomez-Rodriguez, M., Leskovec, J., and Schölkopf, B. (2013). Modeling information propagation with survival theory. *ICML*, 28 :666–674.

Kapoor, J., Vergari, A., Valera, I., and Gomez-Rodriguez, M. (2018). Bayesian nonparametric hawkes processes. *Proceedings of the Bayesian Nonparametrics workshop at the 32nd Conference on Neural Information Processing Systems (NIPS)*.

Lee, C. J. and Sang, H. (2022). Why the rich get richer ? on the balancedness of random partition models. In *ICML*.

Myers, S. A. and Leskovec, J. (2012). Clash of the contagions : Cooperation and competition in information diffusion. *2012 IEEE 12th International Conference on Data Mining*, pages 539–548.

Bibliographie III

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2) :855 – 900.

Poux-Médard, G., Velcin, J., and Loudcher, S. (2021a). Information interactions in outcome prediction : Quantification and interpretation using stochastic block models. *Fifteenth ACM Conference on Recommender Systems (RecSys)*, page 199–208.

Poux-Médard, G., Velcin, J., and Loudcher, S. (2021b). Powered hawkes-dirichlet process : Challenging textual clustering using a flexible temporal prior. *2021 IEEE International Conference on Data Mining (ICDM)*, pages 509–518.

Poux-Médard, G., Velcin, J., and Loudcher, S. (2022). Properties of reddit news topical interactions. In *Complex Networks & Their Applications XI (under press)*. Springer International Publishing.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested dirichlet process. *Journal of the American Statistical Association*, 103(483) :1131–1154.

Rodriguez, M. G., Leskovec, J., and Schoelkopf, B. (2013). Structure and dynamics of information pathways in online media. *WSDM*.

Bibliographie IV

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476) :1566–1581.

Wallach, H., Jensen, S., Dicker, L., and Heller, K. (2010). An alternative prior process for nonparametric bayesian clustering. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 892–899.

Wallach, H., Mimno, D., and McCallum, A. (2009). Rethinking lda : Why priors matter. *Advances in Neural Information Processing Systems*, 22.

Wang, L., Ermon, S., and Hopcroft, J. E. (2012). Feature-enhanced probabilistic models for diffusion network inference. *Machine Learning and Knowledge Discovery in Databases*, pages 499–514.

Wang, W., Liu, Q.-H., Liang, J., Hu, Y., and Zhou, T. (2019). Coevolution spreading in complex networks. *Physics Reports*, 820 :1–51. Coevolution spreading in complex networks.

Welling, M. (2006). Flexible priors for infinite mixture models. *Workshop on learning with non-parametric Bayesian methods*.

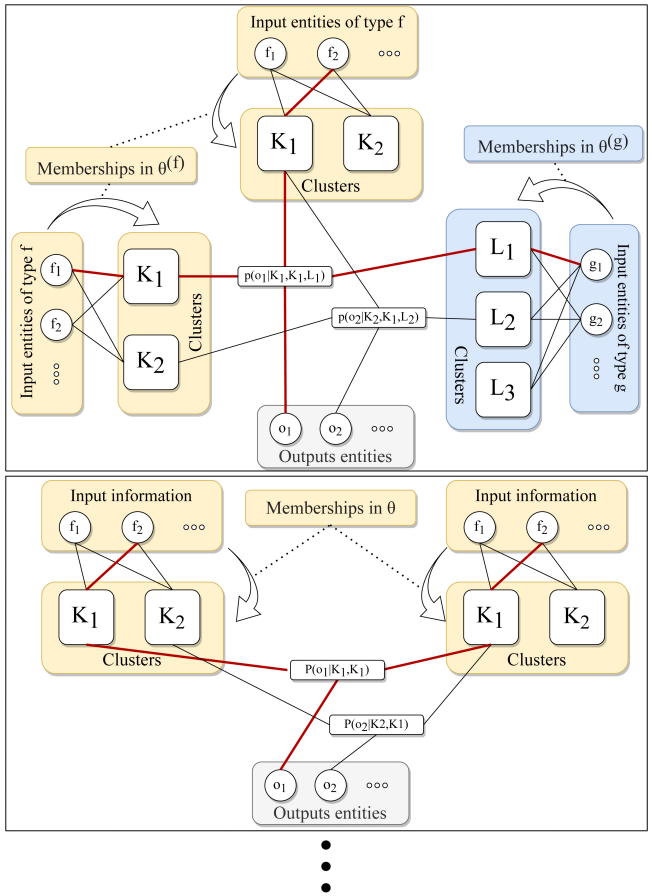
Bibliographie V

Zarezade, A., Khodadadi, A., Farajtabar, M., Rabiee, H. R., and Zha, H. (2017). Correlated cascades : Compete or cooperate. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 238–244.

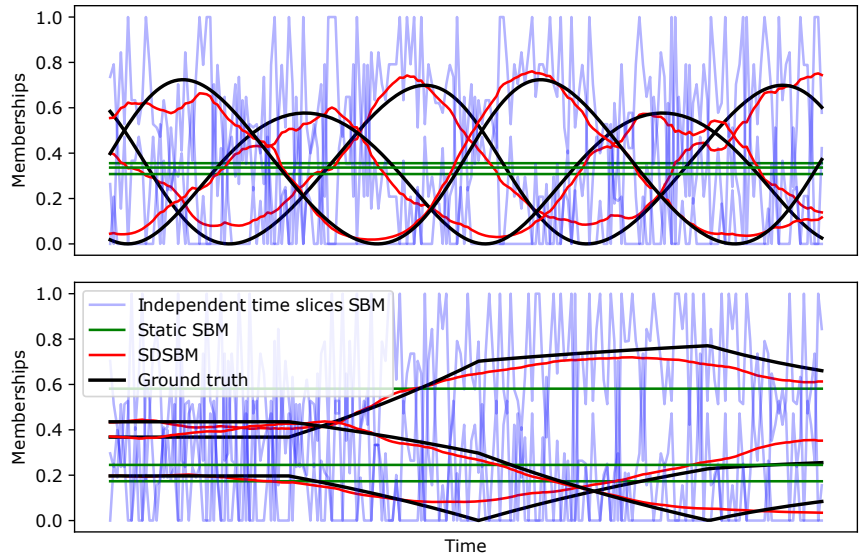
Zhu, Z., Gao, C., and Zhang, Y. (2020). Cooperation and competition among information on social networks. *Nature Sci Rep*, 3103 :12160.

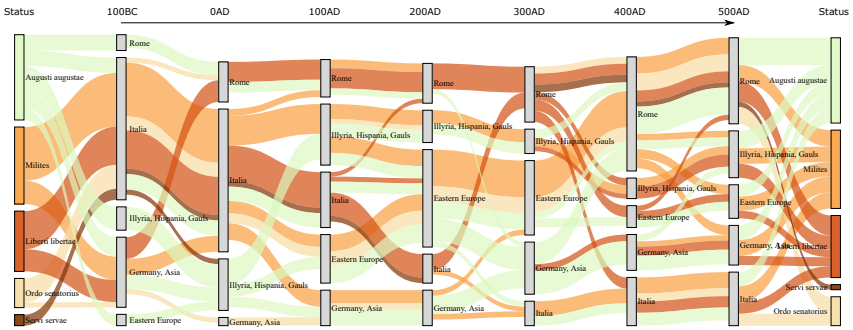
Extensions : SIMSBM et SDSBM

SIMSBM (ICDM 2022)



SDSBM





Modèle graphique MMSBM

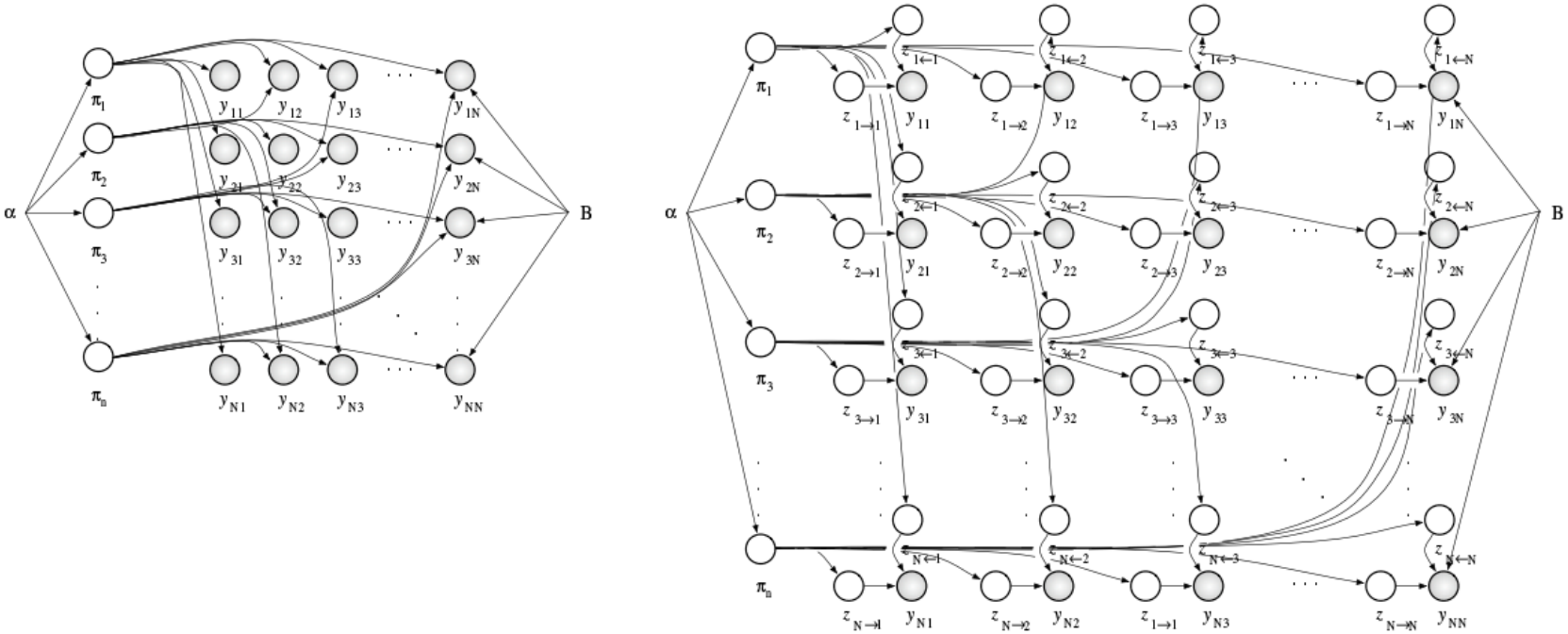


Figure 1: Two graphical model representations of the mixed membership stochastic blockmodel (MMB). Intuitively, the MMB summarized the variability of a graph with the blockmodel B and node-specific mixed membership vectors (left). In detail, a mixed membership, $\pi_n(k)$, quantifies the expected proportion of times node n instantiates the connectivity pattern of group k, according to the blockmodel. In any given interaction, $Y(n,m)$, however, node n instantiates the connectivity pattern of a single group, $z_{n\to m}(k)$. (right). We did not draw all the arrows out of the block model B for clarity; all interactions depend on it.

Inference

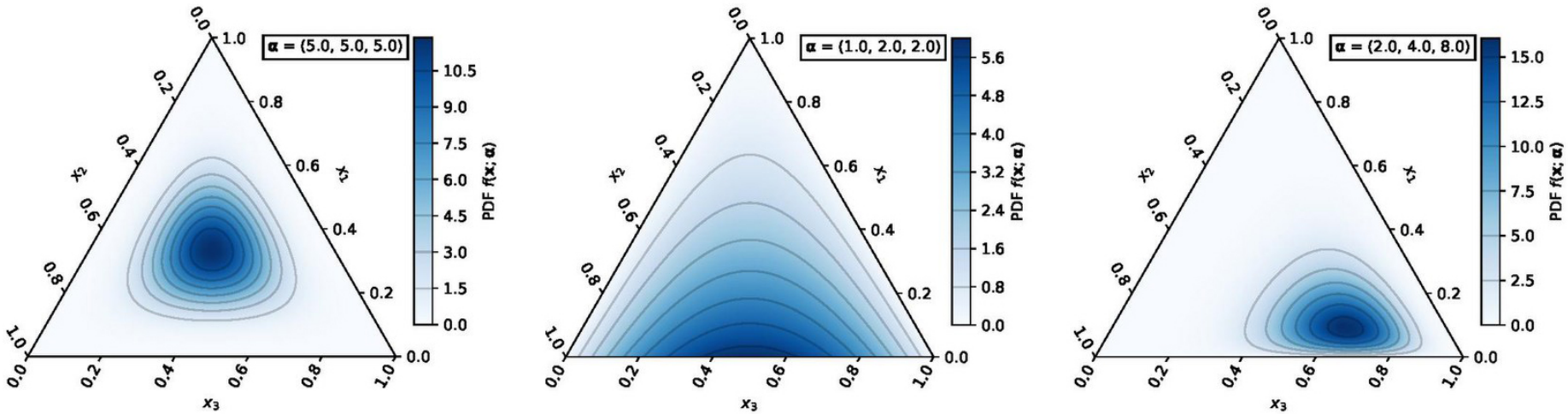
- Log-likelihood d'un flux de données $\mathcal{D} = \{t_0, ..., t_N\}$:

$$\ell(\lambda, \mathcal{D}) = -\int_{t_0}^{t_N} \lambda(t)dt + \sum_{t_i < t_N} \log \lambda(t_i) = \log \lambda(t_1) - \int_{t_0}^{t_1} \lambda(t)dt$$
$$+ \log \lambda(t_2) - \int_{t_1}^{t_2} \lambda(t)dt$$
$$+ ...$$
$$+ \log \lambda(t_N) - \int_{t_{N-1}}^{t_N} \lambda(t)dt$$

- Convexe pour certaines expressions de $\lambda(t)$ (exp, ray, PL, Gaussian, ...).

Processus de Dirichlet

- Distribution de Dirichlet : $\vec{X} \sim Dir(\alpha)$ t.q. $\sum_k X_k = 1$
- Souvent utilisée comme prior dans les modèles de clustering Bayésiens
  - Typiquement X_k est la probabilité d'appartenir au cluster k

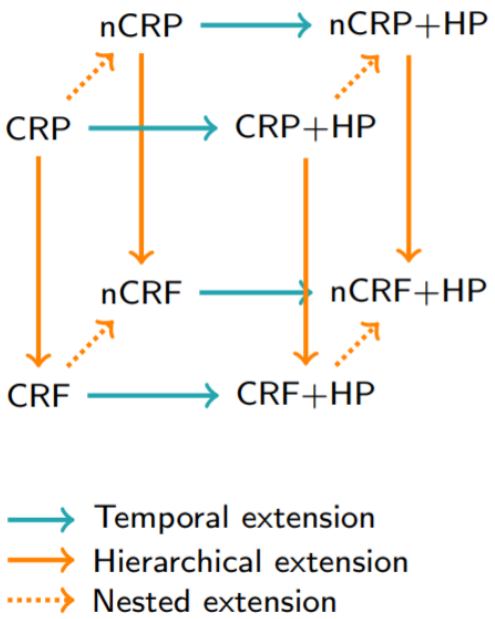


Variants

- D'autres priors basés sur Dirichlet existent :
  - Uniform process (Wallach et al., 2010)
  - Pitman-Yor process (Pitman and Yor, 1997)
  - Hierarchical Dirichlet process (Teh et al., 2006)
  - Nested Dirichlet process (Rodríguez et al., 2008)

→ La plupart ont la propriété "les riches s'enrichissent"

→ Tous considèrent des comptes entiers



Point processes

- Temps souvent "modélisé" en échantillonnant les observations (DTM (Blei and Lafferty, 2006), RCRP (Ahmed and Xing, 2008; Diao and Jiang, 2014), DDCRP (Blei and Frazier, 2010) etc.)
  - Problèmes : comment découper les données, quelle fonction d'échantillonnage utiliser, poids des observations, etc.
- Modéliser le temps explicitement : processus ponctuels

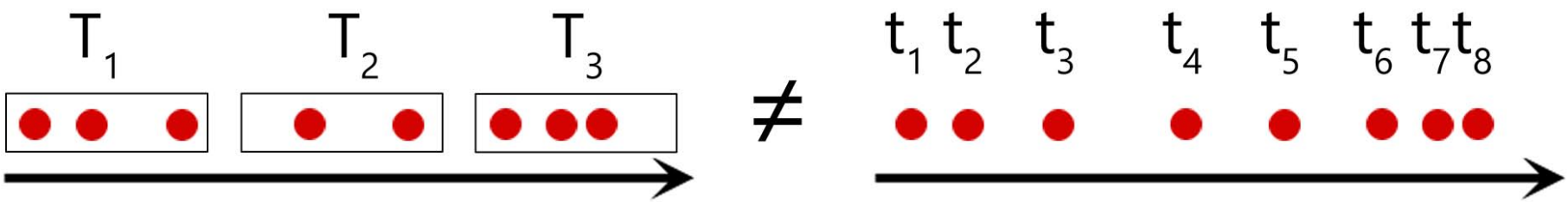


Figure 24 – L'échantillonnage induit des approximations

Processus de Poisson

- Les processus de Poisson sont caractérisés par une intensité λ.
  - $\lambda \Delta t \overset{\Delta t \to 0}{=} P(\mathbb{N}(t + \Delta t) - \mathbb{N}(t) = 1)$
  - Probabilité instantanée d'*un* événement

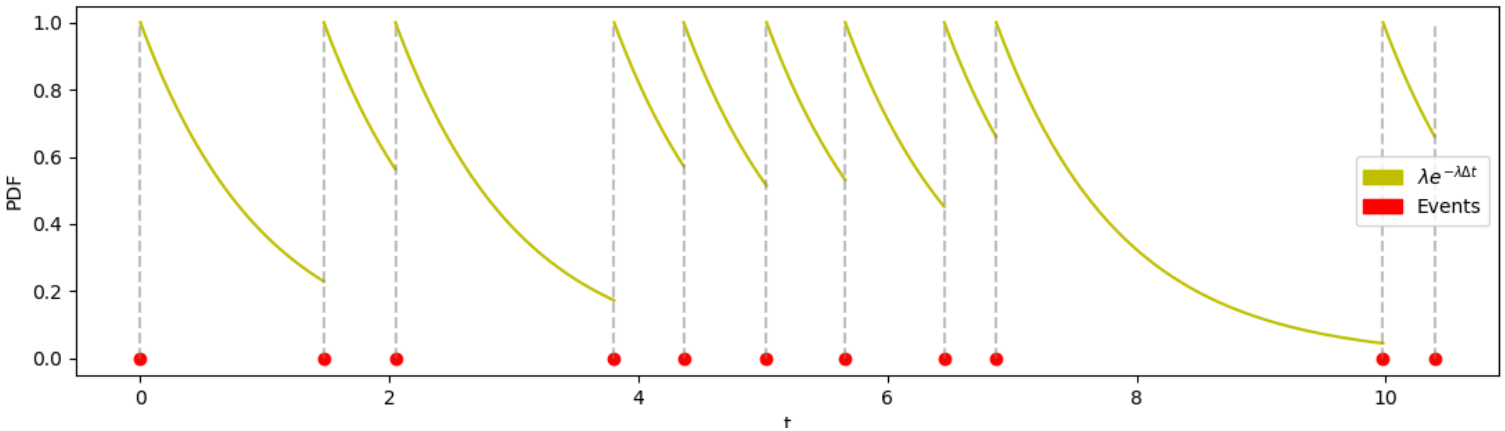


Figure 25 – Pourrait modéliser la désintégration radioactive d'atomes avec une demi-vie de 1

Processus de Poisson non-homogène

- $\lambda(t)$ est fonction du temps
- $\lambda(t)\Delta t \overset{\Delta t \to 0}{=} P(\mathbb{N}(t+\Delta t) - \mathbb{N}(t) = 1)$

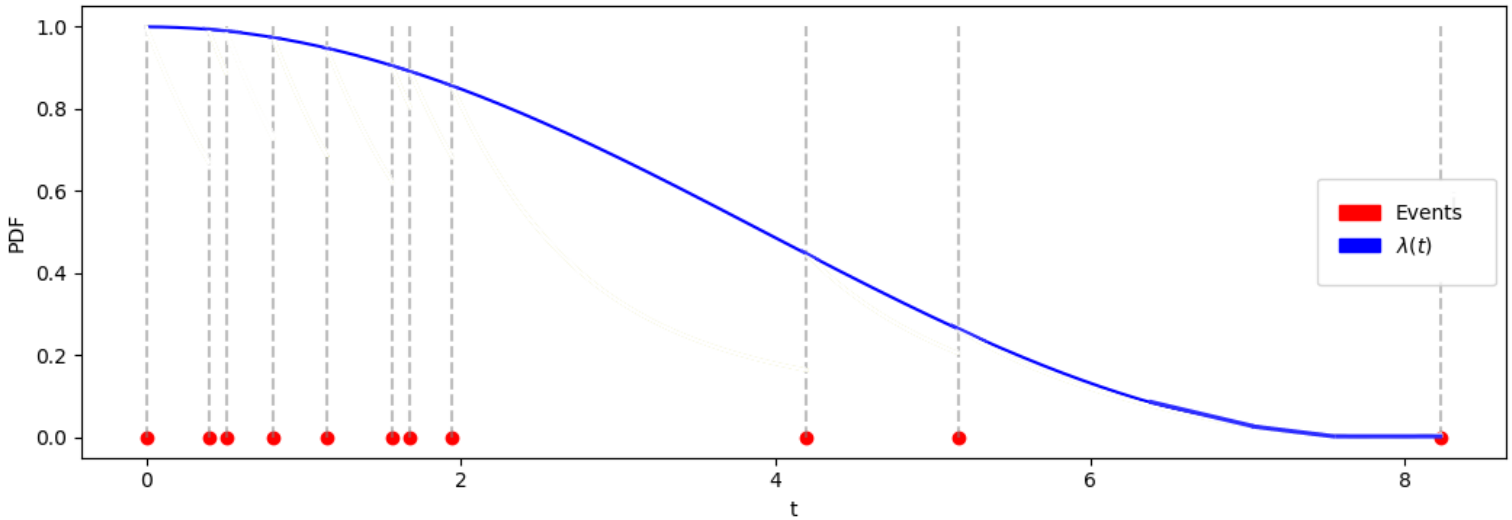


Figure 26 – Pourrait modéliser le passage de voitures sur une route au cours de la journée

Implications de PDP

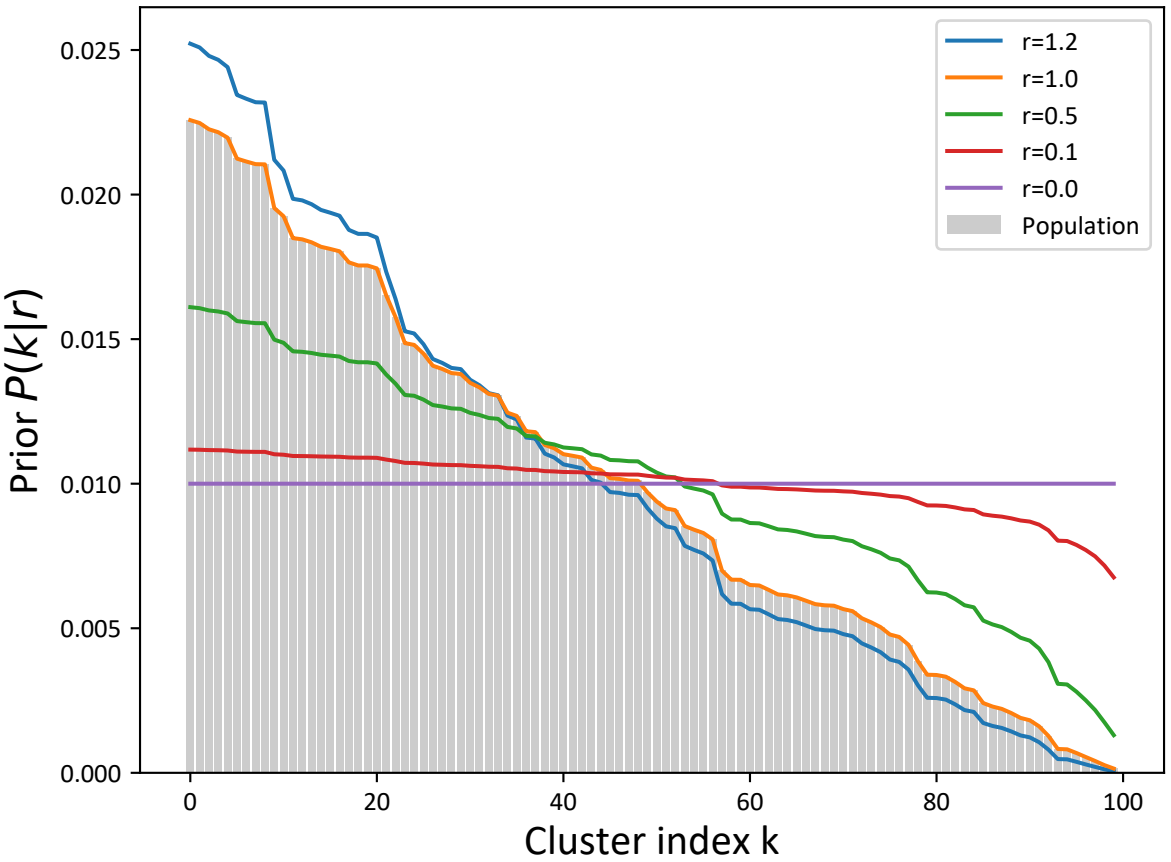
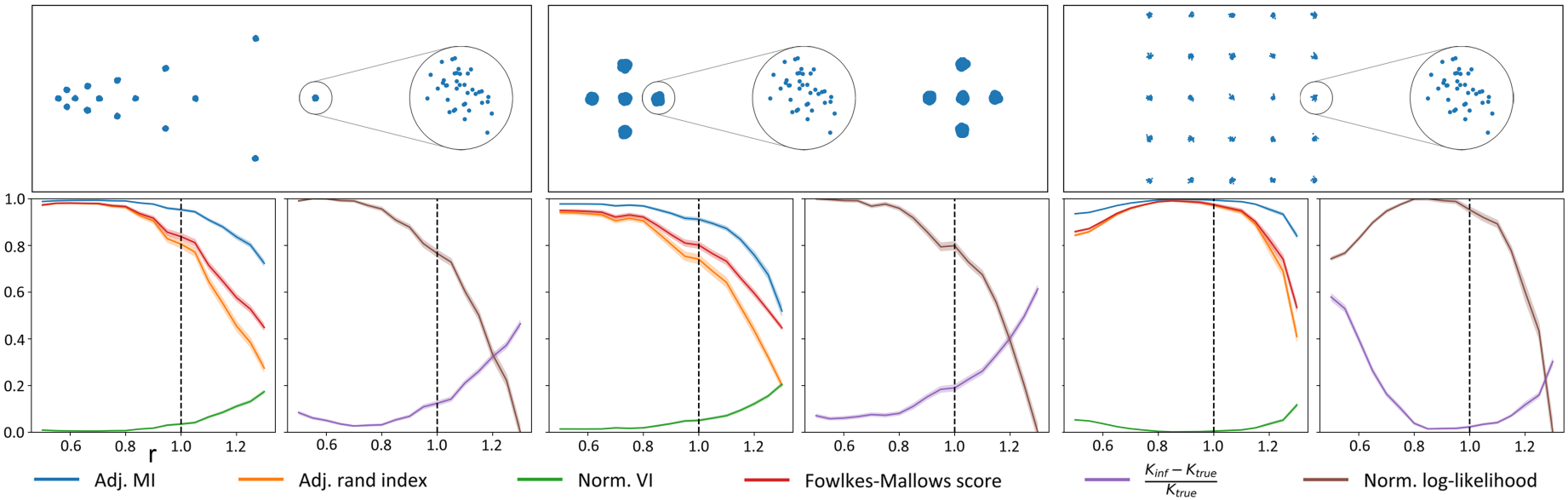


Figure 27 – Probabilité a priori d'appartenir à chacun des 100 clusters représentés ayant une population donnée (barres grises) en fonction de r

Résultats

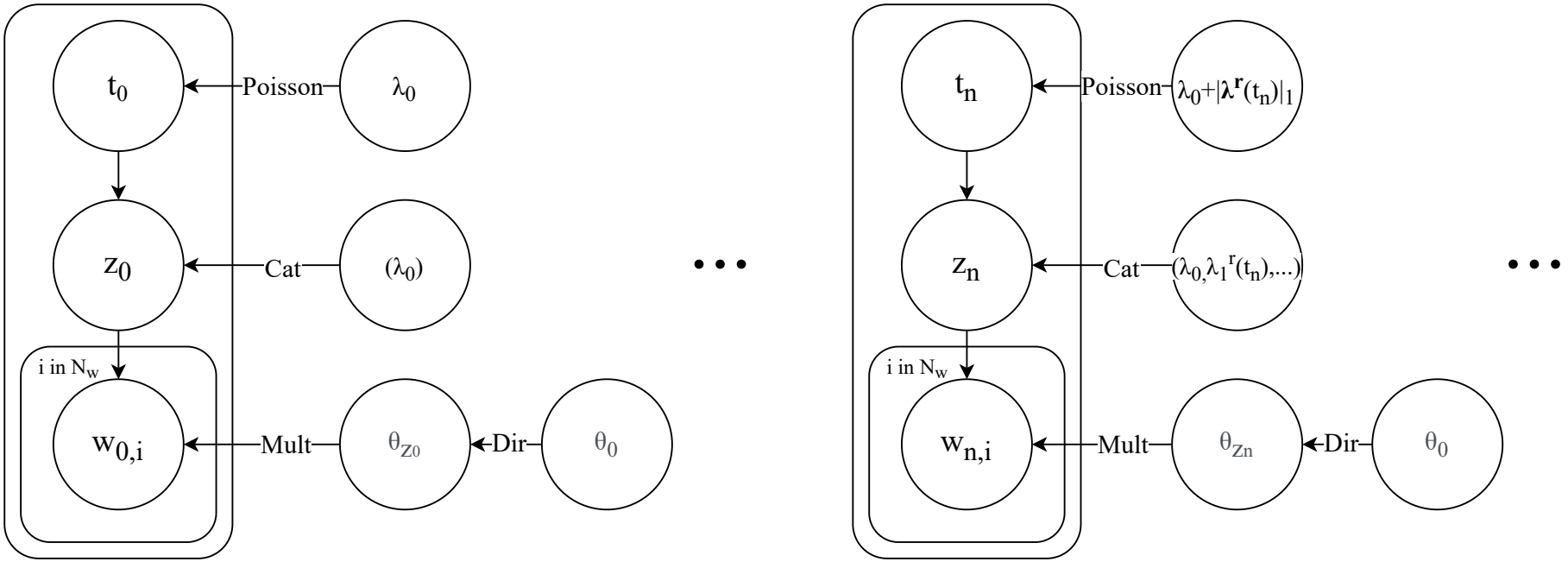
- Utilisation comme prior pour IGMM
- DP n'est pas toujours le meilleur prior



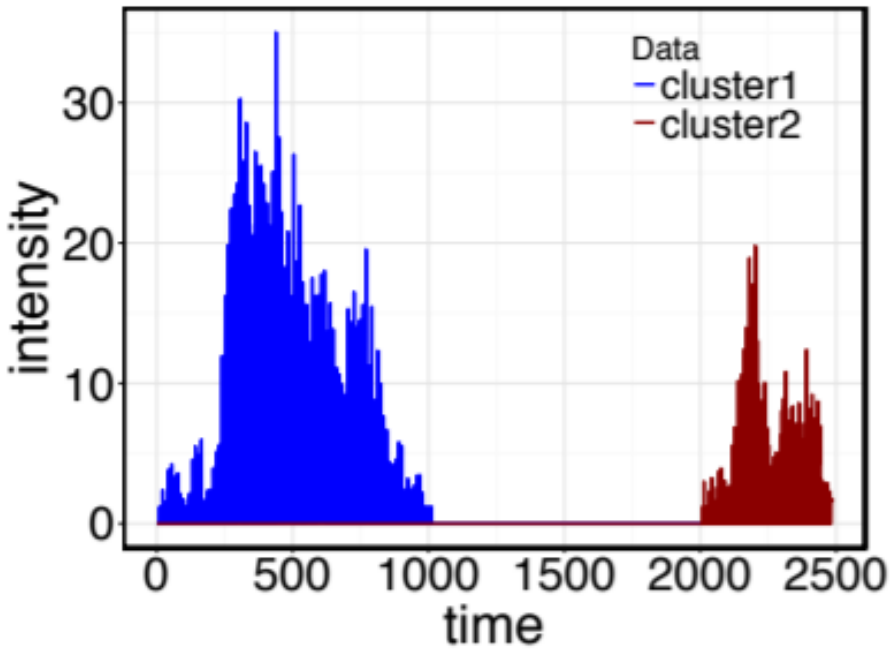
Modèle de langue

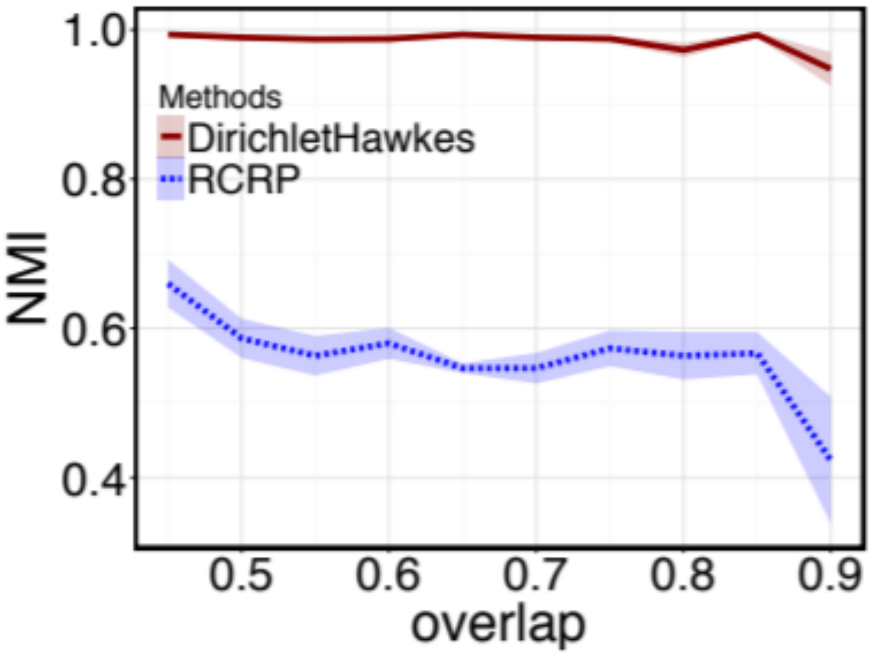
$$
\begin{aligned}
\mathcal{L}(C_i = c | N_{<i,c}, n_i, \theta_0) &= P(n_i | C_i = c, N_{<i,c}, \theta_0) \\
&= \frac{\mathcal{L}_{txt}(\vec{C}_{<i,c} | N_{<i,c}, \theta_0)}{\mathcal{L}_{txt}(\vec{C}_{<i-1,c} | N_{<i,c}, \theta_0)} \\
&= \frac{\cancel{\frac{\Gamma(\theta_0)}{\Gamma(N_c+n_i+\theta_0)}} \prod_v \frac{\Gamma(N_{c,v}+n_{i,v}+\theta_{0,v})}{\cancel{\Gamma(\theta_{0,v})}}}{\cancel{\frac{\Gamma(\theta_0)}{\Gamma(N_c+\theta_0)}} \prod_v \frac{\Gamma(N_{c,v}+\theta_{0,v})}{\cancel{\Gamma(\theta_{0,v})}}} \qquad (1) \\
&= \frac{\Gamma(N_c + \theta_0)}{\Gamma(N_c + n_i + \theta_0)} \prod_v \frac{\Gamma(N_{c,v} + n_{i,v} + \theta_{0,v})}{\Gamma(N_{c,v} + \theta_0)}
\end{aligned}
$$

Représentation graphique DHP



Performances (clusters distincts)

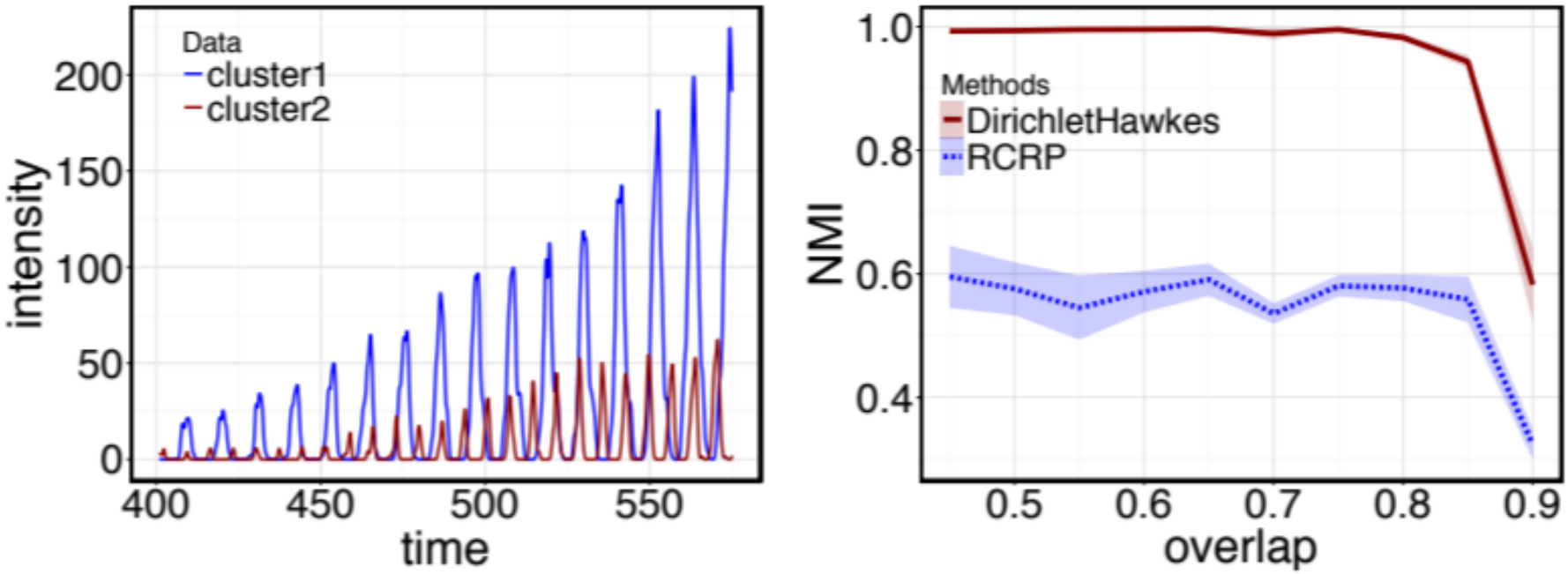




(a) Temporally well-separated clusters.

Figure 28 – (Du et al., 2015)

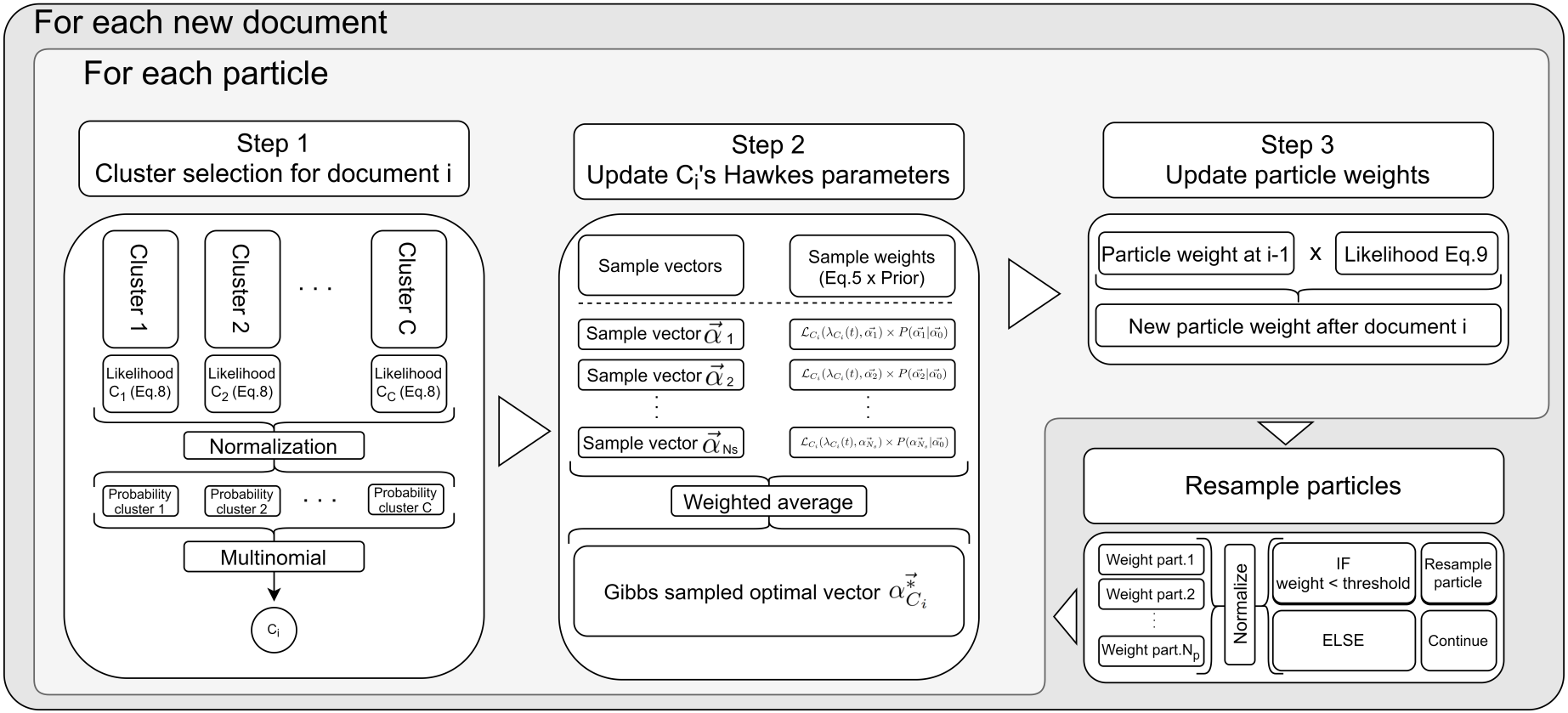
Performances (clusters "non" distincts)



(b) Temporally interleaved clusters.

Figure 29 – (Du et al., 2015)

Inférence (résumée)



Recouvrements

- Souvent, une des informations disponibles est plus informative que les autres :
  - Twitter : textes courts (peu d'information textuelle) mais dynamiques de retweet informatives (information temporelle pertinente)
- Arrive souvent à cause de recouvrements entre informations disponibles :

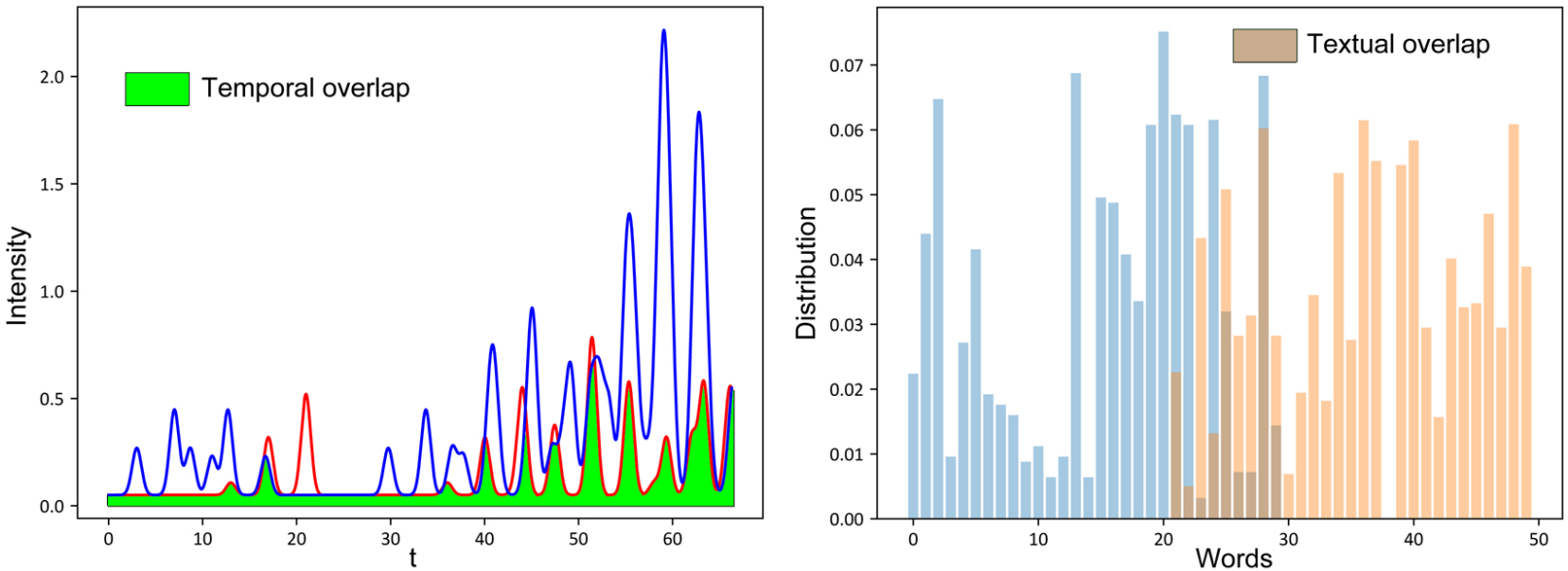


Figure 30 – (Poux-Médard *et al.*, ICDM 2021)

Reddit r/news - Typical output

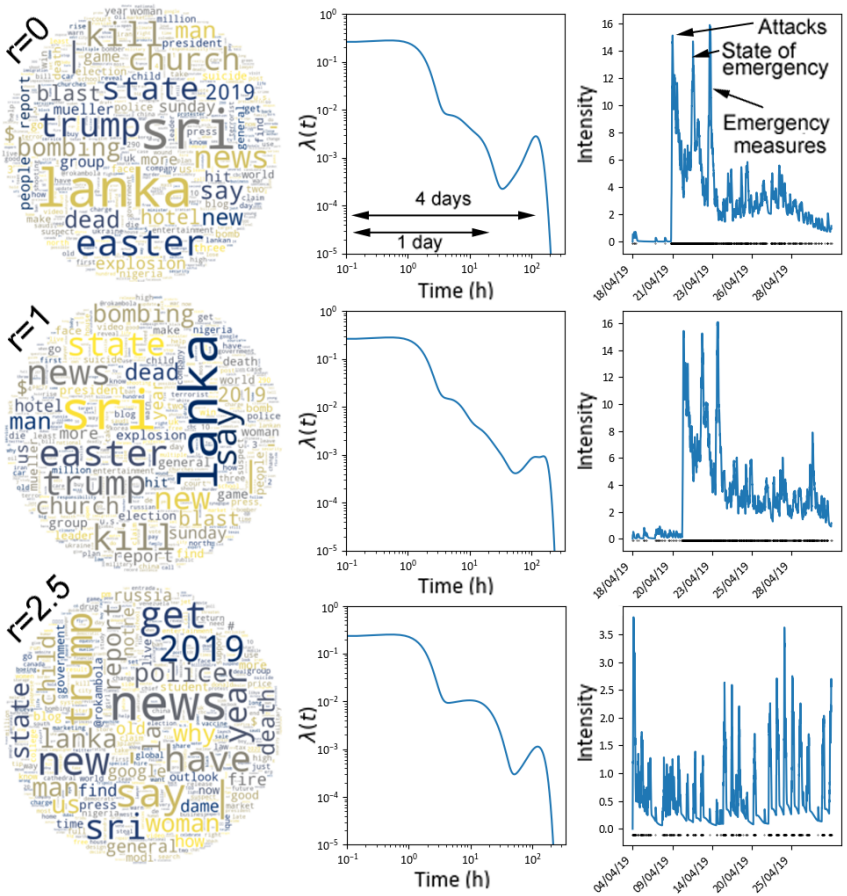
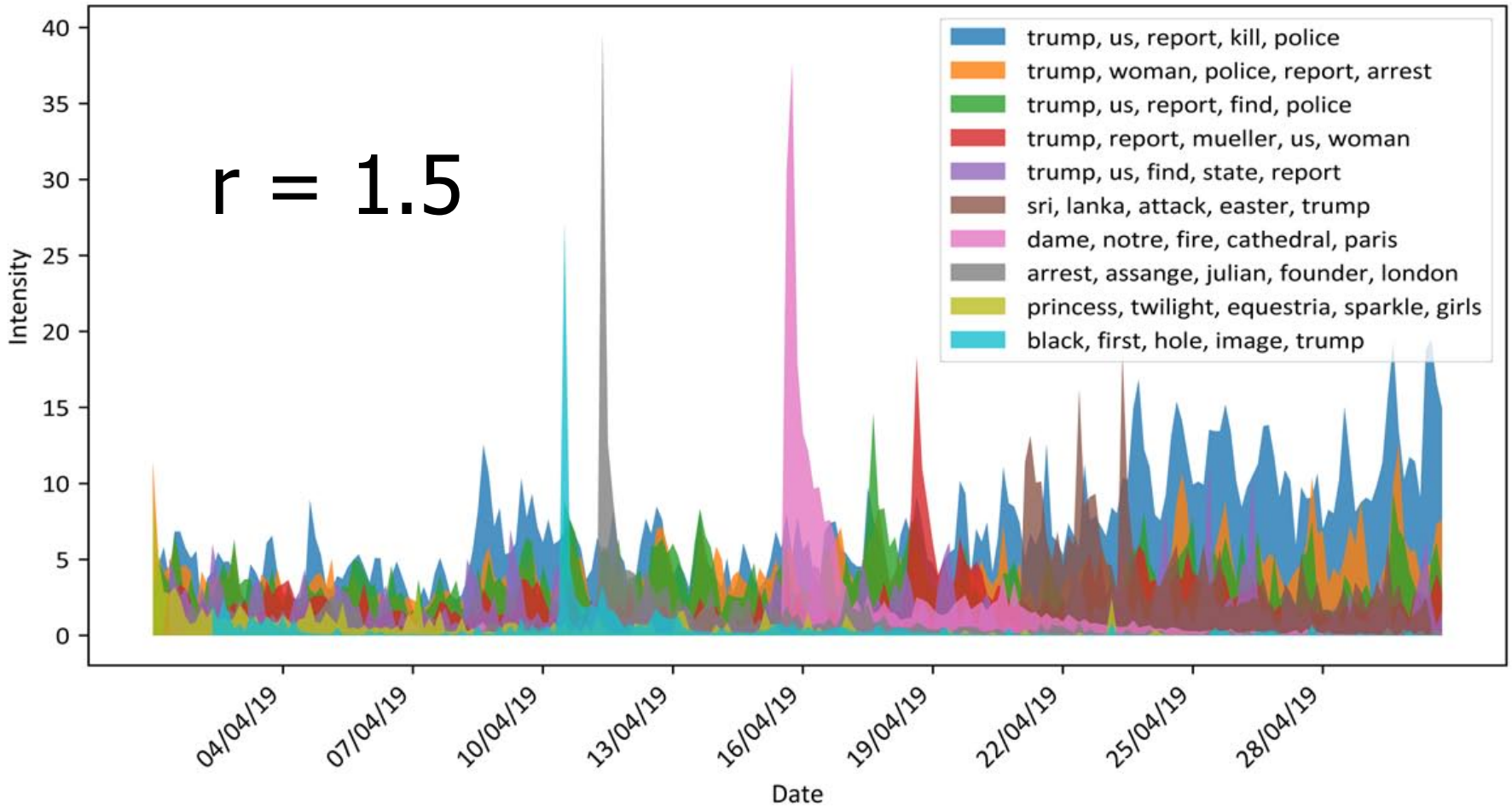


Figure 31 – (Poux-Médard *et al.*, ICDM 2021)

- Données réelles : r/news
- Clusters et leurs dynamiques associées pour plusieurs r
  - Small r : vocabulaire plus restreint
  - Large r : dynamiques plus déterministes

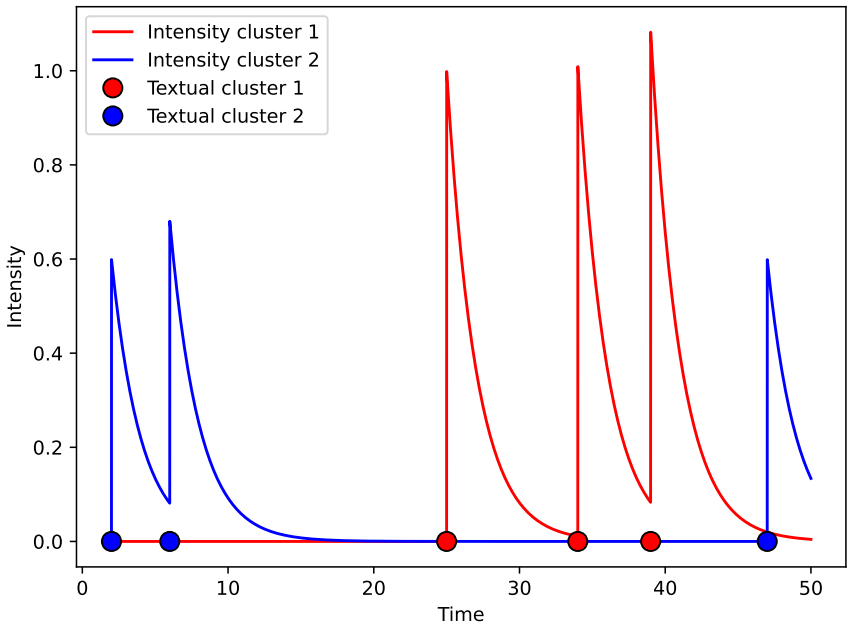
Summary generation

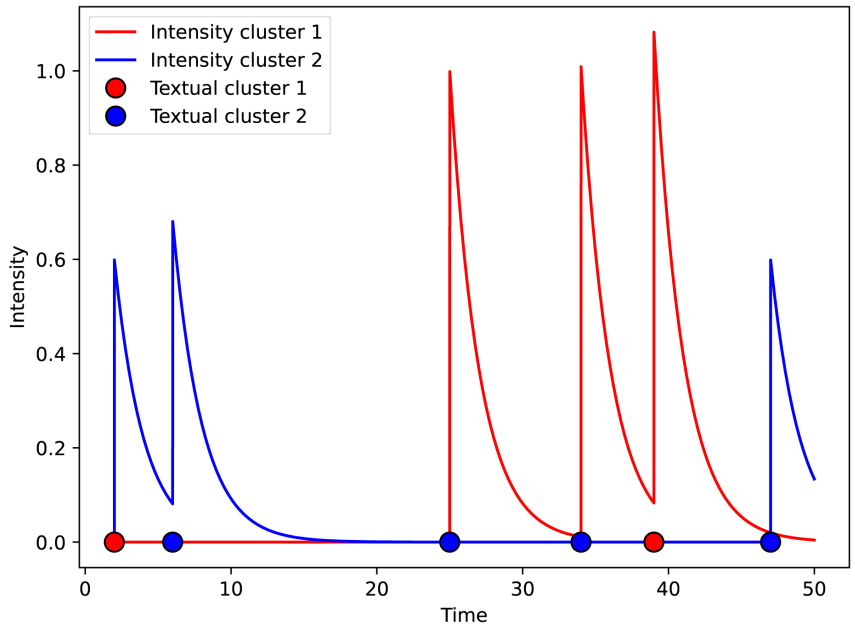
- Powered Dirichlet-Hawkes Process : résumés à partir de flux de données utilisant les interactions temporelles



Décorrelations

- Décorrelations :
  - Ex : un journal influent qui publie un article n'entraîne pas les mêmes dynamiques de réplication qu'un journal moins influent publiant le même article.





Résultats pour plusieurs décorrélations

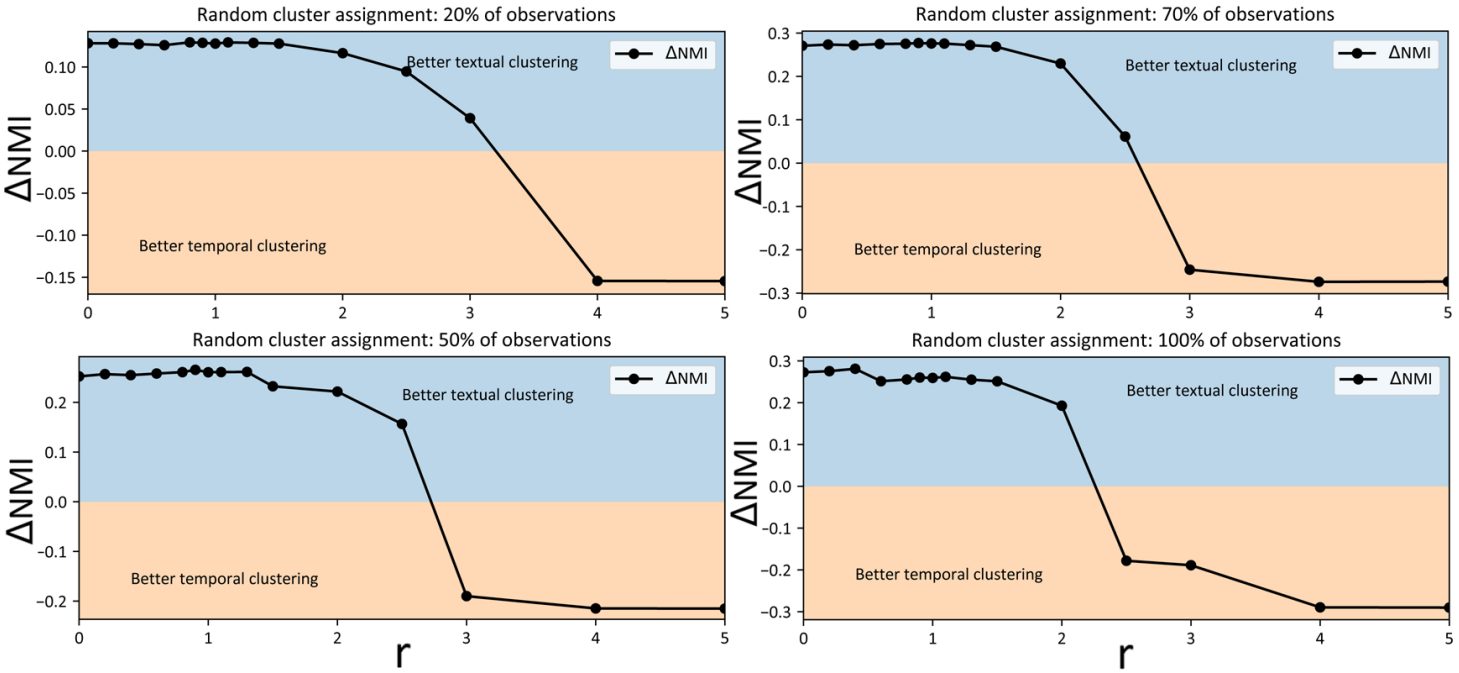


Figure 32 – (Poux-Médard *et al.*, ICDM 2021)

- PDHP retrouve soit les clusters temporels, soit les clusters textuels
  - Small r : bon clusters textuels
  - Large r : bon clusters temporels

Reddit r/news, r/TodayILearned, r/AskScience - Métriques

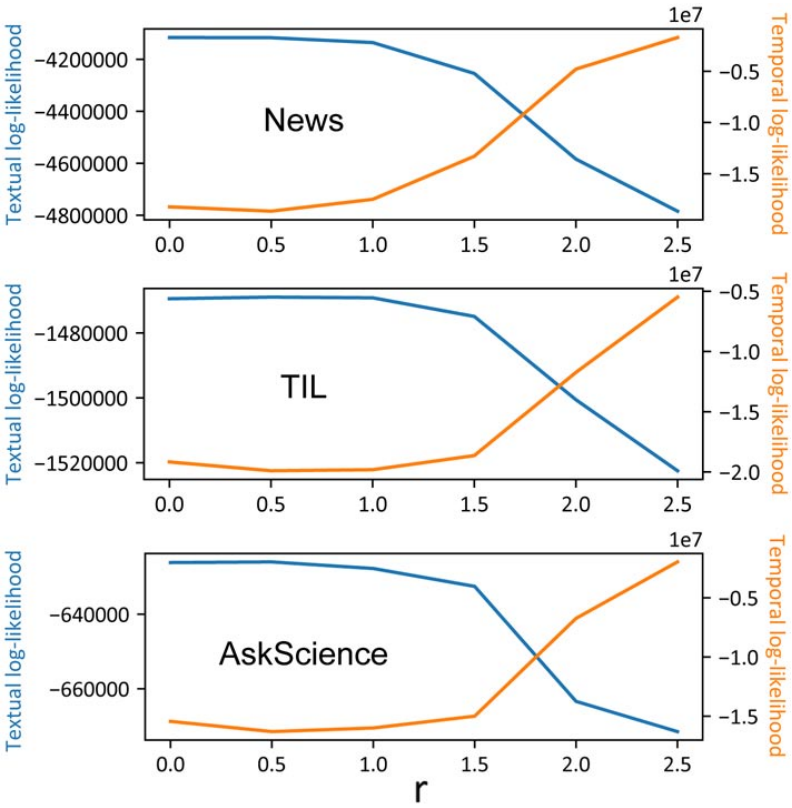


Figure 33 – Vraisemblance textuelle et temporelle vs r (Poux-Médard *et al.*, ICDM 2021)

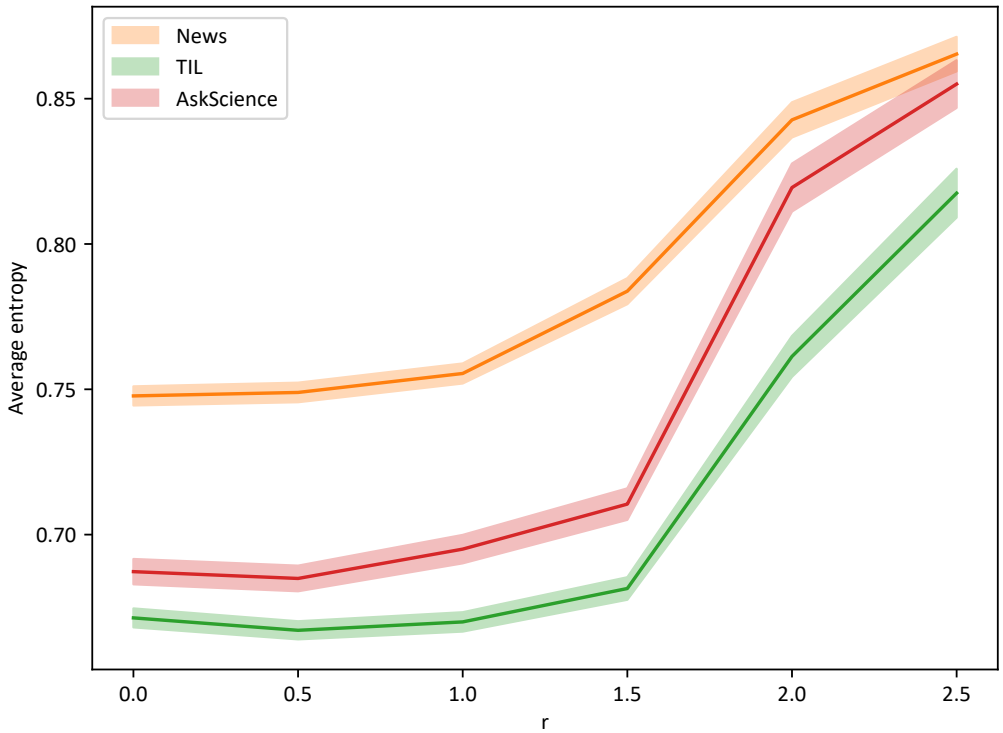


Figure 34 – Entropie des clusters textuels : clusters textuels plus focalisés pour les petits r (Poux-Médard *et al.*, ICDM 2021)

Génération de résumés

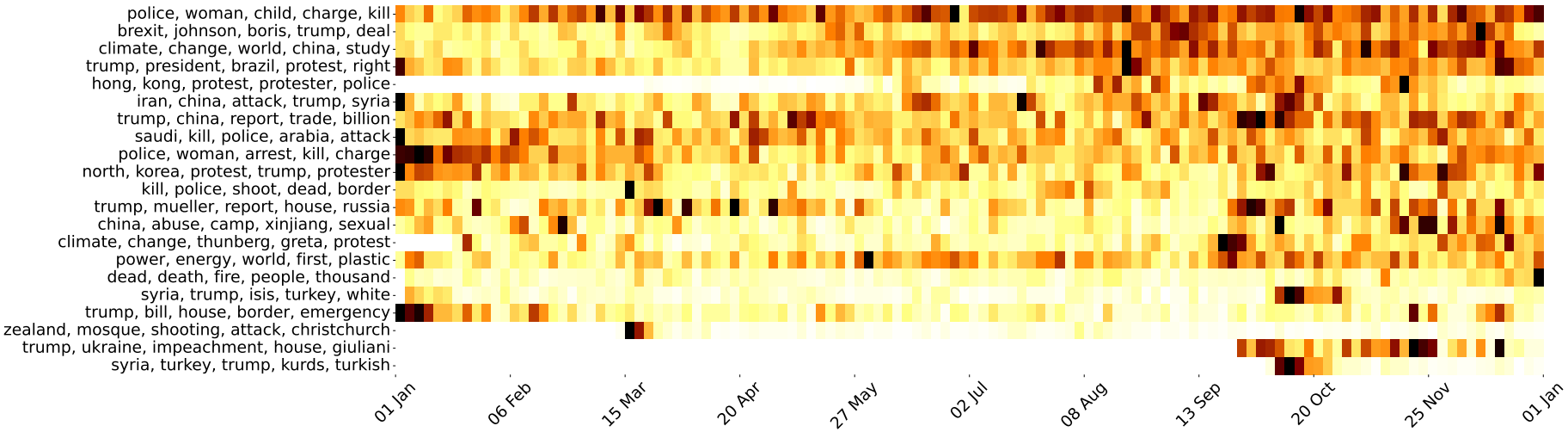


Figure 35 – Échantillon des sujets inférés avec MPDHP sur l'axe des temps réels (01/2019-12/2019)

Perspectives - Prior structurel et temporel

- Plusieurs travaux d'inférence de réseaux via l'analyse de survie :
  - NetRate (Gomez-Rodriguez et al., 2011)
  - KernelCascade (Du et al., 2012)
  - MoNet (Wang et al., 2012)
  - InfoPath (Rodriguez et al., 2013)
  - TopicCascade (Du et al., 2013)
- Tous des cas particuliers de (Gomez-Rodriguez et al., 2013)
  - Formule chacun des modèles précédents comme un **processus de comptage**
  - Créée un pont entre les processus du point et l'inférence de réseaux

Tâche

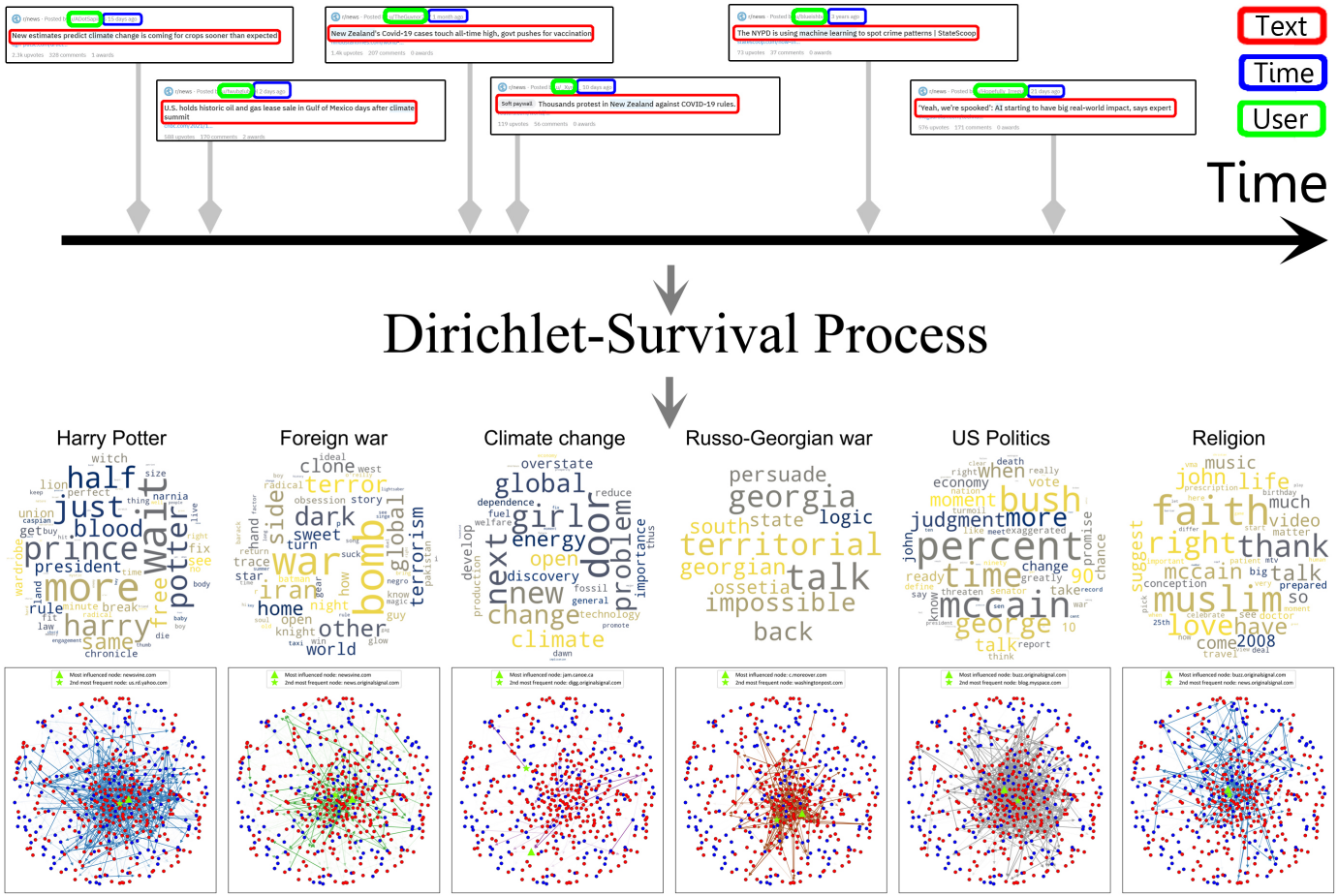
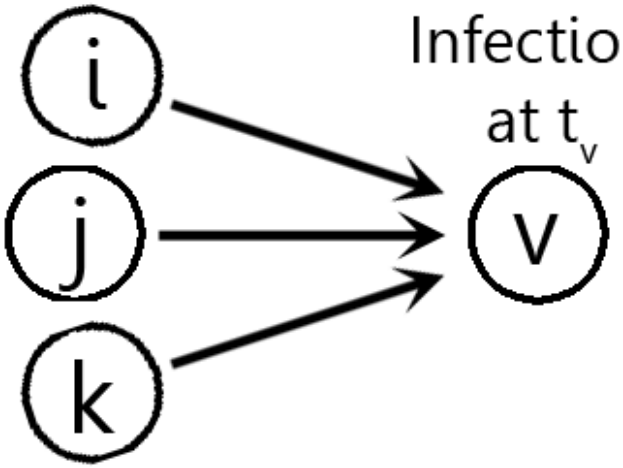
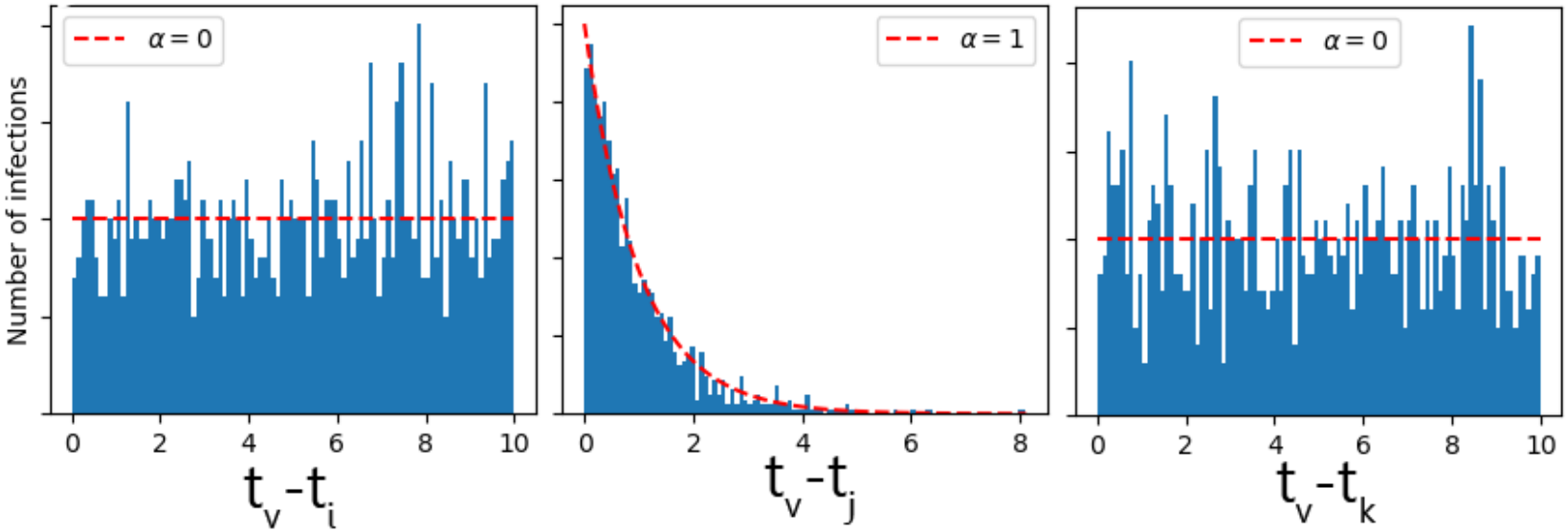


Figure 36 – Processus de Dirichlet-Survival appliqué au jeu de données Memetracker

Inférence de réseau



Exponential model $P(t)=\alpha.e^{-\alpha t}$



Processus ponctuel

- La littérature sur l'inférence des réseaux s'inscrit naturellement dans la littérature des processus ponctuels
  → On peut dériver un a priori Bayésien temporel *et* structurel

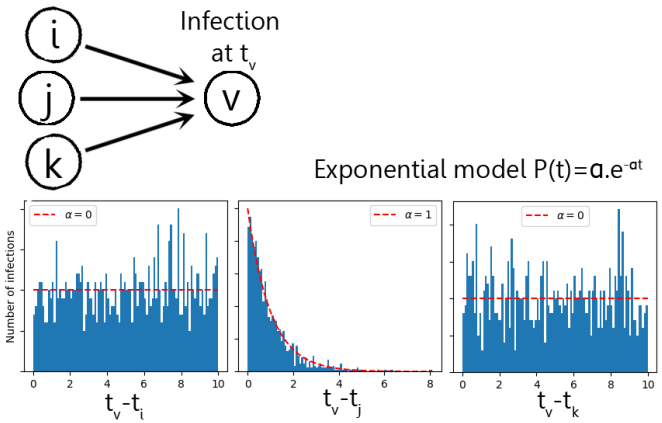


Figure 37 – Survival process

Processus ponctuels
<≈>

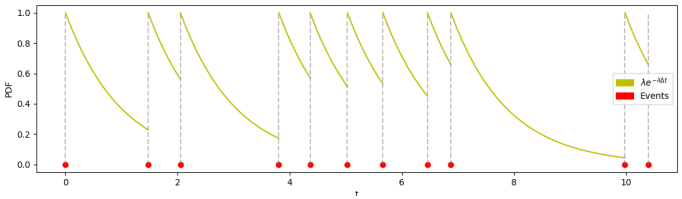


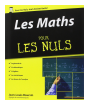
Figure 38 – Hawkes process

Dirichlet-Survival Process

- A priori sur l'appartenance à un cluster C_i pour l'observation i publiée par noeud u au temps t étant donné l'historique $\mathcal{H}$ et les réseaux dépendant des clusters A :

$$P(C_i = k | u, t, \mathcal{H}, A)$$

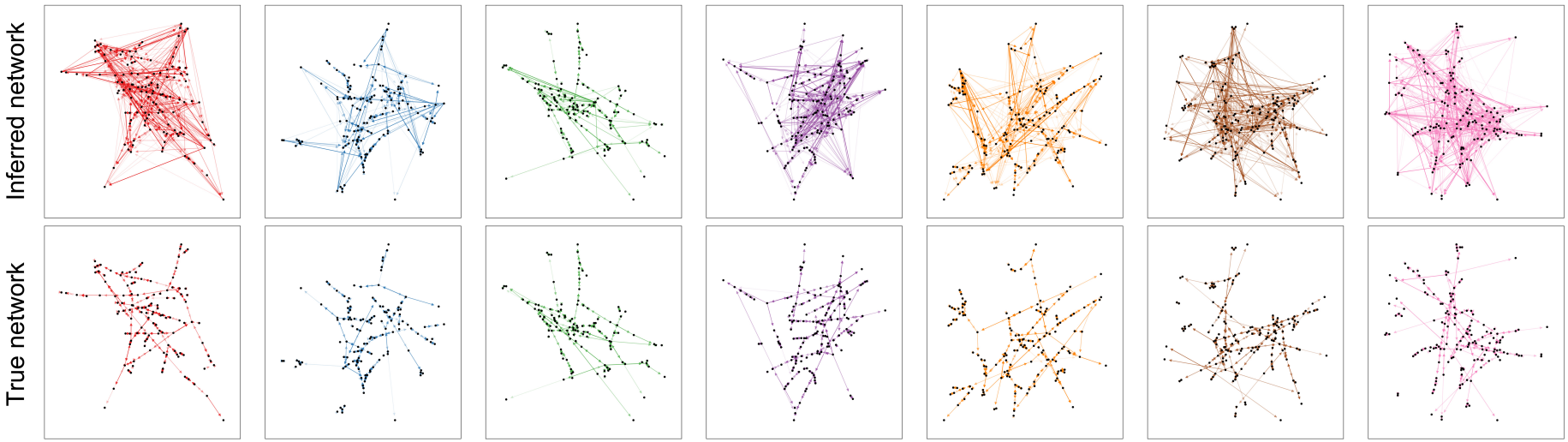
$$= \begin{cases} \frac{\lambda_0^{(k)} + \sum_{\mathcal{H}_{i,c}^{(k)}} H(t_i^c | t_j^c, \alpha_{u_j^c, u_i^c}^{(k)})}{\lambda_0^{(K+1)} + \sum_k^K \lambda_0^{(k)} + \sum_{\mathcal{H}_{i,c}^{(k)}} H(t_i^c | t_j^c, \alpha_{u_j^c, u_i^c}^{(k)})} & \text{if k = 1, ..., K} \\ \frac{\lambda_0^{(K+1)}}{\lambda_0^{(K+1)} + \sum_k^K \lambda_0^{(k)} + \sum_{\mathcal{H}_{i,c}^{(k)}} H(t_i^c | t_j^c, \alpha_{u_j^c, u_i^c}^{(k)})} & \text{if k = K+1} \end{cases}$$



$$= \begin{cases} \frac{\text{Strength of incoming edges of cluster/subnetwork k at time t}}{\text{Normalizing term}} & \text{if k = 1, ..., K} \\ \frac{\text{Probability of a new cluster/subnetwork k+1 at time t}}{\text{Normalizing term}} & \text{if k = K+1} \end{cases}$$

Résultats – Synthétiques

- On simule la diffusion de documents tirés d'un parmi 5 clusters possibles, chacun ayant son propre vocabulaire et sous-réseau de diffusion



Résultats numériques

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