

Interactions entre Informations dans les Processus de Diffusion Culturelle – Mêmes et archéologie

Gaël Poux-Médard

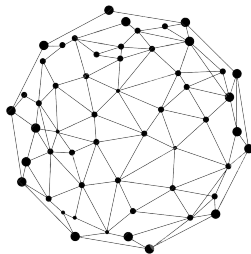
UCBL Lyon 1, Institut Camille Jordan
IARC/WHO, NME branch

6 mars 2026



Introduction

- Grandes quantités de données :
 - 1M de posts Reddit/jour, 500M de tweets, 1B sur Instagram etc.
- Comprendre les mécanismes de diffusion est fondamental
 - Fake news, promotion de contenus, biais, polarisation, etc.
- Littérature concentrée sur les interactions :
 - Humain-humain (liens sociaux, centralité, etc.)
 - Humain-information (préférences, sous-réseaux, etc.)
 - Information-humain (influence, polarisation, etc.)



Interactions

- Modéliser interactions entre informations affine cette compréhension
- Influence des précédents posts sur les suivants ?
 - Réaction tweets climatiques $\propto$ Articles sur des feux de forêt
 - Réaction tweets politiques $\propto$ Articles sur l'achat d'un chiot
 - Réaction à une pub pour du PQ $\propto$ News sur une nouvelle pandémie



Figure 1: 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.

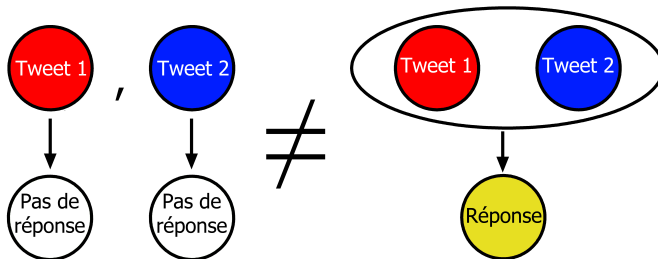


Figure 2: Exemple schématique d'une interaction

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 Usages au delà de l'information en ligne ?



Fréquence

Les interactions significatives sont-elles fréquentes?
→ *Stochastic Block Models*

Stochastic Block Models

- Modéliser un réseau d'interactions :
 - Noeuds : documents (ex. tweets)
 - Liens : résultat d'une relation entre documents (ex. retweet ou non)
- Réduction de dimensionnalité via clustering
 - Regrouper les informations en fonction de leurs (non-)interactions.

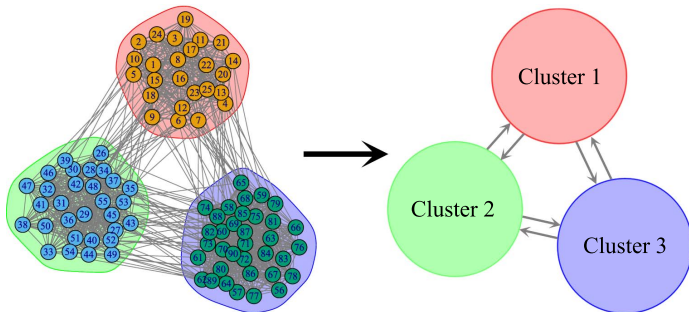
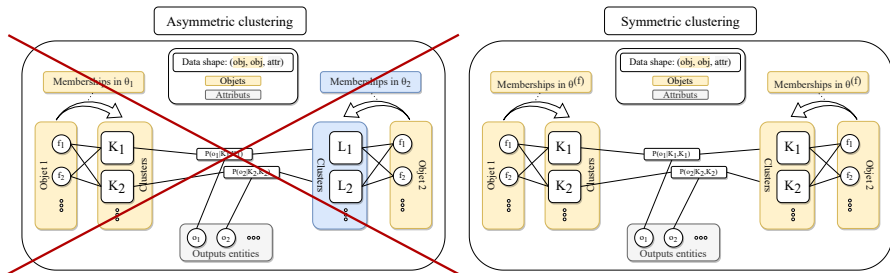


Figure 3: SBM - Supposer l'existence de groupes simplifie le système

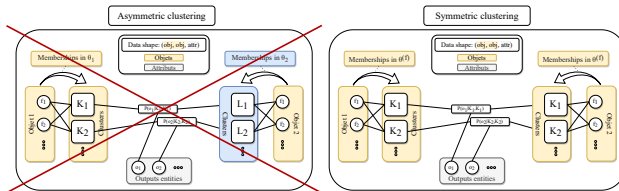
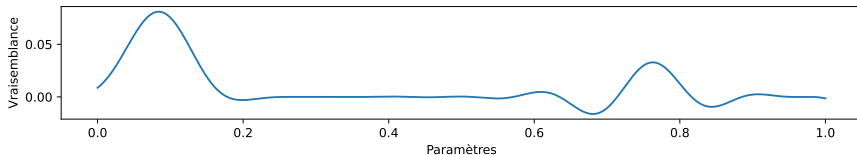
Interactions

- Schématiquement
 - Clustering partagé par toutes les informations
 - Interactions symétriques
 - $P(\text{action} | \text{info1}, \text{info2}) = P(\text{action} | \text{info2}, \text{info1})$



Apprentissage

- On spécifie la structure du modèle, mais les paramètres sont appris à partir des données
 - $p(\text{données}|\text{params}) = f(\text{données})$ est notre modèle
 - $p(\text{données}|\text{params}) = f(\text{params}) \propto p(\text{params}|\text{données})$ est sa "vraisemblance"
- Trouver params qui maximisent cette probabilité



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



Reviews | 7 Pol Marcin Lekarski, 2016 Feb;6(2):329-129-33.

[Post-Lyme disease syndrome]

[Article in Polish]
Lustyna Blaut-Jurkowska ¹, Marcin Jurkowski [†]

Affiliations: a - expand
PMD: 270006020

Abstract

Lyme disease is a chronic infectious disease caused by the bacteria, spirochete of the *Borrelia* type. Skin, nervous system, musculoskeletal system and heart may be involved in the course of the disease. The prognosis for properly treated Lyme disease is usually good. However, in about 5% of patients so called Post-Lyme disease syndrome (PLSD) develops. It is defined as a syndrome of subjective symptoms persisting despite proper treatment of *Borrelia burgdorferi* infection. The most common symptoms include fatigue, muscle and joint pain and problems with concentration. The pathogenesis of PLSD remains unknown. The differential diagnosis should include neurological, rheumatic and mental diseases. Till now there is no causative treatment of PLSD. In relieving symptom, rehabilitation, painkillers, anti-inflammatory and antidepressants medicines are recommended. Emotional and psychological supports are also necessary. Non-specific symptoms reported by patients with post-Lyme disease syndrome raise the suspicion of other pathologies. This can lead to misdiagnosis and implementation of unnecessary, potentially harmful to the patient's therapy. An increase in tick-borne diseases need to increase physicians awareness of these issues.

Keywords: Borrelia burgdorferi; post-Lyme disease syndrome (PLSD).

© 2016 WJEDPRESS.

Résultats

- Nos résultats en prédiction sont meilleurs que ceux des baselines
- Pourtant, les interactions entre clusters semblent peu significatives

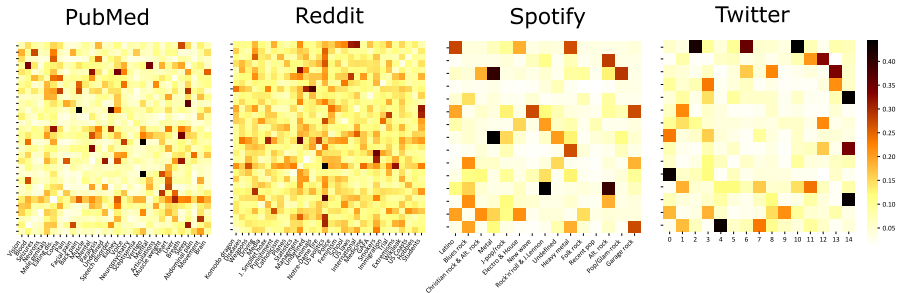


Figure 4: Influence des interactions entre clusters

P1: Les interactions significatives sont rares

Un problème de temps ?

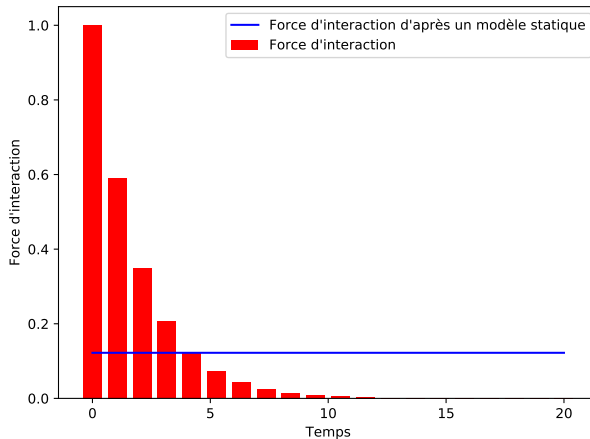


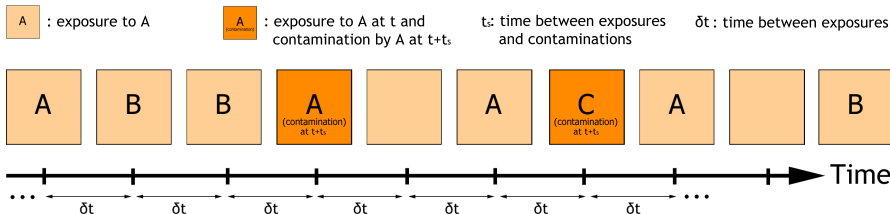
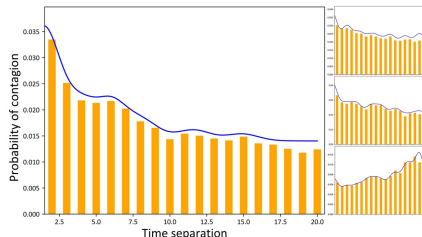
Figure 5: Les modèles statiques peuvent omettre des interactions.

Durée

Les interactions perdurent-elles dans le temps ?

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

- Deux jeux de données :
 - Twitter : anciens tweets influencent nouveaux retweets ?
 - Ads : exposition à des pubs influencent les probabilités de click ?

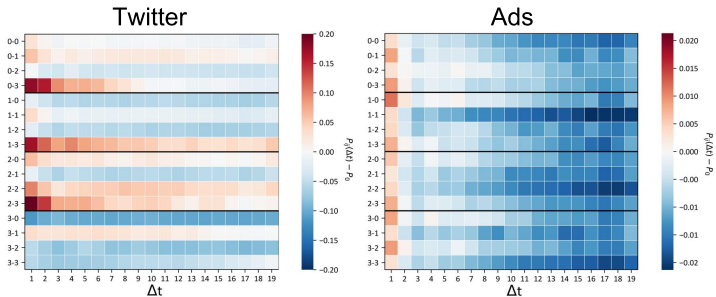


Figure 6: Conclusion : les interactions entre informations ont un effet bref

P2: Les interactions sont brèves

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



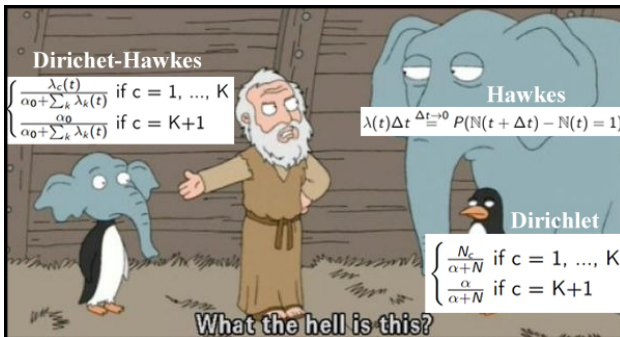
À grande échelle

*Modéliser des interactions brèves et rares à grande échelle ?
→ Processus de Dirichlet-Hawkes*

Processus de Dirichlet-Hawkes

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

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



Processus de Dirichlet - Modèles séquentiels

- Processus de Dirichlet:

$$P(cluster) = \begin{cases} \frac{N_c}{\alpha + N} & \text{si } c = 1, \dots, K \\ \frac{\alpha}{\alpha + N} & \text{si } c = K+1 \end{cases}$$

- Un a priori utile pour modéliser des données séquentielles

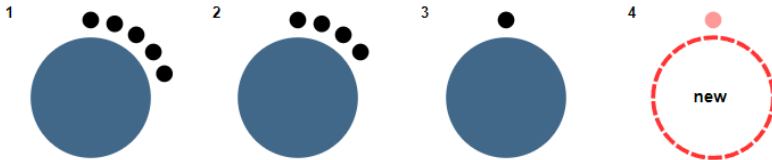


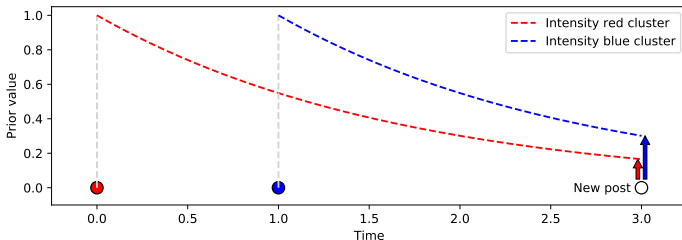
Figure 7: Réalisation d'un processus de Dirichlet (10 étapes)

Processus de Dirichlet-Hawkes - Mathématiquement

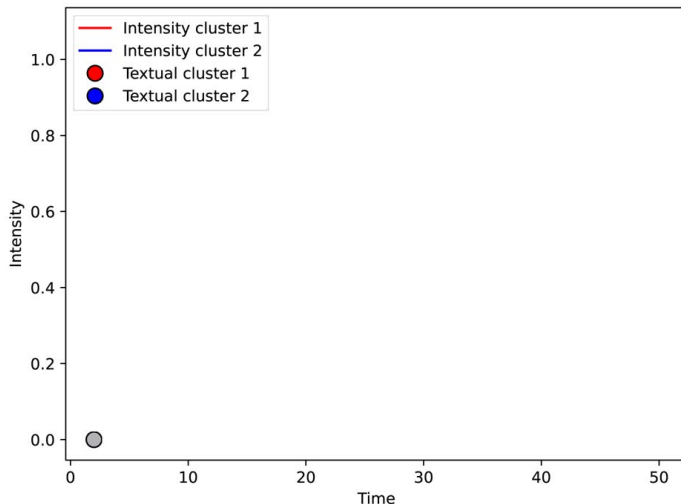
- Dirichlet-Hawkes Process (Du *et al.*, KDD 2015)
- Comptes N_c de Dirichlet remplacés par intensités $\lambda_c(t)$:

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

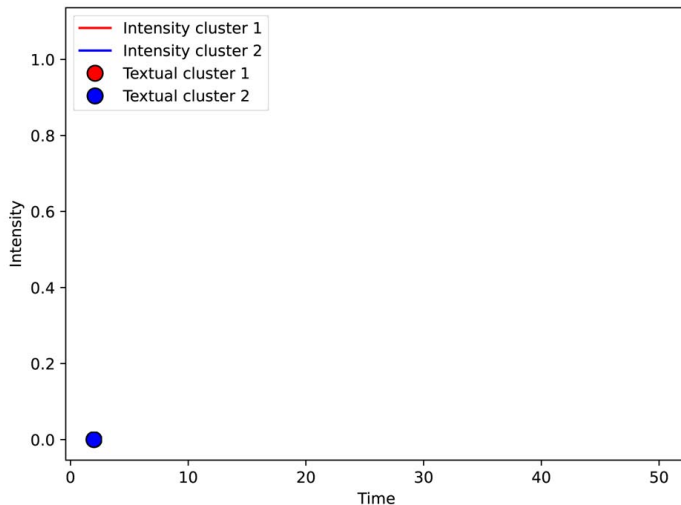
↓



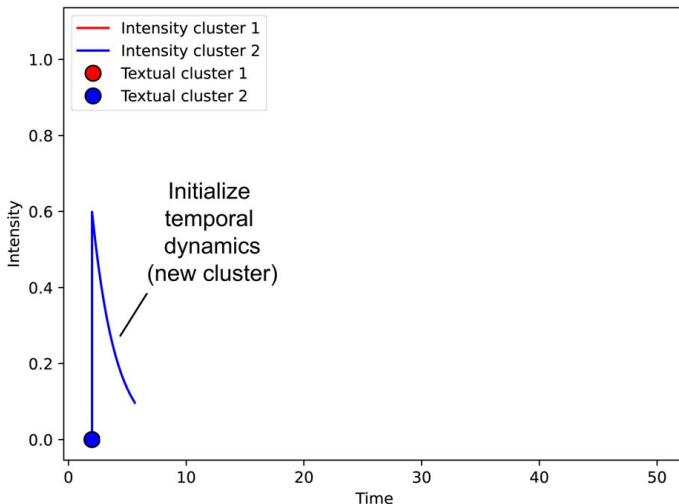
Processus de Dirichlet-Hawkes - Inférence (1 instance)



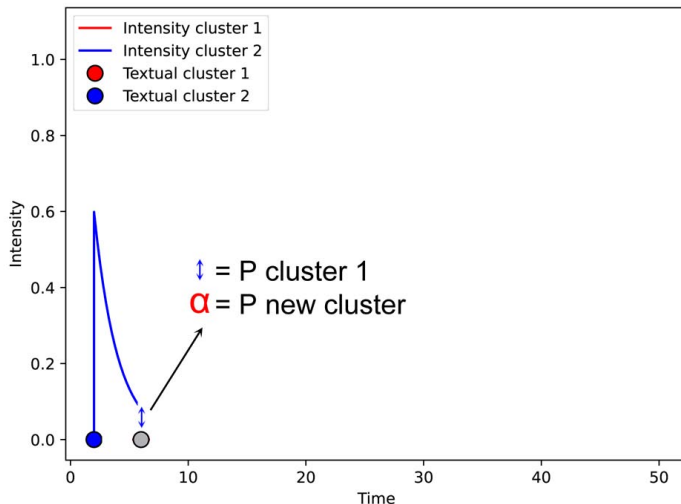
Processus de Dirichlet-Hawkes - Inférence (1 instance)



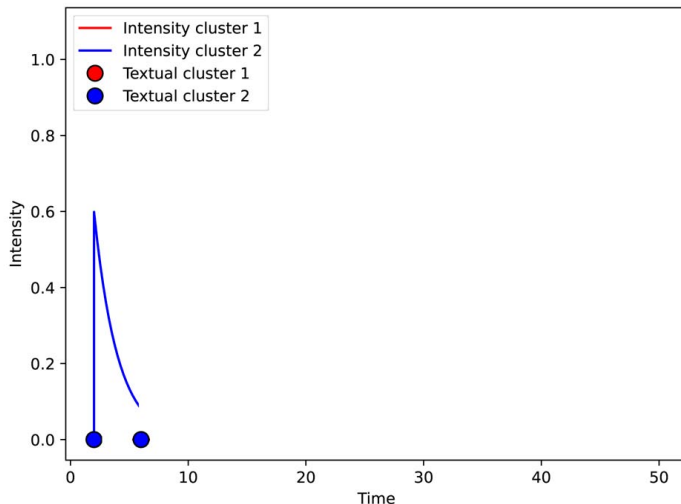
Processus de Dirichlet-Hawkes - Inférence (1 instance)



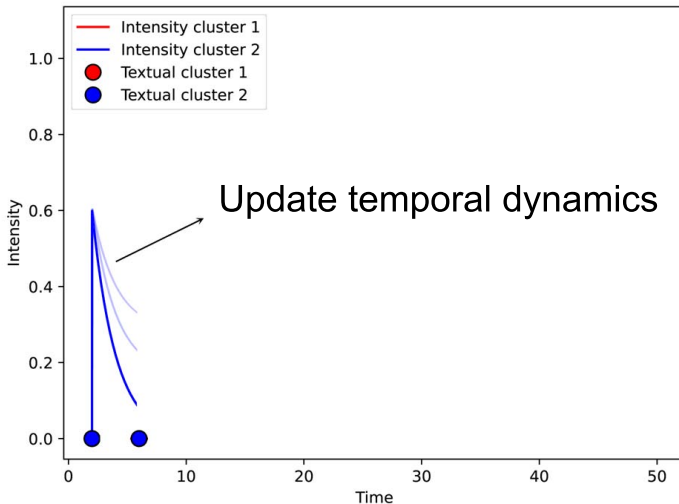
Processus de Dirichlet-Hawkes - Inférence (1 instance)



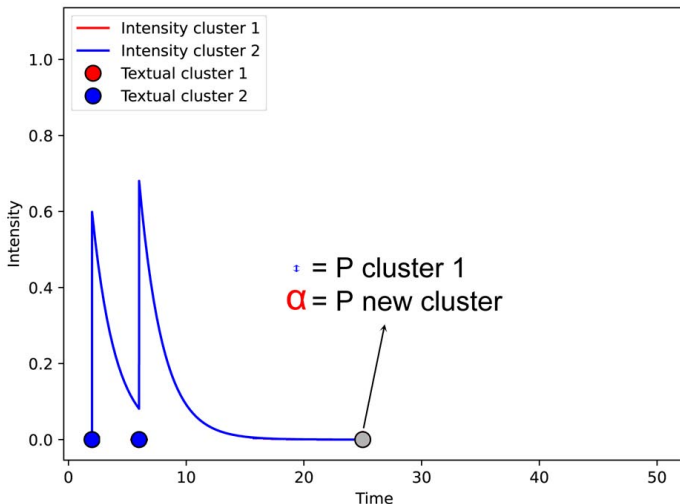
Processus de Dirichlet-Hawkes - Inférence (1 instance)



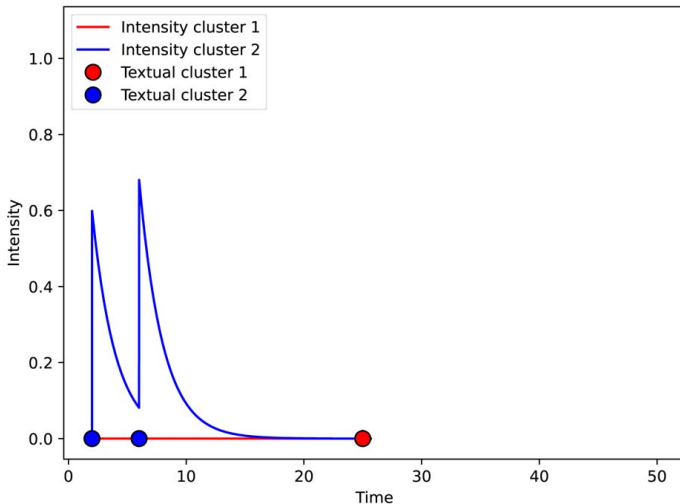
Processus de Dirichlet-Hawkes - Inférence (1 instance)



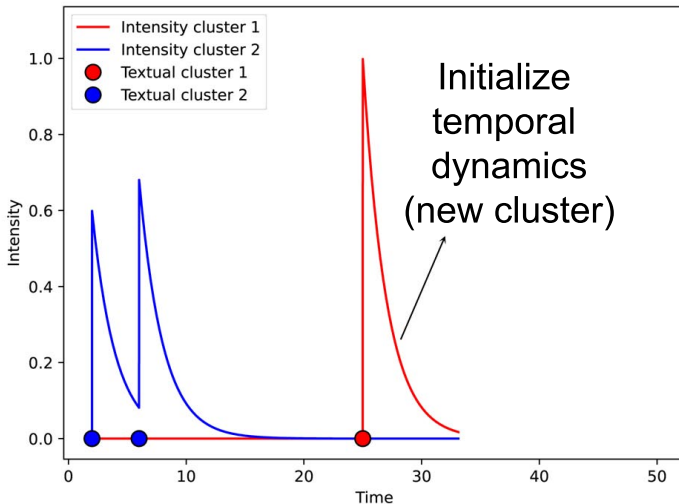
Processus de Dirichlet-Hawkes - Inférence (1 instance)



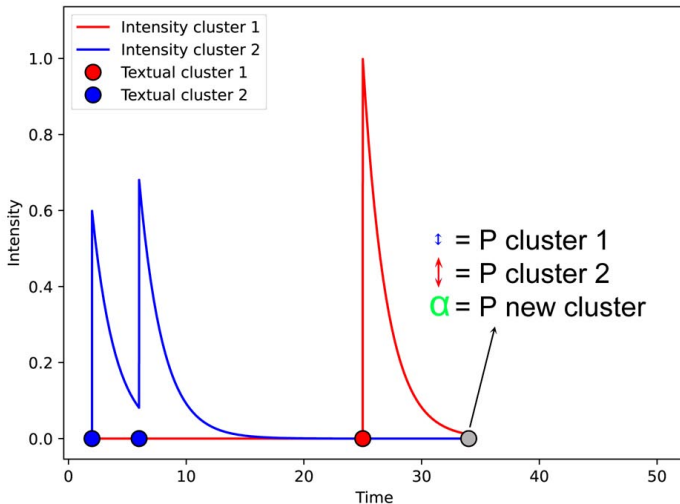
Processus de Dirichlet-Hawkes - Inférence (1 instance)



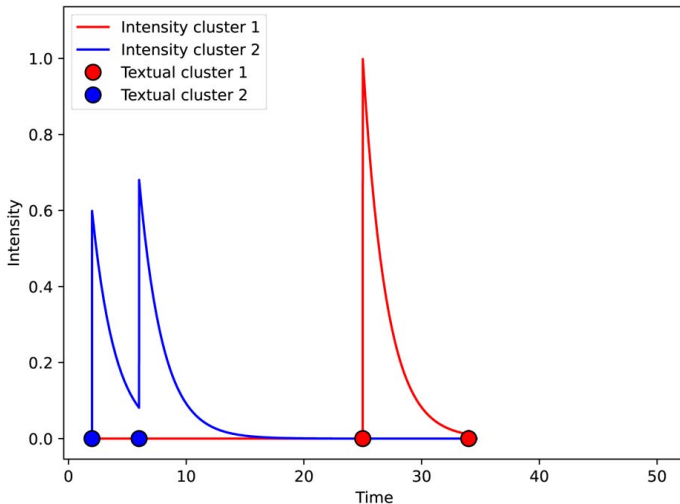
Processus de Dirichlet-Hawkes - Inférence (1 instance)



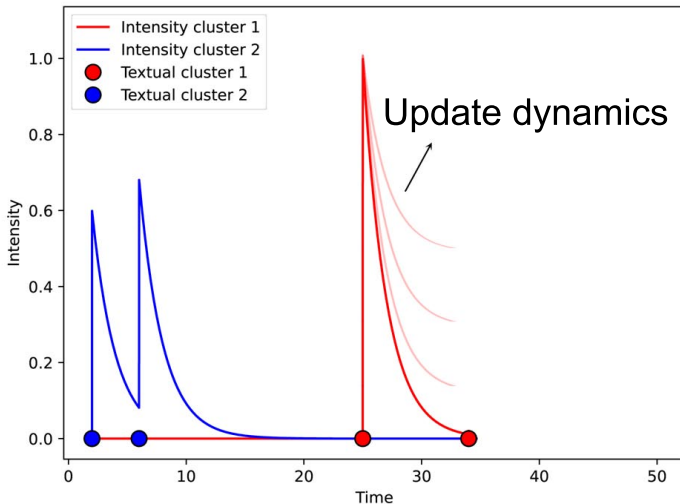
Processus de Dirichlet-Hawkes - Inférence (1 instance)



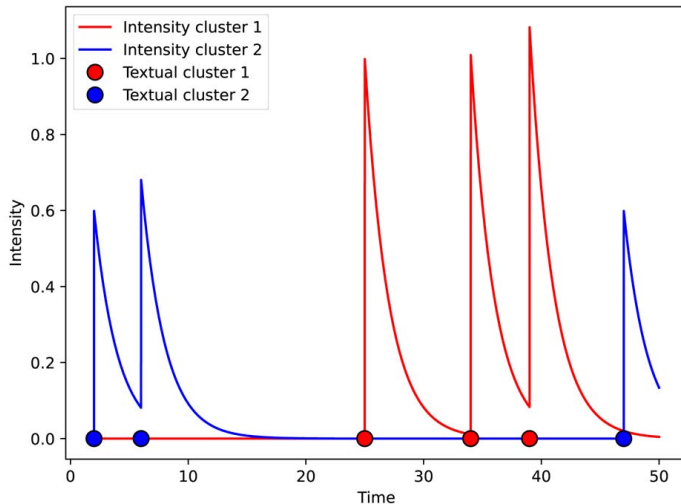
Processus de Dirichlet-Hawkes - Inférence (1 instance)



Processus de Dirichlet-Hawkes - Inférence (1 instance)

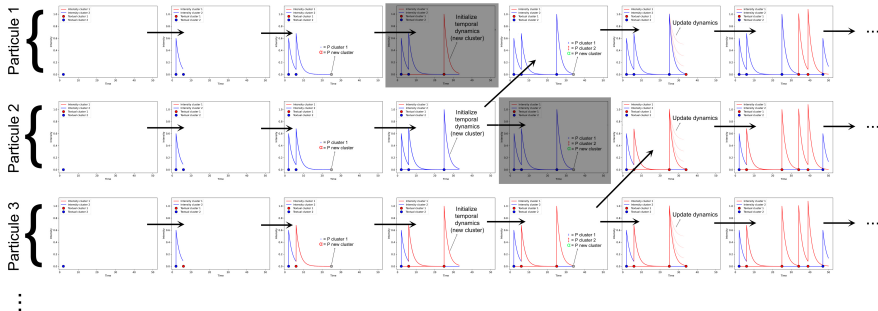


Processus de Dirichlet-Hawkes - Inférence (1 instance)



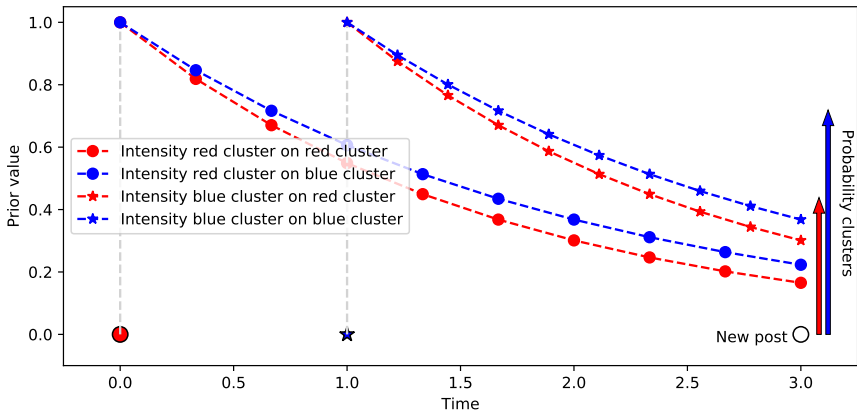
Processus de Dirichlet-Hawkes - Inférence (SMC)

- Algorithme de Monte Carlo séquentiel



Multivariate Powered Dirichlet-Hawkes Process

- Multivariate Powered Dirichlet-Hawkes prior
 - *Multivariate*: modélise des interactions intra-cluster et extra-cluster
 - *Powered*: contrôle l'importance de l'information temporelle



Multivariate Powered Dirichlet-Hawkes Process - Données synthétiques

- Données synthétiques (2 clusters, 20 mots/doc., 5000 obs.)
- Récupération des clusters sous-jacents ?

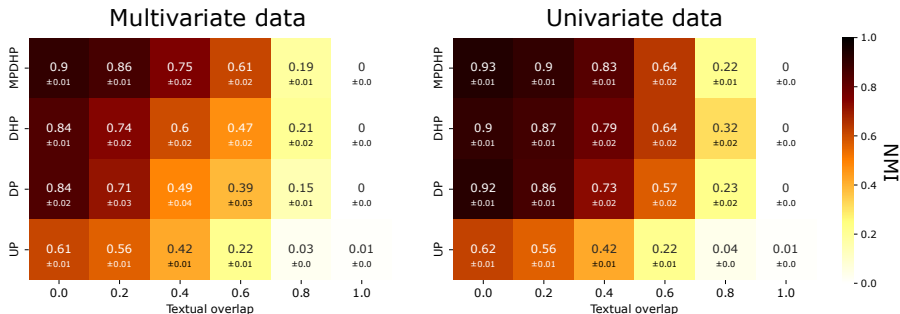


Figure 8: Résultats de MPDHP sur des données synthétiques (100 runs)

P3: Modélisation efficace d'interactions rares et brèves

Étude de cas

*Quantifier le rôle des interactions dans la diffusion de l'information ?
Étude de cas : le dataset Reddit News*

Inferred clusters

- 100 000 titres de news sur Reddit en 2019

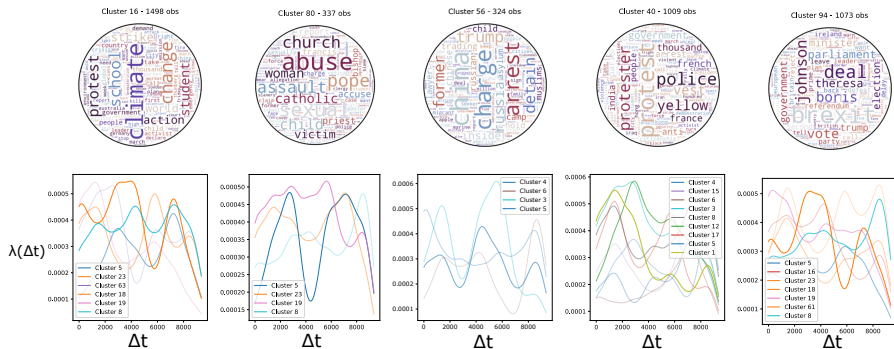


Figure 9: Quelques clusters inférés par MPDHP (haut), et leur dynamique d'interaction avec les autres clusters (bas). Ex. : un document du cluster 16 influe fortement la probabilité de publication d'un document du cluster 23 environ 4000s après sa publication.

Réseau d'interaction entre clusters

- Réseau d'interaction entre clusters

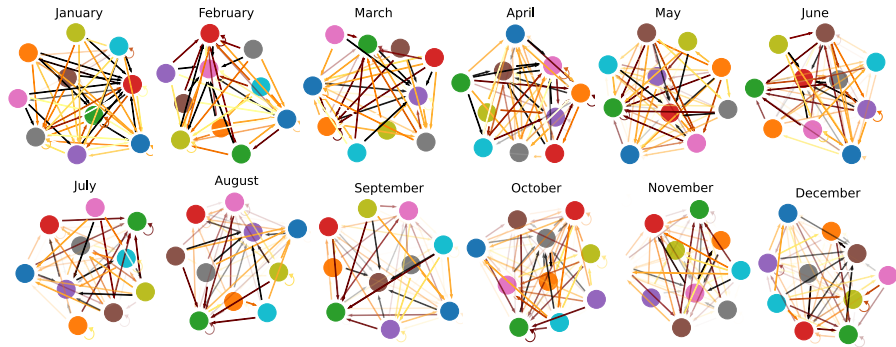


Figure 10: Évolution du réseau d'interaction à courte portée entre les clusters sur toute l'année. Noeuds = clusters ; Liens = force d'interaction

Résultats consistants

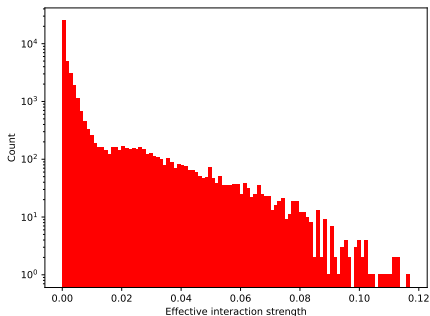


Figure 11: Distribution typique de la force des interactions dans Reddit News.

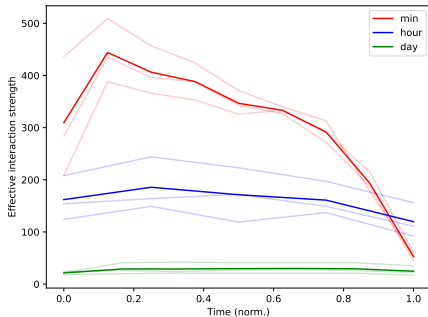


Figure 12: La force des interactions effectives diminue dans le temps

- Nous retrouvons nos premières conclusions.
- Peu d'interactions → Dataset généré par un processus exogène ?

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

→ Un nouveau jeu d'outils dédié à l'étude de cas spécifiques où les interactions pourraient jouer un rôle.

[/thèse]
- ... Et justement !

Une autre application

- Dans la diffusion en ligne :
 - Le réseau est disponible : relations, acteurs, activité, etc.
 - L'information est accessible : mêmes, news, posts, etc.
- En archéologie :
 - Le réseau a disparu : relations sociales, commerciales, politiques, etc.
 - Seule subsiste l'information : artefacts, inscriptions, noms, etc.

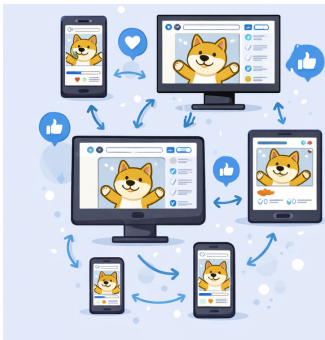


Figure 13: PCR Gd Est

1. Forêt-Hell - Scherwood
2.9 Montserrat - Schwaabacher et Au Wessertal
4. Götterwald - Herten
5. Neudorf - Hohenberg/Hungary
6. Jagelwald - Hohenberg
7. Grotte - Hohenberg et Hohenberg
8. Dürrenwald - Scherwood
9.5. Scherwood - Hohenberg
11. Hohenberg - Hohenberg
12. Hohenberg - Hohenberg
13. Langenwald - ZAC Zoon commercial Nord
14. Götter - Scherwood/Lefly
15. Hohenberg - Hohenberg
16. Hohenberg - Hohenberg
17. Hohenberg - Hohenberg
18. Hohenberg - Hohenberg
19. Hohenberg - Hohenberg
20. Hohenberg - Hohenberg
21. Hohenberg - Hohenberg
22. Hohenberg - Hohenberg
23. Hohenberg - Hohenberg
24. Hohenberg - Hohenberg
25. Hohenberg - Hohenberg
26. Hohenberg - Hohenberg
27. Hohenberg - Hohenberg
28. Hohenberg - Hohenberg
29. Hohenberg - Hohenberg
30. Hohenberg - Hohenberg
31. Hohenberg - Hohenberg
32. Hohenberg - Hohenberg
33. Hohenberg - Hohenberg
34. Hohenberg - Hohenberg
35. Hohenberg - Hohenberg
36. Hohenberg - Hohenberg
37. Hohenberg - Hohenberg
38. Hohenberg - Hohenberg
39. Hohenberg - Hohenberg
40. Hohenberg - Hohenberg
41. Hohenberg - Hohenberg
42. Hohenberg - Hohenberg
43. Hohenberg - Hohenberg
44. Hohenberg - Hohenberg
45. Hohenberg - Hohenberg
46. Hohenberg - Hohenberg
47. Hohenberg - Hohenberg
48. Hohenberg - Hohenberg
49. Hohenberg - Hohenberg
50. Hohenberg - Hohenberg
51. Hohenberg - Hohenberg
52. Hohenberg - Hohenberg
53. Hohenberg - Hohenberg
54. Hohenberg - Hohenberg
55. Hohenberg - Hohenberg
56. Hohenberg - Hohenberg
57. Hohenberg - Hohenberg
58. Hohenberg - Hohenberg
59. Hohenberg - Hohenberg
60. Hohenberg - Hohenberg
61. Hohenberg - Hohenberg
62. Hohenberg - Hohenberg
63. Hohenberg - Hohenberg
64. Hohenberg - Hohenberg
65. Hohenberg - Hohenberg
66. Hohenberg - Hohenberg
67. Hohenberg - Hohenberg
68. Hohenberg - Hohenberg
69. Hohenberg - Hohenberg
70. Hohenberg - Hohenberg
71. Hohenberg - Hohenberg
72. Hohenberg - Hohenberg
73. Hohenberg - Hohenberg
74. Hohenberg - Hohenberg
75. Hohenberg - Hohenberg
76. Hohenberg - Hohenberg
77. Hohenberg - Hohenberg
78. Hohenberg - Hohenberg
79. Hohenberg - Hohenberg
80. Hohenberg - Hohenberg
81. Hohenberg - Hohenberg
82. Hohenberg - Hohenberg
83. Hohenberg - Hohenberg
84. Hohenberg - Hohenberg
85. Hohenberg - Hohenberg
86. Hohenberg - Hohenberg
87. Hohenberg - Hohenberg
88. Hohenberg - Hohenberg
89. Hohenberg - Hohenberg
90. Hohenberg - Hohenberg
91. Hohenberg - Hohenberg
92. Hohenberg - Hohenberg
93. Hohenberg - Hohenberg
94. Hohenberg - Hohenberg
95. Hohenberg - Hohenberg
96. Hohenberg - Hohenberg
97. Hohenberg - Hohenberg
98. Hohenberg - Hohenberg
99. Hohenberg - Hohenberg
100. Hohenberg - Hohenberg

Diffusion culturelle en archéologie

- De ce travail acharné, émergent peu à peu de nouvelles bases
- En particulier :
 - Artefacts : 200k petits objets en Europe de l'ouest, datés et géolocalisés
 - Clauss-Slaby : 200k inscriptions en Europe de l'ouest, datées et géolocalisées
 - EDH, RGZM, Trismegistos, etc.

Imperium Romanum

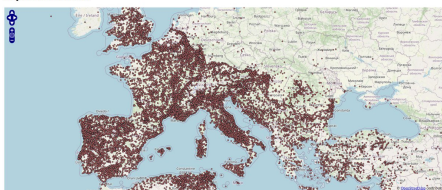
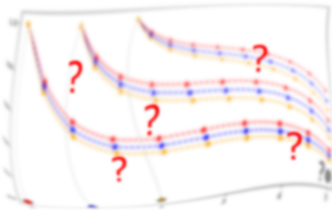
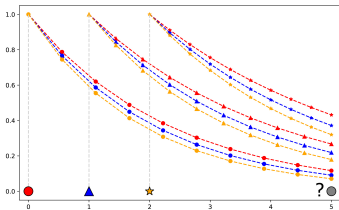


Figure 14: Clauss-Slaby (BdD épigraphique) & Artefacts (BdD petit mobilier)

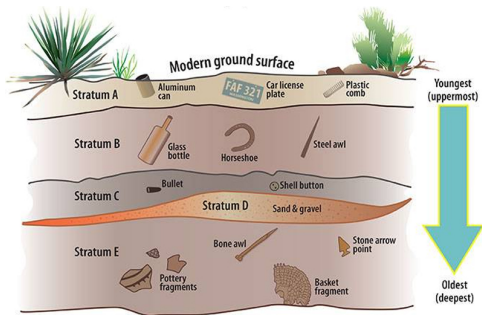
Approche frontale ?

- Diffusion culturelle : comment les artefacts, langages, noms, co-évoluent dans l'espace et le temps.
 - Plutôt centrale : "par quels mécanismes en sommes nous arrivés là"
- Il n'y aurait plus qu'à rentrer les données dans nos modèles ? Non :
 - Données lacunaires, biaisées, incertaines, bruitées, hétérogènes
 - Besoin d'outils dédiés, l'un des axes du projet
 - Trop d'incertitude pour une approche frontale
 - Groupes spatio-temporels, les 3/4 du projet



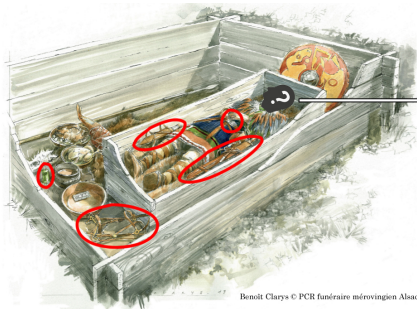
Typochronologie automatisée

- Diffusion culturelle multivariée : artefacts, langages, noms
- Typochronologie :
 - Datation par co-occurrence spatiale et temporelle (=strate)
 - On en connaît pour des amphores, céramiques, verres, etc.
- Autres typochronologies à découvrir dans les données ?
 - ① Approximer des “strates” à partir d’objets datés et localisés
 - ② Déterminer quels *types* de co-occurrences sont localisantes



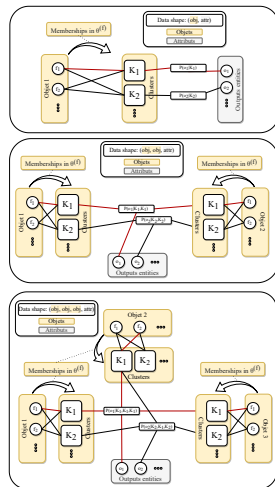
Interactions entre objets

- “Petit mobilier mérovingien en Alsace”



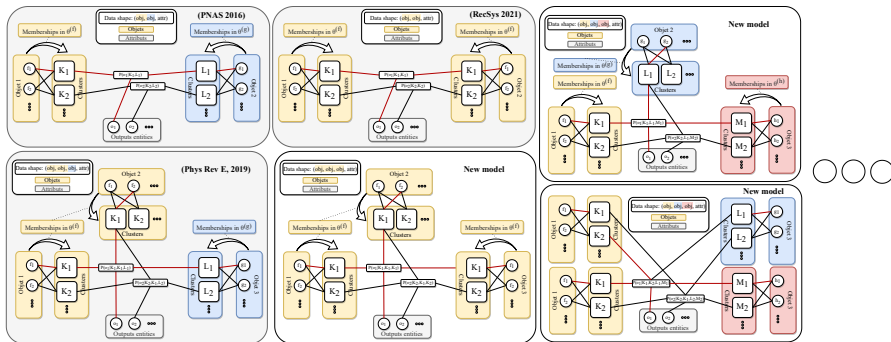
Benoît Clarys © PCR funéraire mérovingien Alsace

- 2000 petits objets
- 2 sexes (50/50)
- ~ 90% précision avec des paires



Interactions entre objets dans l'espace

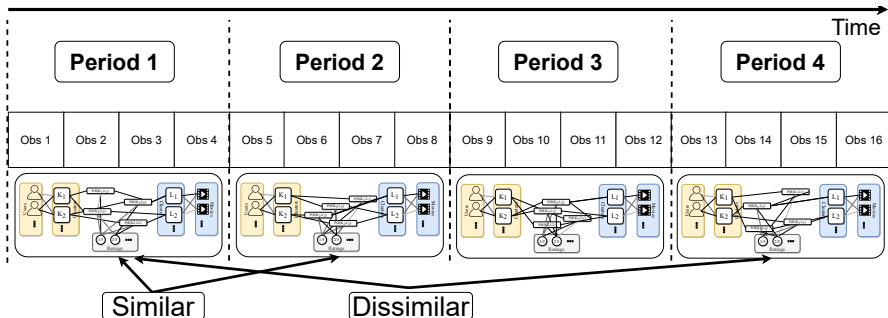
- Généralisation de (PNAS 2016, Phys.Rev.E, 2019, RecSys 2021)
 - Autant d'informations et d'interactions qu'on veut
 - Ex: (objets, ~~region~~, ~~date~~, présence), ou objets, ~~date~~, region



- “Petit mobilier mérovingien en Alsace”
 - Le temps ?

Interactions entre objets dans l'espace et le temps

- Modéliser le temps comme un continuum

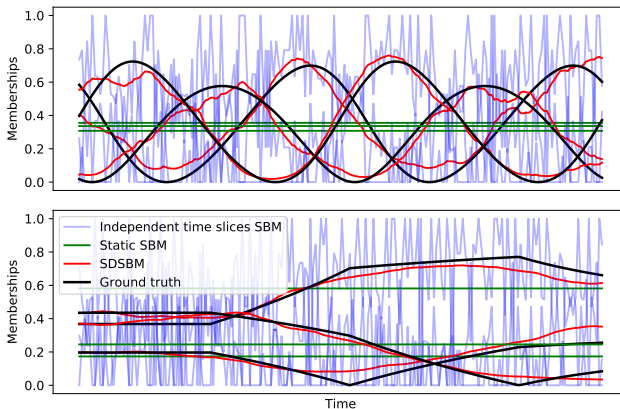


- Une hypothèse: les modèles proches dans le temps sont similaires
 - (Priors dépendants des voisins)

→ Evolution douce

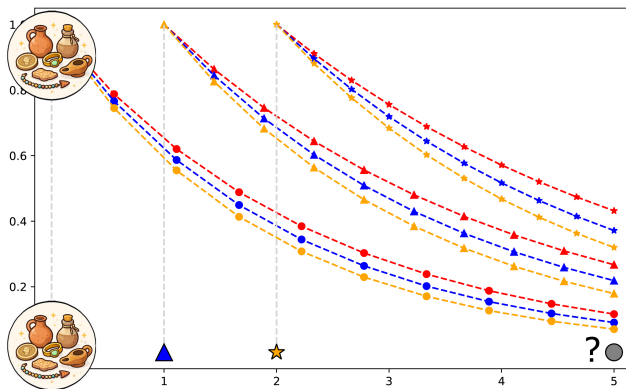
Résultats de l'approche

- “Petit mobilier mérovingien en Alsace”
- En sortie : probabilité que deux groupes d'objets co-occurrent à un instant et un endroit donnés, et composition de ces groupes.



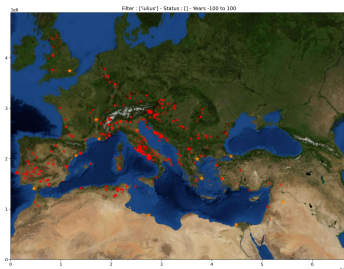
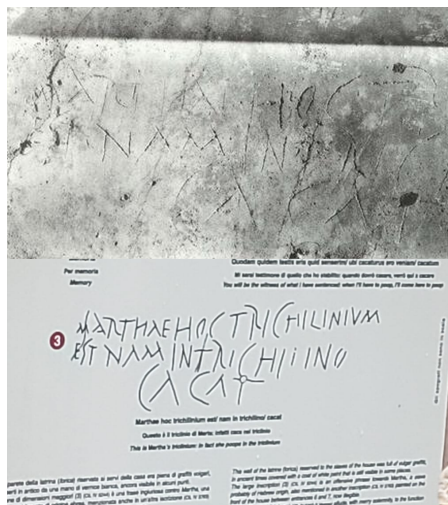
Résultat

- Diffusion culturelle multivariée : ~~artefacts~~, langages, noms
- “Petit mobilier mérovingien en Alsace”
 - On a nos groupes d'objets, ou typochronologies
 - Premier pas vers notre diffusion multifactorielle



Epigraphies

- Diffusion culturelle multivariée : ~~artefacts~~, langages, noms



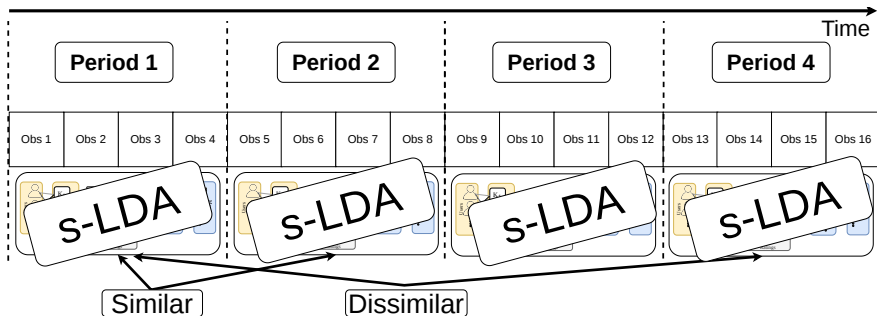
- ↑ epimap.msh-lse.fr/
- ← “Marthae hoc trichilinium est nam in trichilinio cacat”

Typochronologie linguistique automatisée

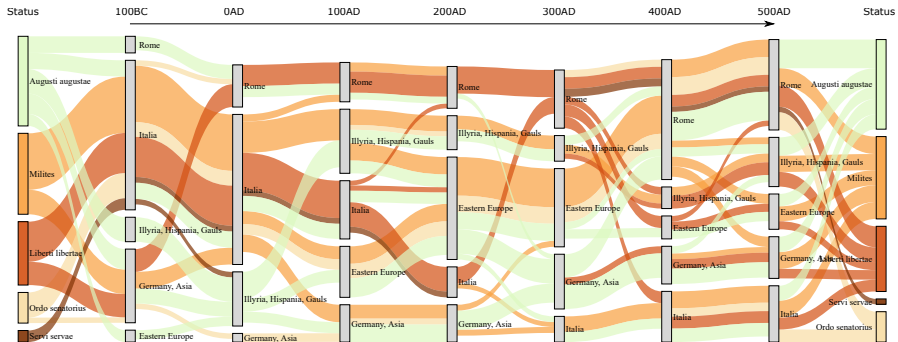
- Diffusion culturelle multivariée : ~~artefacts~~, langages, noms
- Typochronologie *linguistique*:
 - On sait que le type d'objet la locution X est usitée entre t_a et t_b
 - Locution X ($t_a < t < t_b$) dans une inscription mal datée = inscription bien datée
 - Fonctionne par co-occurrence *de mots* (locutions) spatiale et temporelle
- Exemples (connus, et débattus) :
 - DM/Dis Manibus : II-III^s. dans l'empire romain
 - HSE/Hic Situs Est : >I^{er}s.
 - Matronae : >I^{er}s. en Germanie, plus ancien en Italie
- Autres typochronologies à découvrir dans les données ?
 - ① Déterminer quelles ~~typologies~~ locutions sont datantes

Interactions entre mots dans l'espace et le temps

- Un problème similaire, mais interactions entre mots particulières
 - LDA : trouve des “topics” dans un corpus de documents textuels
 - Supervised LDA : les forme en utilisant d'autres infos (e.g. région)
- Manque le temps, mais :



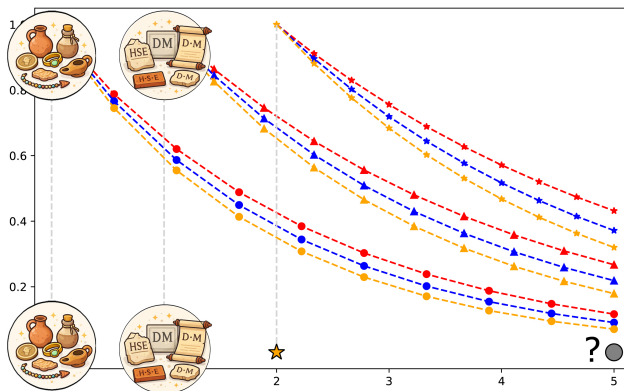
Un exemple illustratif



- 100k épigraphies avec date (de -100 à 500), région et statut associés
- Prédire la région à l'aide du texte au cours du temps
- Statut (militaire, esclave, ...) plutôt que mots, mais idée similaire

Résultat

- Diffusion culturelle multivariée : ~~artefacts~~, ~~langages~~, noms
 - On a nos groupes d'objets, ou typochronologies
 - On a nos locutions, ou typochronologies textuelles



Onomastie épigraphique

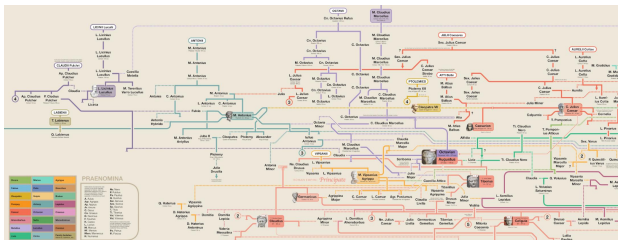
- Diffusion culturelle multivariée : ~~artefacts~~, ~~langages~~, noms



- ← “Dis Manibvs sacrvm Postumia Cnaei filia Sperata sacerdos Cererum publica pia vixit annis LXXI hic sita est”
- ← “Aux dieux Mânes sacrés. Postumia Sperata, fille de Cnaeus, prêtresse publique des Cérès, pieuse, a vécu 71 ans. Elle repose ici.”
- Père, métier, âge, parfois époux/se, origine, légion, etc.
 - Liens sociaux
- ECDS-00000008, Tunisie, non datée, mais “Dis Manibus”
 - I-III s.

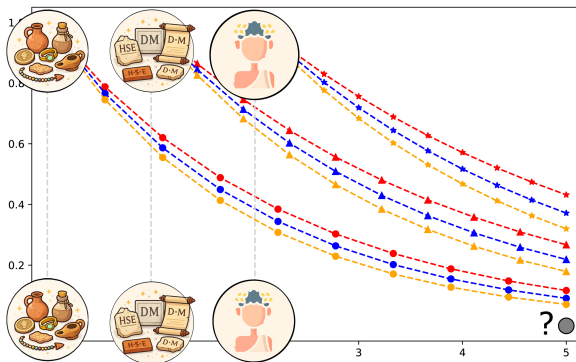
Typochronologie onomastique ?

- Diffusion culturelle multivariée : ~~artefacts~~, ~~langages~~, **noms**
- Typochronologie *onomastique* ?
 - Les individus ont tendance à transmettre leur nom à leur descendance
 - Les patronymes apparaissent en foyers localisés
 - Information spatiale et temporelle
- Objectif :
 - Extraire une base de donnée onomastique (datée, localisée)
 - Identifier des (groupes de) patronymes localisés
 - (long-shot) Inférer un arbre généalogique probabiliste



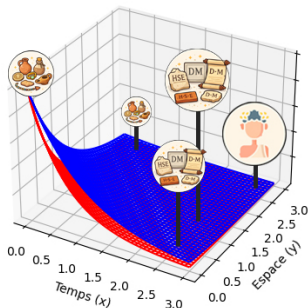
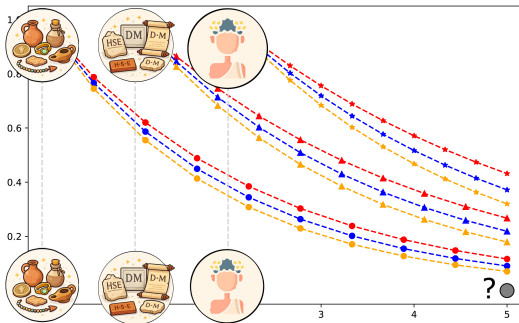
Résultat

- Diffusion culturelle multivariée : ~~artefacts~~, ~~langages~~, ~~noms~~
 - On a nos groupes d'objets, ou typochronologies
 - On a nos locutions, ou typochronologies textuelles
 - On a nos acteurs de diffusion
- Maintenant, on peut tenter de tout mettre ensemble



Diffusion culturelle multivariée

- Interactions spatio-temporelles entre objets, locutions et individus



- Horizon : 6 ans
- Candidature Access-ERC “EX-AEQUOS” pour venir au CAMS
 - Explaining Cultural Spread through Archaeological, Epigraphic and Onomastic Quantities at Scale

Merci pour votre attention !

Site web, articles et autres : <https://gaelpouxmedard.github.io/>

Code et données : <https://github.com/GaelPouxMedard/>

