Résumé des travaux de thèse

Gaël Poux-Médard

October 4, 2021

Mon sujet

Intitulé

Rôle de l'interaction entre informations dans les processus de diffusion

L'information

Une information peut être :

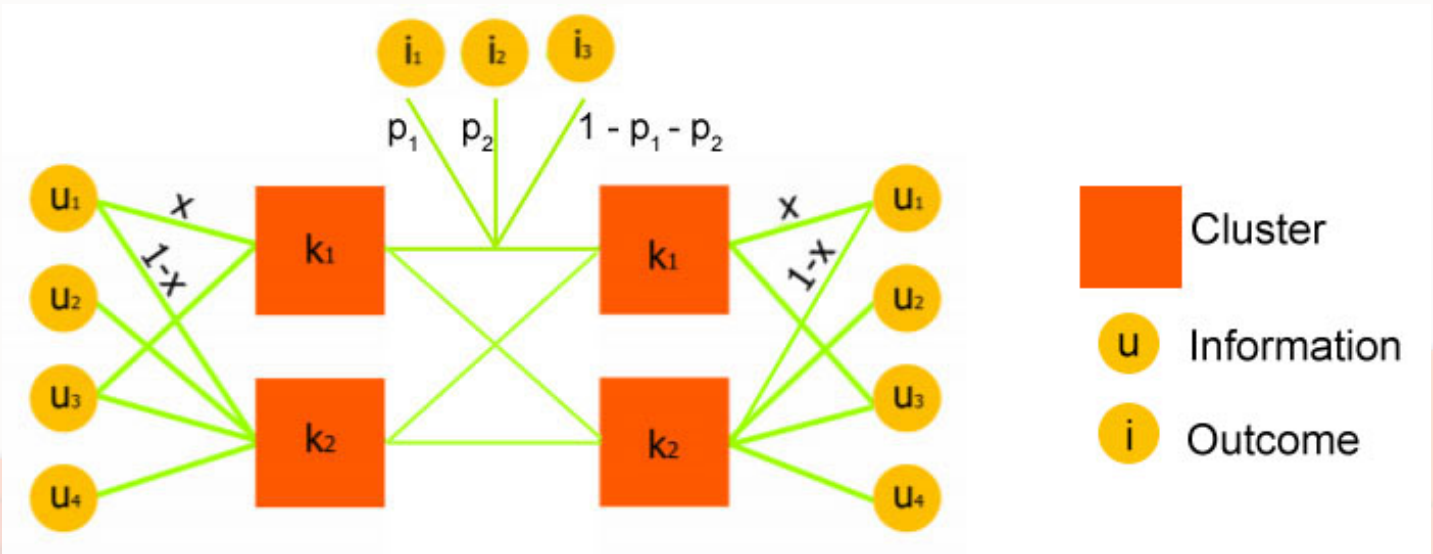
- Un hashtag (comme sur Twitter)
- Un mème (comme sur Internet)
- Un topic (comme sur Reddit)
- Une image (comme sur Instagram)
- Une situation (comme dans un dilemme du prisonnier itéré)
- Une pub (comme à la télé)
- ...

Sommaire

IMMSBM

IMMSBM

- Interactions de paires modélisées par un MMSBM appliqué à un graphe biparti symétrique.
- Si un individu est exposé à une paire d'informations (A,B), le modèle infère p(décision sur A | (A,B))
- Le modèle impose p(décision sur A | (A,B))=p(décision sur A | (B,A))
- p(décision sur X | (A,A))=viralité intrinsèque de A vàv de X
- Inférence via un algorithme EM.



Résultats

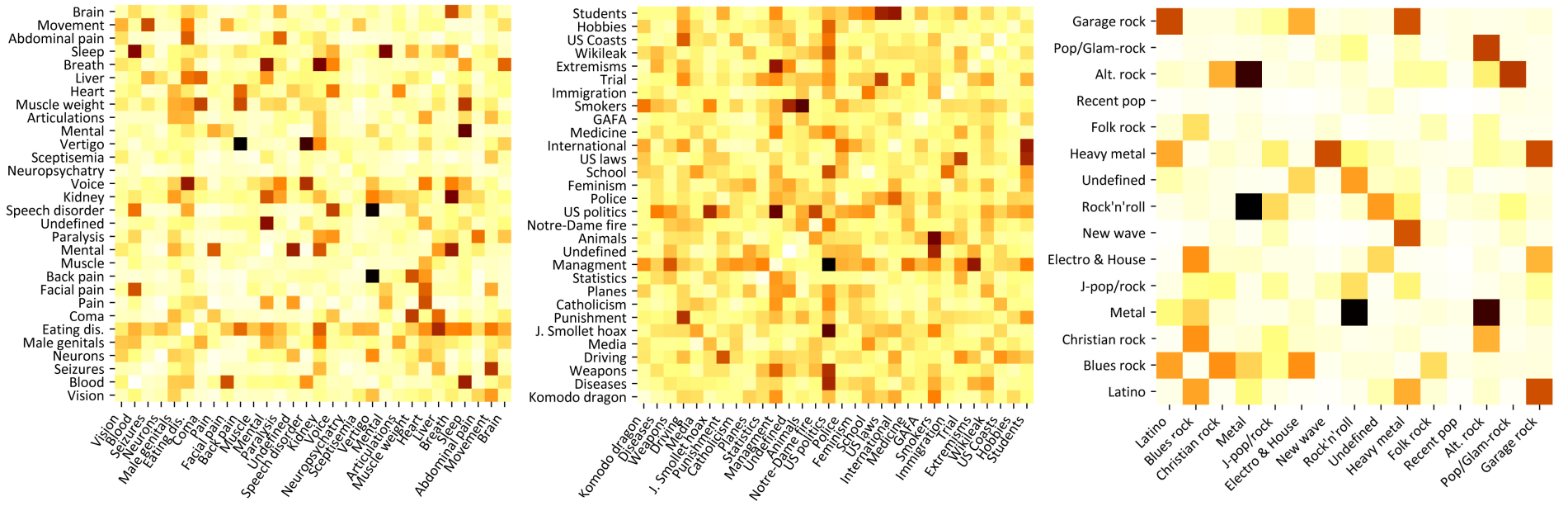


Figure 1: Conclusion : les interactions significatives entre clusters d'informations sont rares.

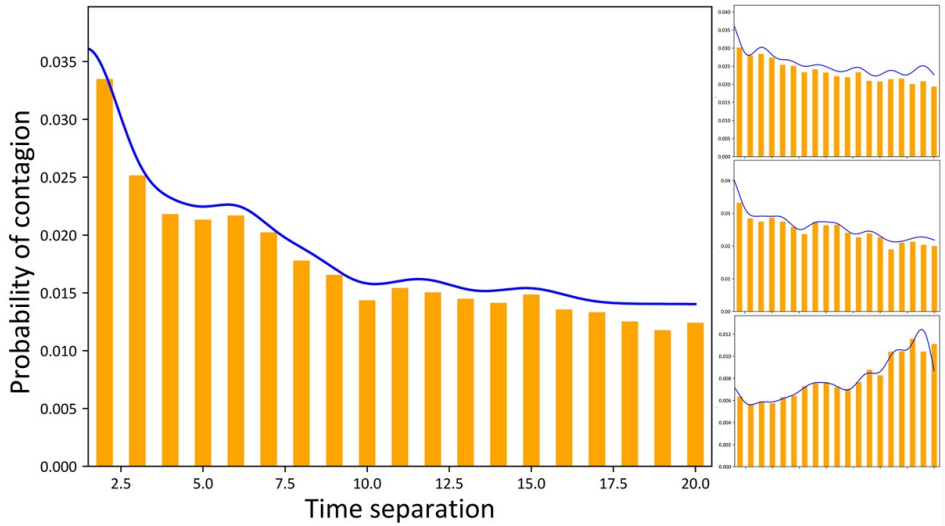
Papier possiblement accepté à RecSys 🤞

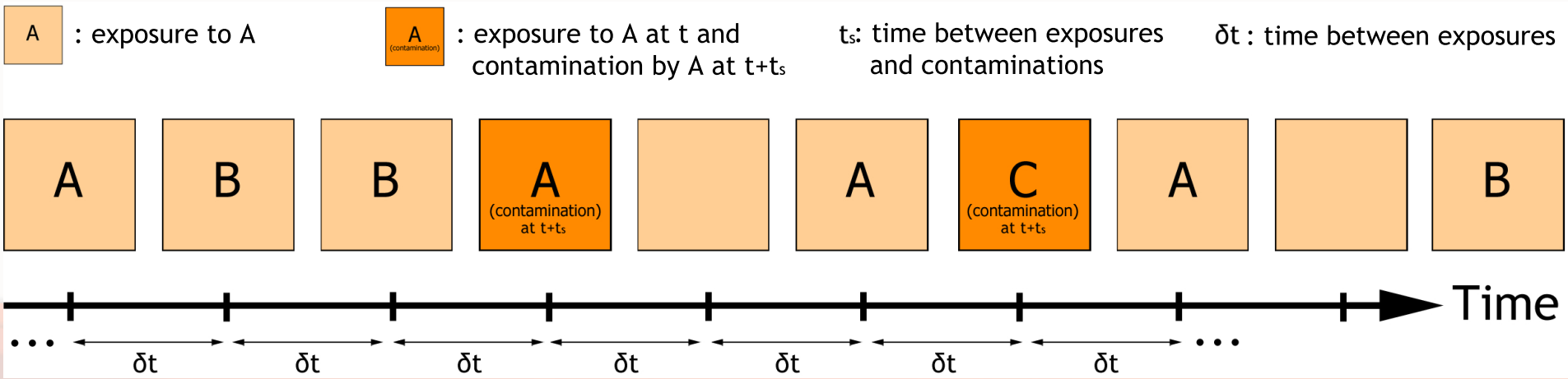
Sommaire

InterRate

Aspect temporel des interactions

- Dans quelle mesure les interactions subsistent dans le temps ?
  - Information A à temps t_A et B à temps $t_B > t_A$: comment A interagit avec B en fonction de $\Delta t = t_B - t_A$?
- L'interaction est toujours définie entre paires d'informations.
- Développement d'un modèle en temps continu, convexe et parallélisable.





InterRate : le modèle en équations

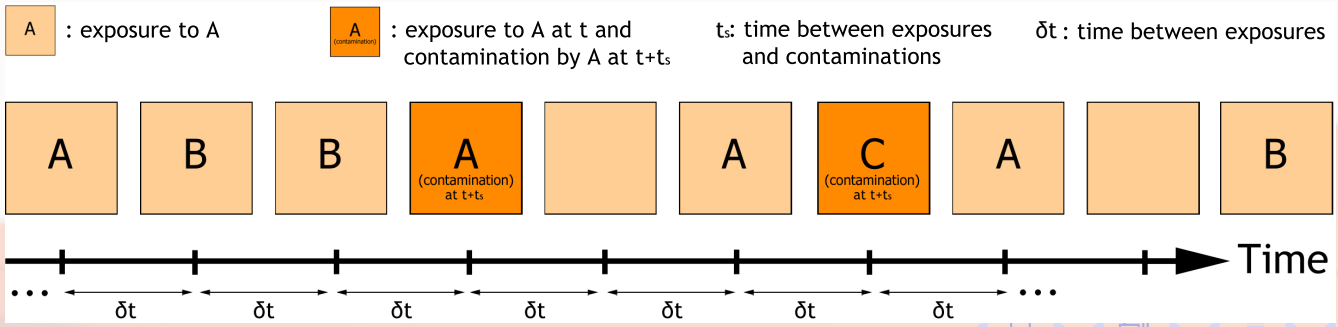
Likelihood

$$\ell(\beta|\mathcal{D}, t_s) = \sum_{\mathcal{D}} \sum_{t_j^{(y)} \in \mathcal{H}_i^{(x)}} c_{t_i}^{(x)} \log\left(H(t_i^{(x)} + t_s | t_j^{(y)}, \beta_{xy})\right) + (1 - c_{t_j}^{(y)}) \log\left(1 - H(t_i^{(x)} + t_s | t_j^{(y)}, \beta_{xy})\right)$$

Kernel functions

RBF : $\log H(t_i^{(x)} + t_s | t_j^{(y)}, \beta_{ij}) = -\beta_{ij}^{(bg)} - \sum_{s=0}^{S} \frac{\beta_{ij}^{(s)}}{2} (t_i + t_s - t_j - s)^2$

EXP : $\log H(t_i^{(x)} + t_s | t_j^{(y)}, \beta_{xy}) = -\beta_{ij}^{(bg)} - \beta_{ij}(t_i + t_s - t_j)$



Conclusions

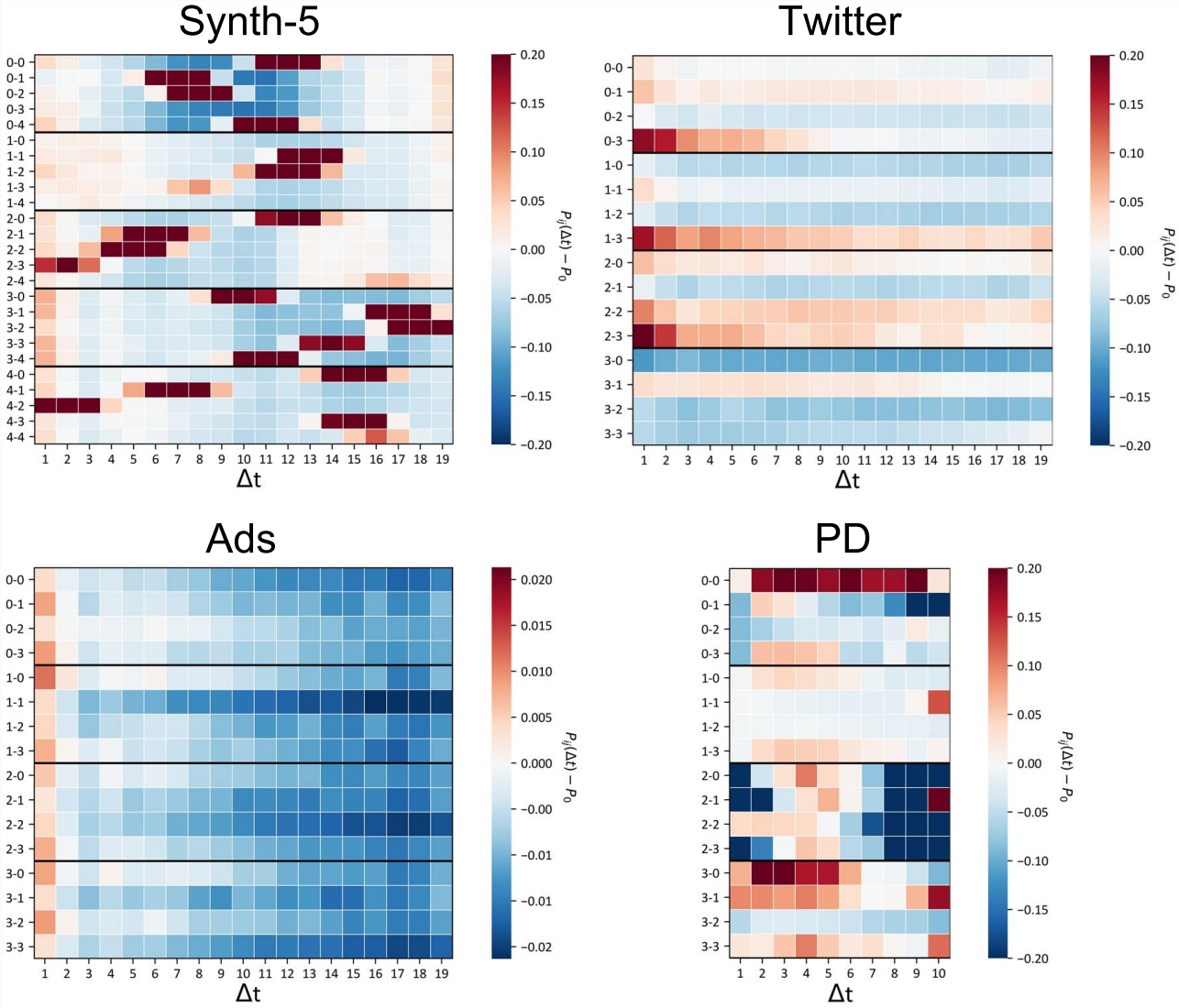


Figure 2: Les interactions entre informations sont brèves. Papier accepté à ECML-PKDD.

Synthèse

Synthèse

- Les interactions prennent place entre des groupes d'information spécifiques $\rightarrow$ nécessité de créer des clusters.
- Les interactions entre informations individuelles ne perdurent pas dans le temps $\rightarrow$ nécessité de considérer l'aspect temporel.
- Cependant, elles jouent un rôle non-négligeable dans la plupart des corpus considérés et améliorent significativement la compréhension de corpus provenant du monde réel.

Solution

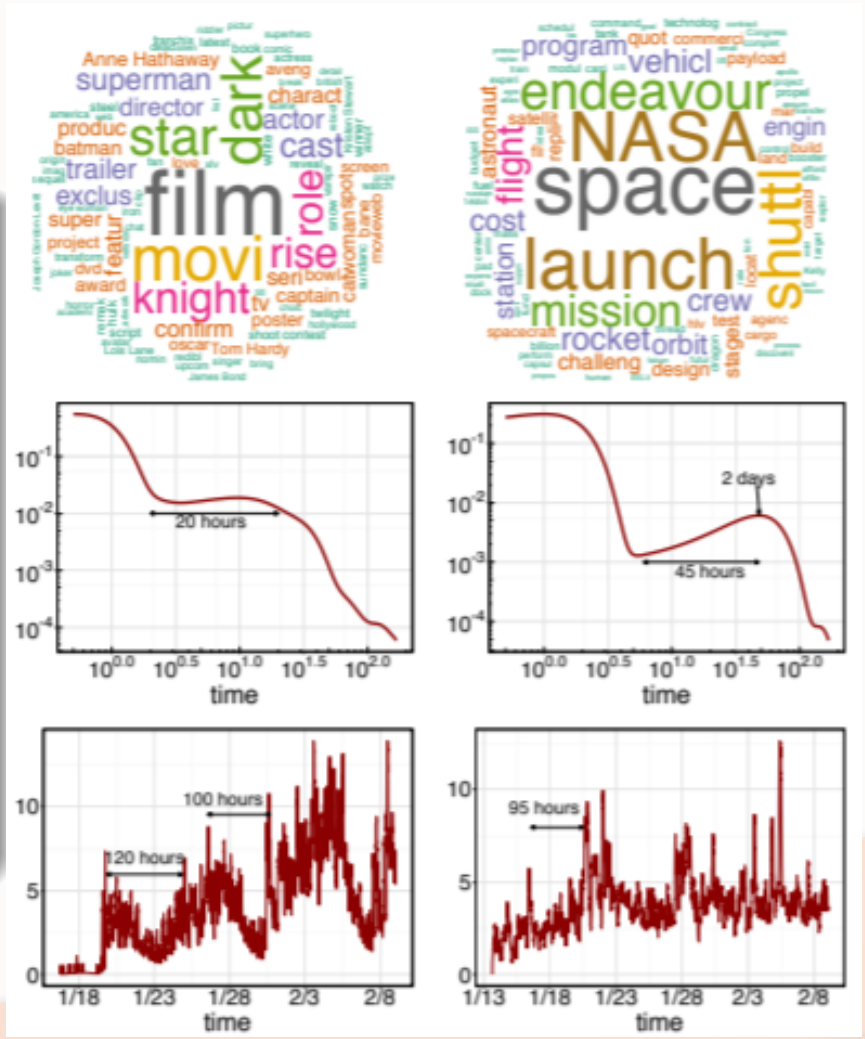
- Nécessité de créer des clusters + nécessité de considérer l'aspect temporel = nécessité de clustering temporel
- Une publication de N. Du à KDD en 2015 fait cela : *Dirichlet-hawkes processes with applications to clustering continuous-time document streams*
- Cette méthode introduit les processus de Hawkes-Dirichlet permettant de créer des clusters de documents se basant à la fois sur leur contenu et leur date.

Sommaire

Powered Dirichlet Process

Hawkes-Dirichlet process

- (N.Du, KDD 2015) : *Dirichlet-hawkes processes with applications to clustering continuous-time document streams*. Le modèle :
- $p(\text{cluster}|date, texte) \propto \underbrace{p(\text{cluster}|texte)}_{\text{Likelihood modèle de langue}} \times \underbrace{p(\text{cluster}|date)}_{\text{Prior temporel (Hawkes-Dir)}}$
- Le prior temporel définit arbitrairement la dépendance temporelle du clustering



Powered Dirichlet Process

Powered Dirichlet process

- Le prior temporel définit arbitrairement la dépendance temporelle du clustering ; nous voulons contrôler cette dépendance en ajoutant un paramètre r au processus de Hawkes-Dirichlet.
- Ajouter ce paramètre revient à définir une version ”puissance” du processus de Dirichlet, ce qui permet de contrôler l'intensité de son hypothèse ”rich-get-richer”.
- Typiquement, $p(\text{n}^{th}$ obs appartient au cluster c$|\vec{N}, \alpha) = \frac{N_c^r}{\alpha + \sum_k N_k^r}$
  - $r = 1$: Dirichlet process ; $r = 0$: Uniform process ; $0 < r < 1$: ”rich-get-less-richer” ; $r > 1$: ”rich-get-more-richer”.
- Ma contribution : mise en évidence du besoin de définir ce processus, dérivation de ce processus, analyse de convergence et du nombre moyen de clusters observés à la n-ième observation.

Résultats

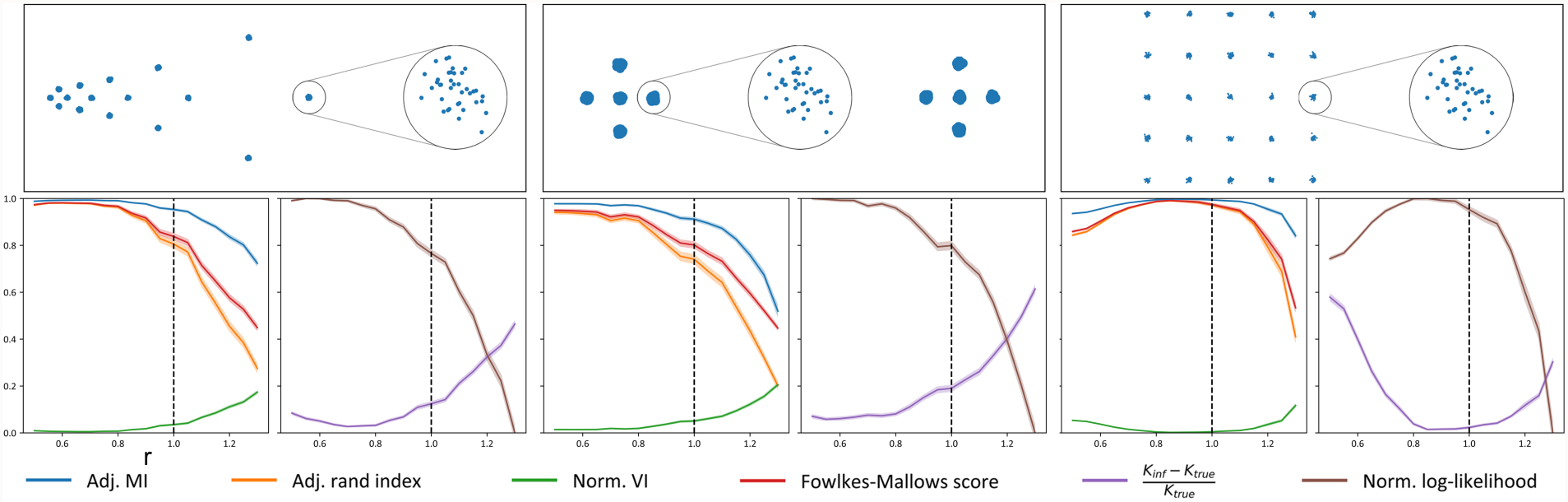


Figure 3: Varier le paramètre r donne de meilleurs résultats de clustering dans plusieurs situations. Papier refusé à ECML-PKDD, mais avec de très bonnes reviews.

Sommaire

PDHP

Powered Hawkes-Dirichlet Process

- On utilise le Powered Dirichlet Process pour définir le PDHP
- Contrôle de l'importance donnée au prior temporel

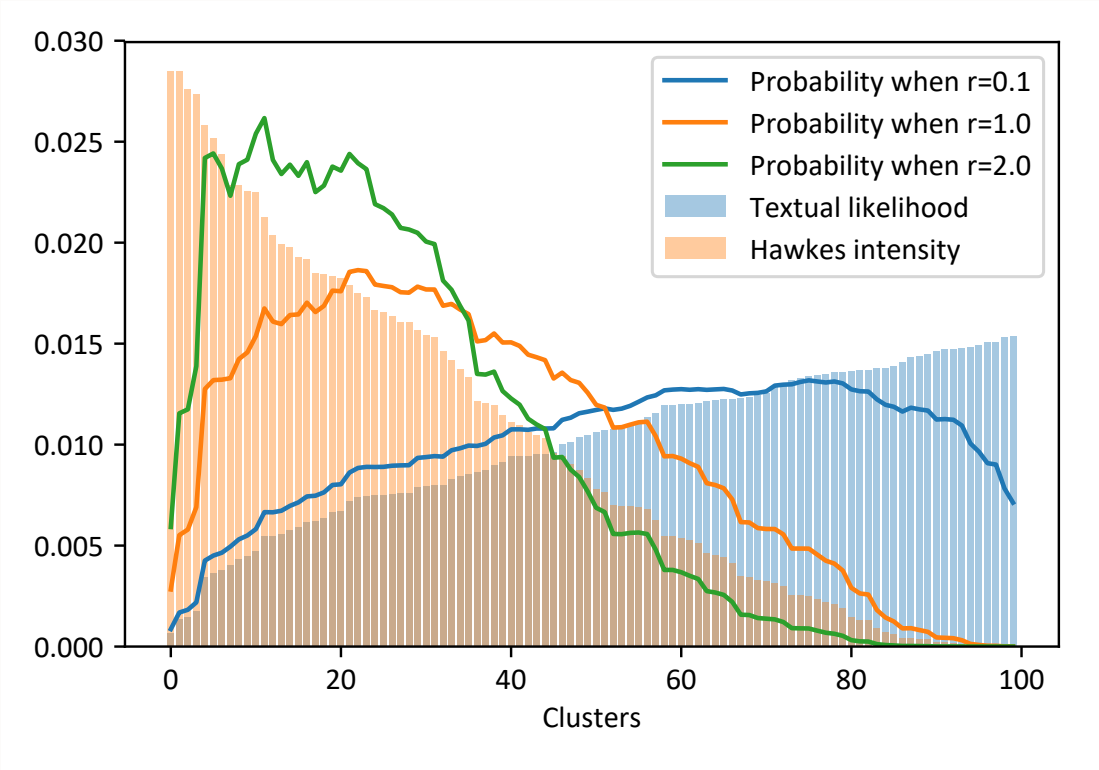
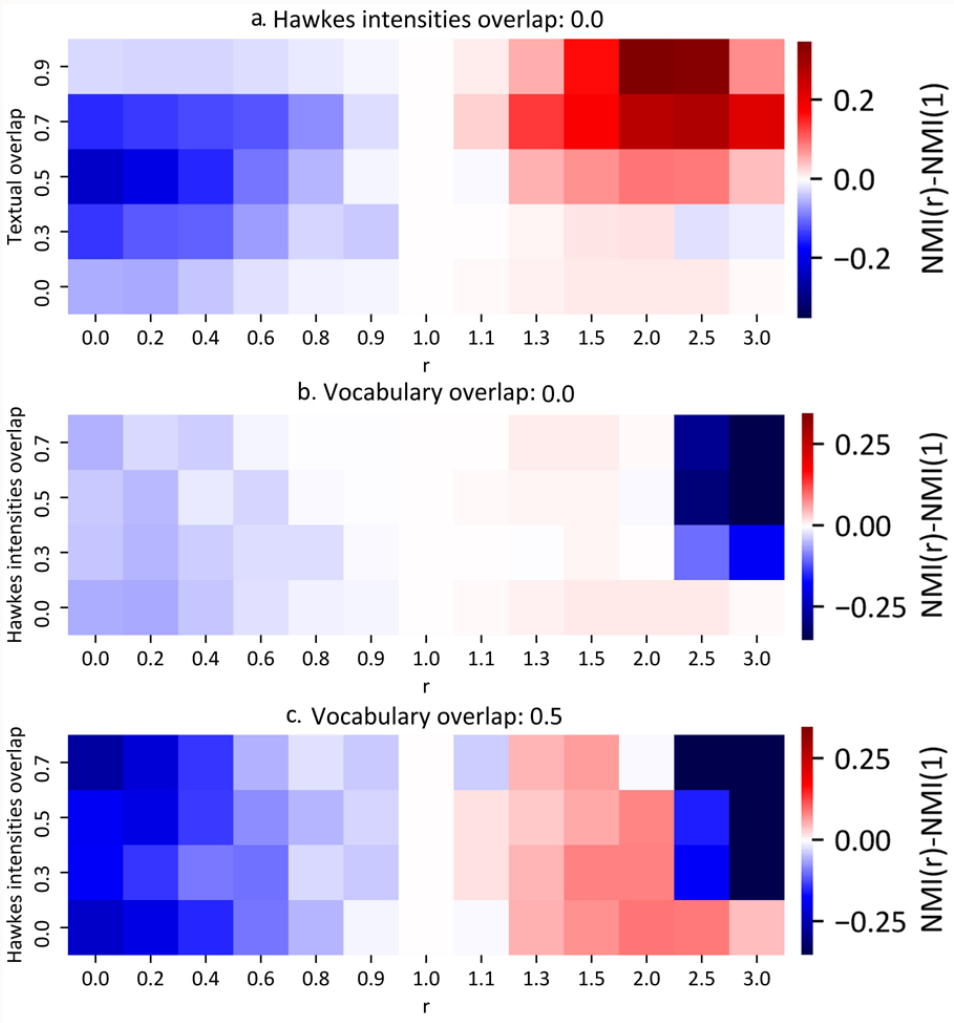


Figure 4: Varier le paramètre r donne $\pm$ d'importance à l'aspect temporel du clustering

PDHP



Explications

- Rappel : $r = 1$ est le processus de Hawkes-Dirichlet classique
- Meilleurs résultats quand le texte est peu informatif (overlap textuel > 1)
- Résultats similaires quand le texte est informatif (overlap textuel ≈ 0)

Figure 5: Différence de NMI entre le PDHP et le DHP pour différents overlaps textuels et temporels

PDHP

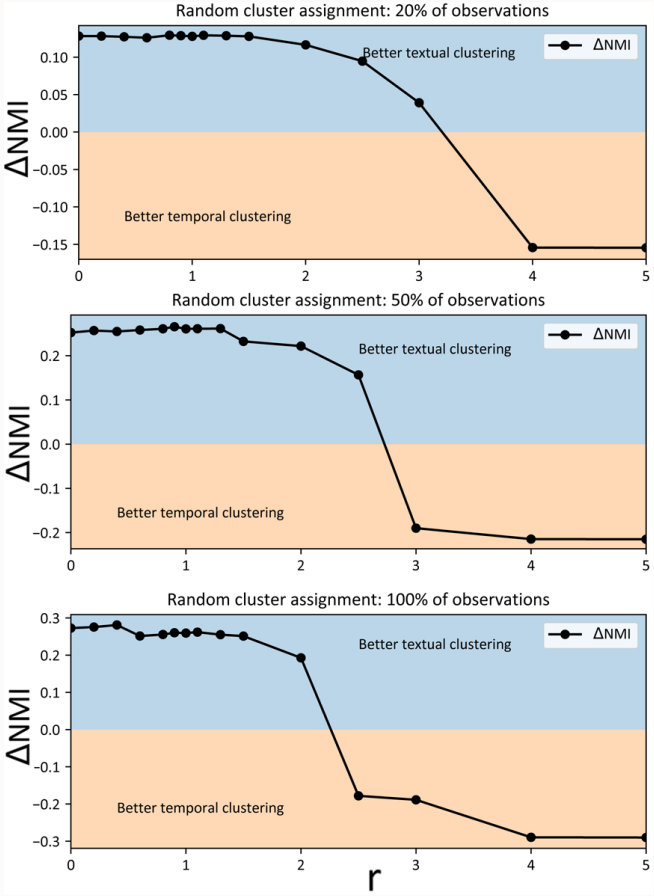


Figure 6: ΔNMI = NMI textuelle - NMI temporelle, pour différentes valeurs de décorrélation

Explications

- On décorrèle les clusters textuels et temporels : des documents au contenu textuel identique ne suivent pas nécessairement la même dynamique de publication.
- Fig de gauche : on décorrèle 20%, 50% et 100% des observations.
- Varier r permet de récupérer l'un ou l'autre clustering (en fonction du texte ou de la dynamique de publication).

PDHP

Et quid du monde réel ?

- Faire varier r permet de privilégier un clustering plutôt textuel ou temporel sur des données du monde réel.
- On considère trois corpus provenant de Reddit : plusieurs subreddits de news (**News**), le subreddit **TodayILearned** et le subreddit **AskScience**. On considère l'intégralité des titres de posts publiés en avril 2019.

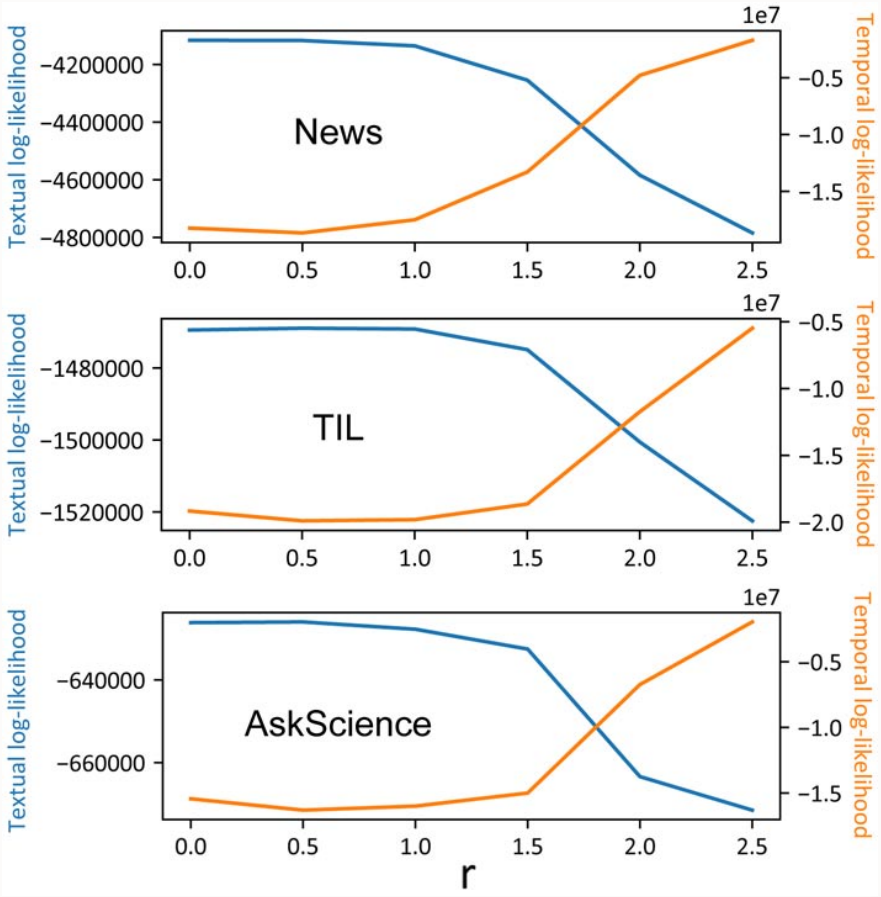


Figure 7: Likelihood textuelle (modèle Dirichlet-Multinomial) et likelihood temporelle (Hawkes process) versus r sur des données de Reddit.

PDHP

Et quid du monde réel ?

- Illustration avec le cluster inféré le plus proche du vocabulaire associé aux attentas du Sri Lanka pour différents r.
- Varier r modifie les dynamiques inférées et le contenu textuel du cluster.
- Le modèle a également repéré d'autres événements d'avril, comme l'incendie de Notre Dâme, l'arrestation de Julien Assange, le rapport Mueller à charge de Trump, la première photo d'un trou noir, ...

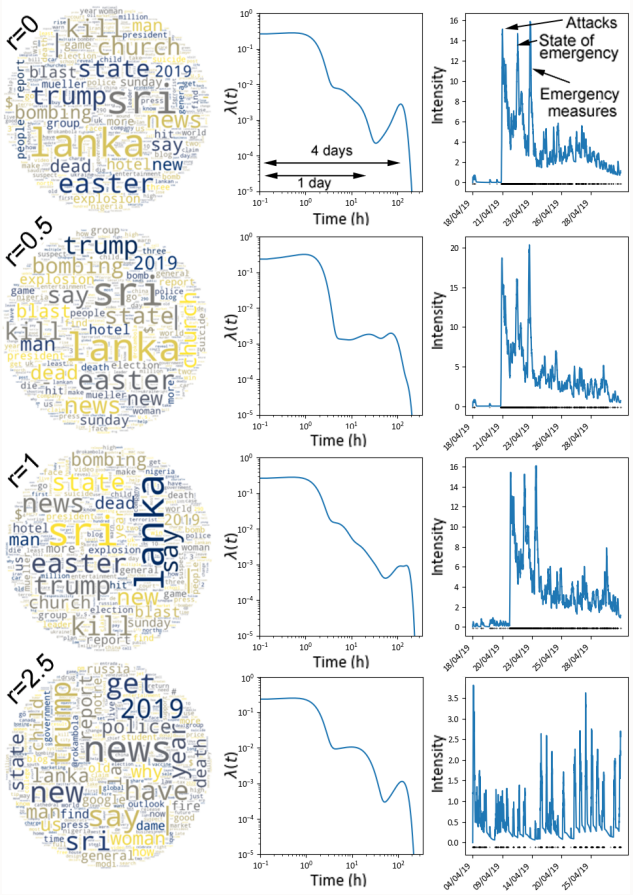


Figure 8: Likelihood textuelle (modèle Dirichlet-Multinomial) et likelihood temporelle (Hawkes process) versus r sur des données de Reddit.

Sommaire

A venir

Projets

- OUsToN (**O**nline **US**er-**TO**pic **N**etwork) : Survival-Dirichlet process pour inférer des clusters d'information et le réseau sous-jacent permettant leur diffusion.
- Création et obtention du prix nobel d'informatique.
- Version multivariée des PDHP pour inférer une interaction entre clusters d'information.
- Version de OUsToN incluant une notion d'interaction entre informations au niveau des noeuds individuels. **M**ultivariate **O**nline **U**ser **TO**pic **N**etwork ?

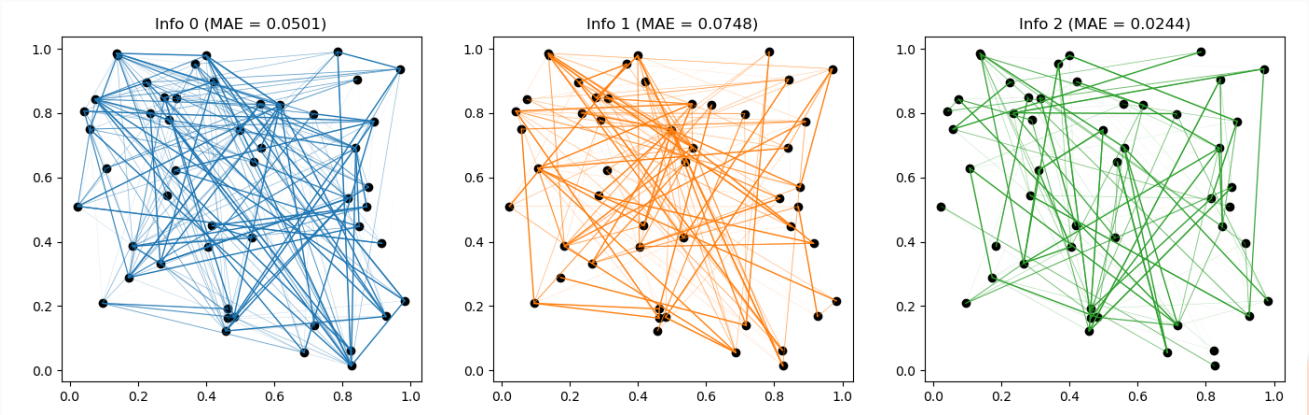


Figure 9: Résultats préliminaires de OUsToN sur données synthétiques

Merci de votre attention !