Dirichlet-Point processes

Gaël Poux-Médard

Université de Lyon, France
Lyon 2, ERIC UR 3083

November 2021

Introduction

- Every minute:

400h of video
350 000 tweets
500 000 comments
4 200 000 searches

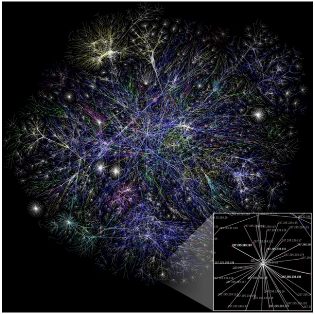


Figure 1: Snapshot of the internet (Wikipedia)

Motivation

- Every minute:

400h of video
350 000 tweets
500 000 comments
4 200 000 searches

- How to make sense out of *that*?

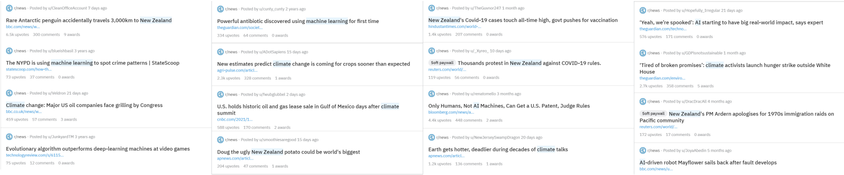


Figure 2: A typical stream from r/news

Motivation

- Every minute:

  400h of video
  350 000 tweets
  500 000 comments
  4 200 000 searches

- How to make sense out of *that*?
  → Hidden semantic links

Figure 2: A typical stream from r/news – with topics

Available information

- Main clues:
  - Textual information

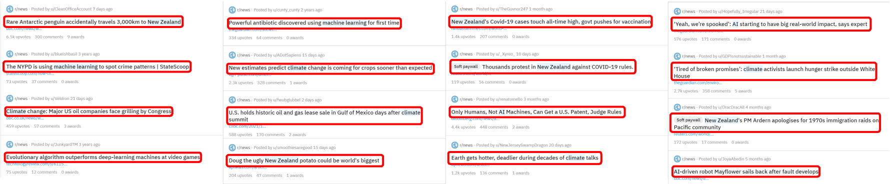


Figure 3: We can use textual information

Available information

- Main clues:
  - ◇ Textual information
  - ◇ Temporal information

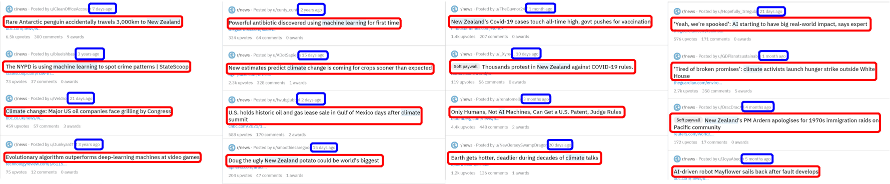


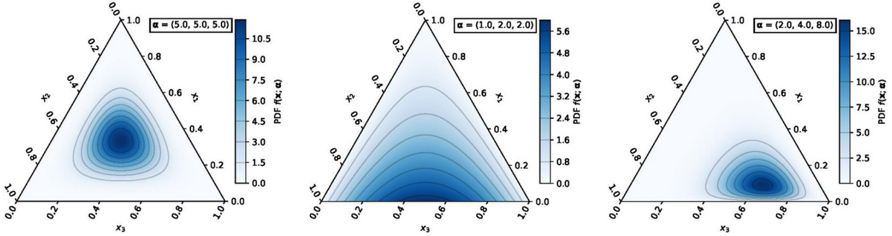
Figure 3: We can use textual information and temporal information

Documents stream

- The data is therefore a documents stream

Dirichlet process

- Dirichlet processes fit to consider streams as inputs
- Dirichlet distribution: $\vec{X} \sim Dir(\alpha)$ s.t. $\sum_k X_k = 1$
- Often used as a prior distribution in Bayesian clustering
  - Typically X_k is the probability to belong to cluster k
- Can be represented in several ways:
  - Stick-breaking process
  - Polya-Urn process
  - Chinese restaurant process



Chinese restaurant process

- Chinese Restaurant Process:

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

Handling a stream of documents

- Chinese Restaurant Process:

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

- Useful for sequential modeling (explicit posterior at each step, allows Gibbs sampling)

$$\underbrace{P(n^{th} obs = c|D, history)}_{Posterior} \propto \underbrace{P(D|n^{th} obs = c)}_{Likelihood} \times \underbrace{P(n^{th} obs = c|history)}_{CRP \;\; prior}$$

- Hypothesis: “rich-get-richer”

Variants

- Variants of DP exist:
  - ◇ 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]

→ Most exhibit "rich-get-richer" property

→ All consider counts, none consider temporal dimension

Modeling time as a continuous variable

- Time often "modeled" by sampling observations (DTM [Blei and Lafferty, 2006], TOT [Wang and McCallum, 2006], RCRP [Ahmed and Xing, 2008], DDCRP [Blei and Frazier, 2010], etc.)
  - ◇ Problems: how to slice data, which sampling function use, how to weight observations, etc.
- Whole literature modeling time explicitly: point processes

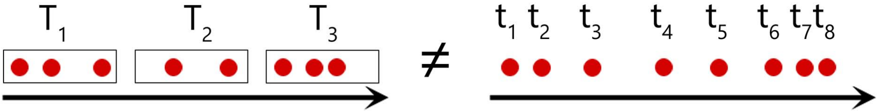


Figure 4: Data sampling/slicing is an approximation

Poisson process

- Poisson processes are characterized by an **intensity** λ.
  - $P(\mathbb{N}(t) = n) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$ = probability for n events to happen within a time t
- Instantaneous PDF of **one** event (or inter-arrival time PDF):

$$f(t) = \frac{P(\mathbb{N}(t) = 1)}{t} = \lambda e^{-\lambda t}$$

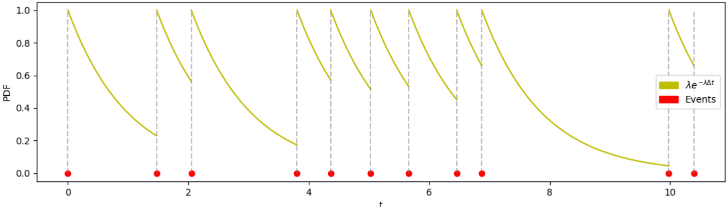


Figure 5: Could model radioactive decay events of atoms whose half-life is 1

Non-homogeneous Poisson process

- $\lambda(t)$ is a function
- $\lambda(t) = \lim_{\Delta t \to 0} \frac{P(\mathbb{N}(t+\Delta t)-\mathbb{N}(t)=1)}{\Delta t}$

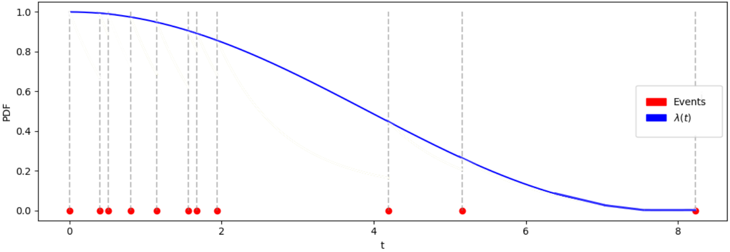


Figure 6: Could model cars arrival at gas station throughout a day

Hawkes process

- Hawkes processes: $\lambda(t)$ depends on past events $\mathcal{H}_t = \{t_i | t_i < t\}$
  $\rightarrow$ "Self-exciting process"
- Typically: $\lambda(t) = \lambda_0 + \sum_{t_i \in \mathcal{H}_t} \phi(t - t_i)$

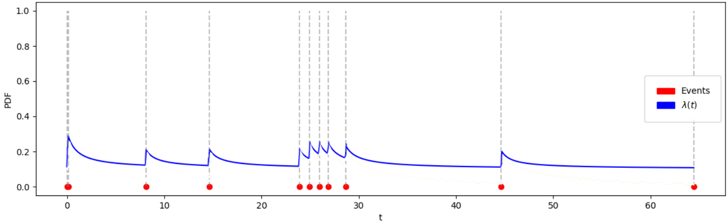


Figure 7: Could model online posting dynamics

Inference

- Log-likelihood $\ell(\lambda)$ fit for data streams:

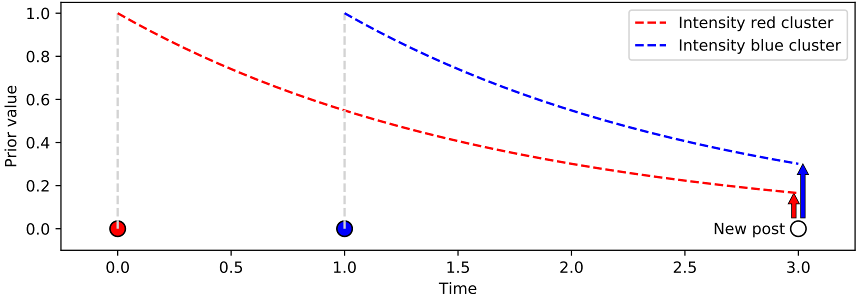
$$\ell(\lambda) = -\int_{t_0}^{t_N} \lambda(t)dt + \sum_{t_i<t_N} \log\lambda(t_i) = \log\lambda(t_1) - \int_{t_0}^{t_1} \lambda(t)dt$$
$$+ \log\lambda(t_2) - \int_{t_1}^{t_2} \lambda(t)dt$$
$$+ \dots$$
$$+ \log\lambda(t_N) - \int_{t_{N-1}}^{t_N} \lambda(t)dt$$

- Convex for certain shapes of $\lambda(t)$ (exp, ray, PL, Gaussian, ...).

Dirichlet-Hawkes process

- [Du et al., 2015]: Dirichlet-Hawkes prior (Bayesian inference)
- Merges Dirichlet priors and Hawkes processes

$$P(\text{cluster}|\text{text, time, H}) \propto \underbrace{P(\text{text}|\text{cluster})}_{\substack{\text{Textual likelihood} \\ \text{(Dirichlet-Multinomial)}}} \times \underbrace{P(\text{cluster}|\text{time, H})}_{\substack{\text{Temporal prior} \\ \text{(Dirichlet-Hawkes)} \\ \downarrow}}$$

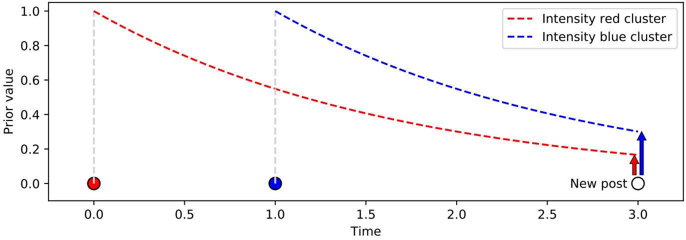


Dirichlet-Hawkes process – Explicit

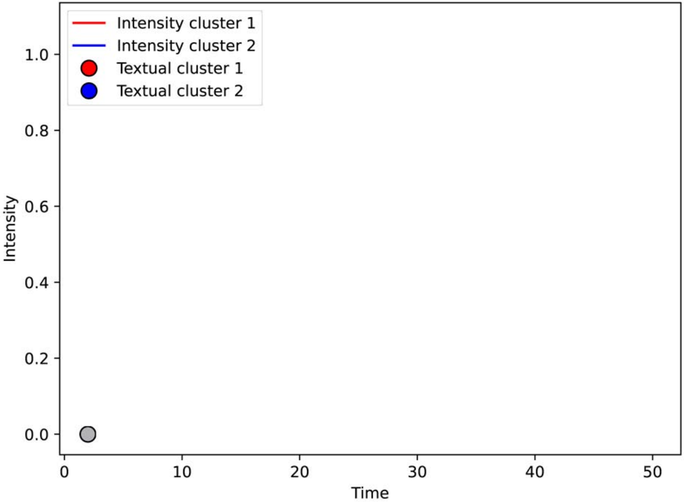
- $P(c|t,\mathcal{H})$: prior probability of cluster c at time t given history $\mathcal{H}$
- $\lambda_c(t)$: intensity of cluster c at time t
- Dirichlet process with counts N_c replaced by $\lambda_c(t)$

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

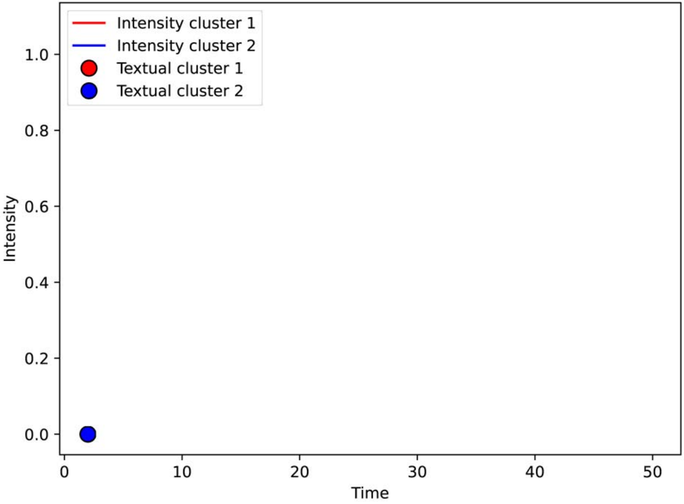
$\downarrow$



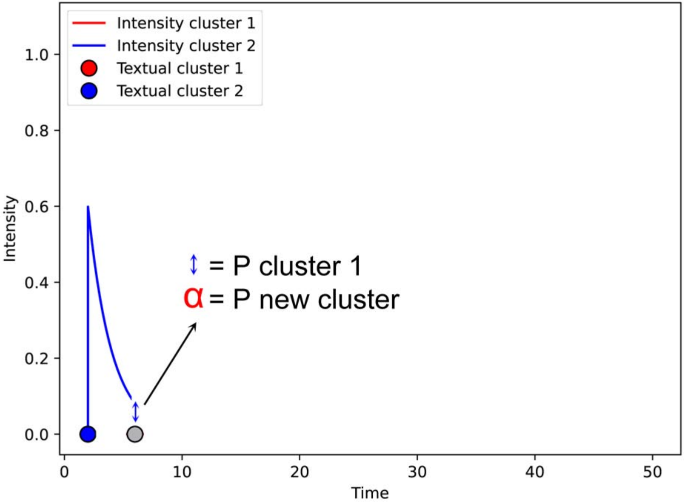
Inference (1 particle)



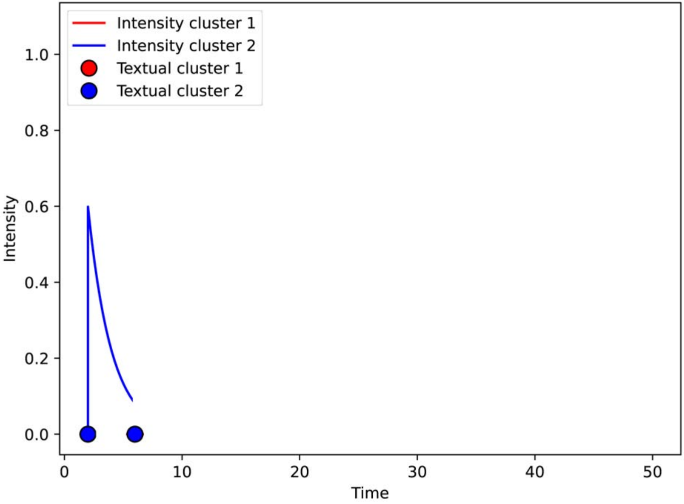
Inference (1 particle)



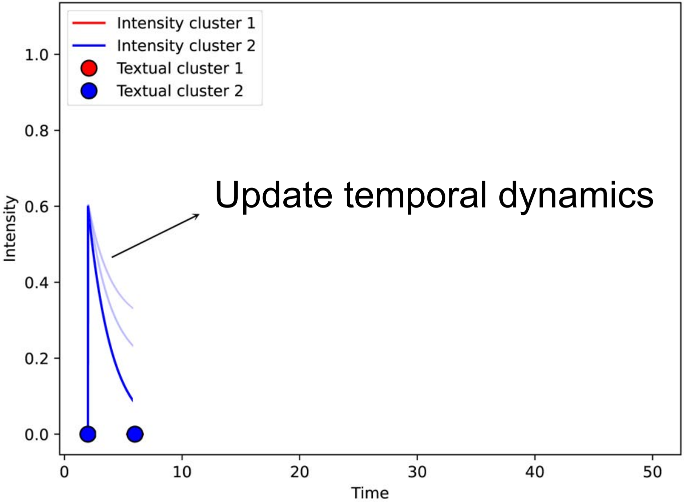
Inference (1 particle)



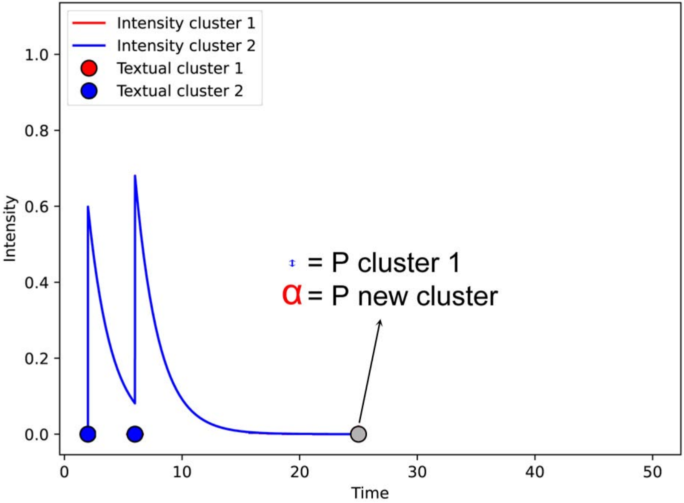
Inference (1 particle)



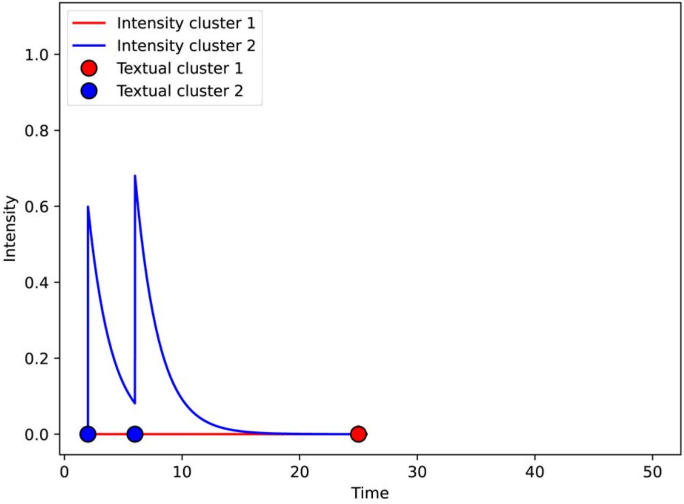
Inference (1 particle)



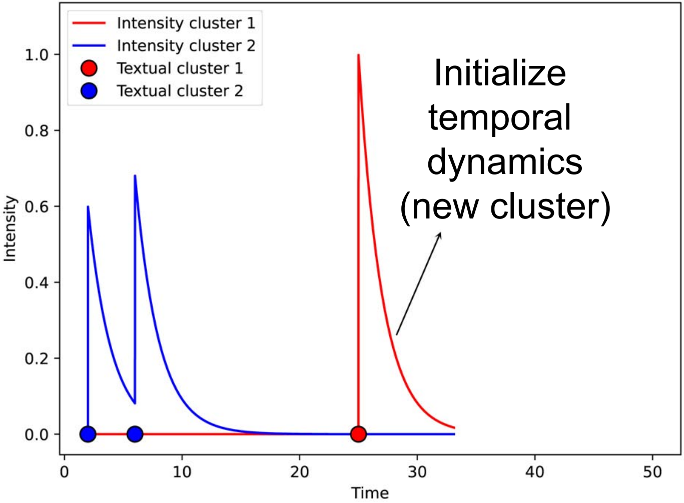
Inference (1 particle)



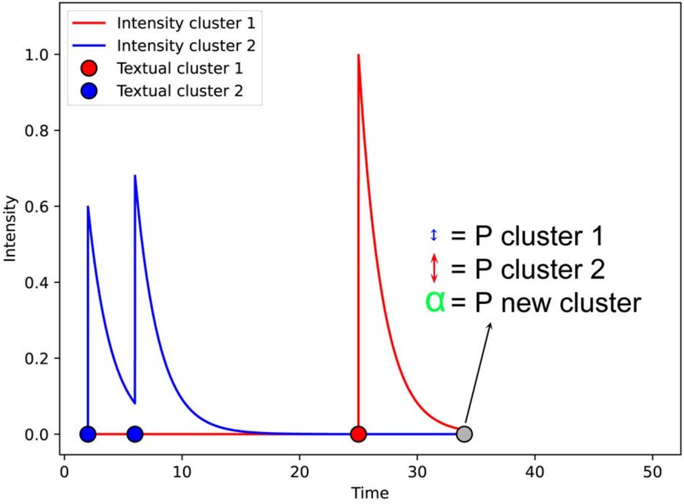
Inference (1 particle)



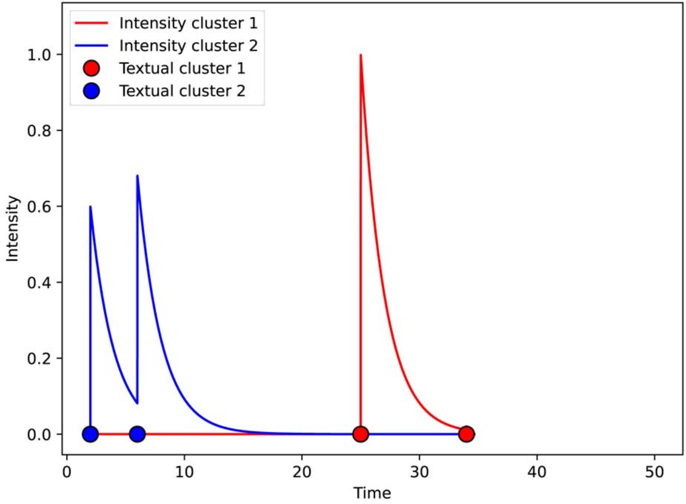
Inference (1 particle)



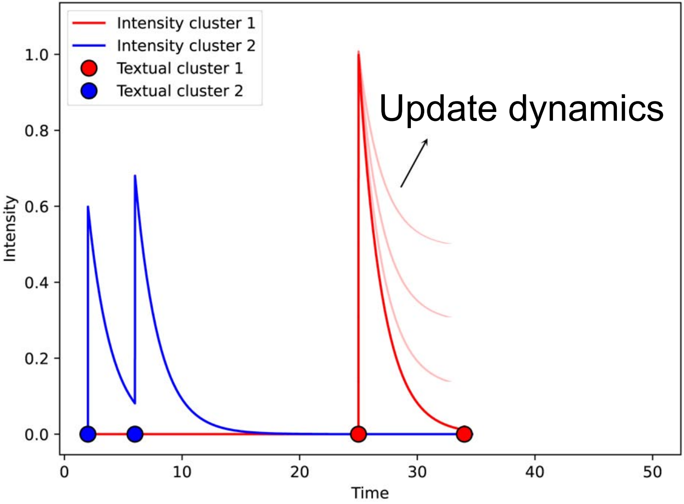
Inference (1 particle)



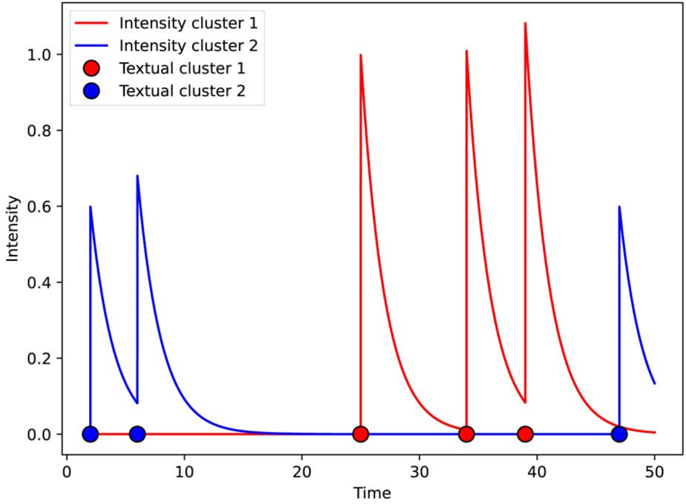
Inference (1 particle)



Inference (1 particle)

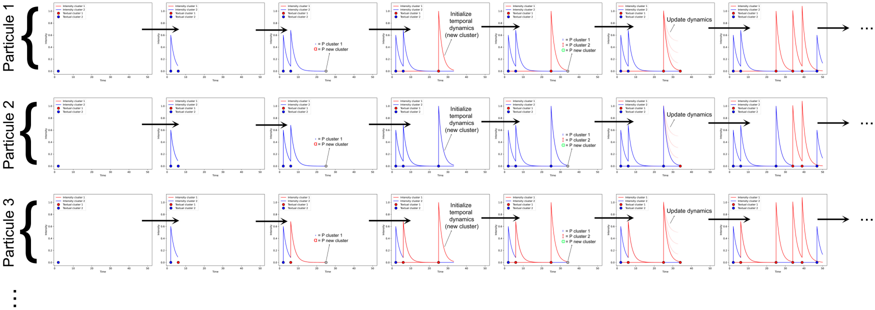


Inference (1 particle)



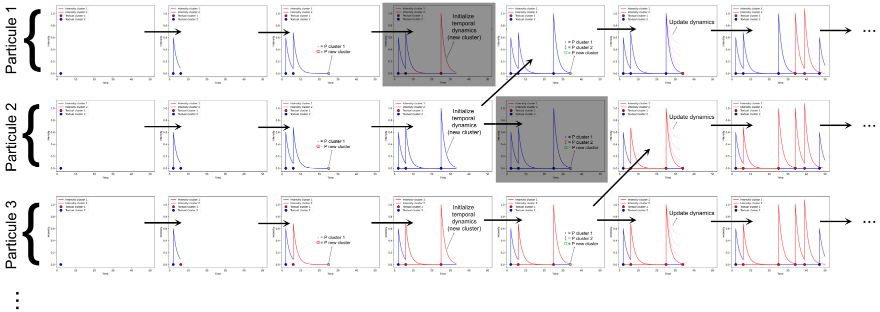
Inference (all particles)

- Run simultaneously on several *particles*

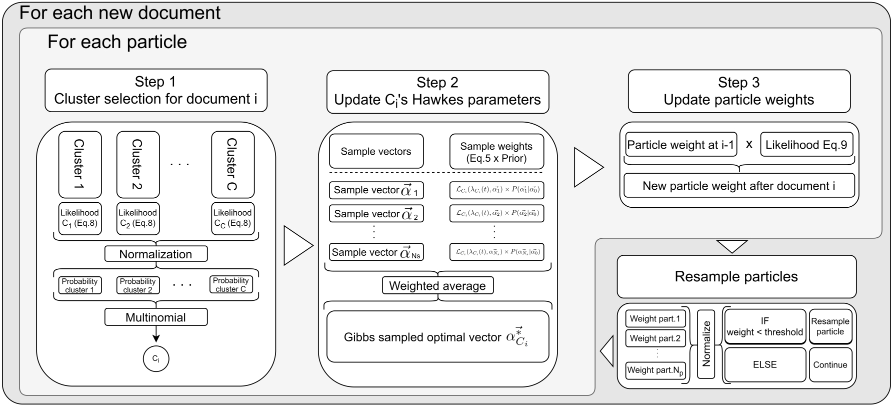


Inference (all particles)

- Discard unlikely particles and replace them by more likely ones



Inference (summarized)



Performances (well-separated)

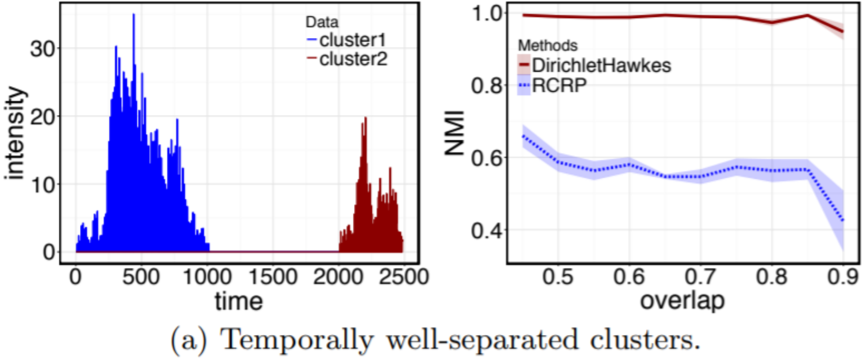


Figure 10: [Du et al., 2015]

Performances ("not" well-separated)

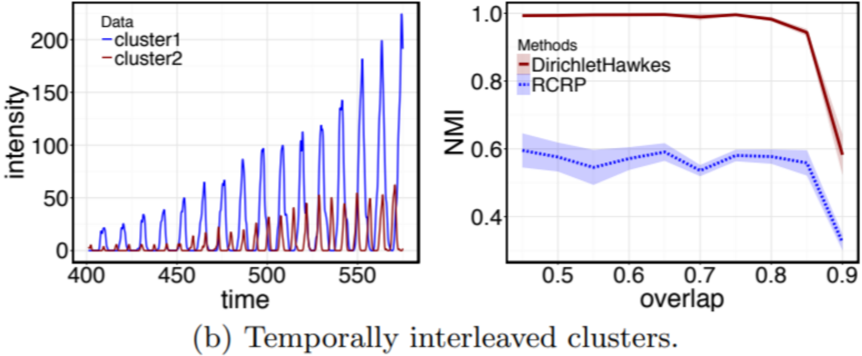


Figure 11: [Du et al., 2015]

Output

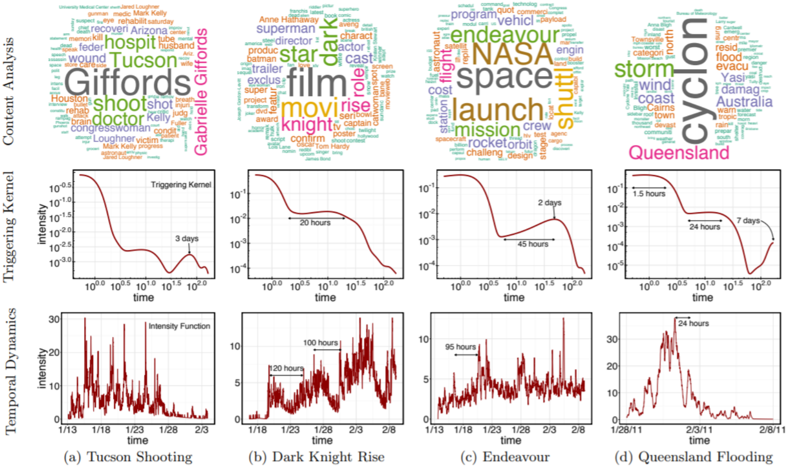


Figure 12: [Du et al., 2015]

Variants

- Numerous variants based on Dirichlet-Hawkes process
  - ◇ Hierarchical (CRF) and Nested (nCRP) extensions of DHP
  - ◇ Multivariate DHP [Zheng et al., 2021]
  - ◇ Not-vanishing DHP prior [Kapoor et al., 2018]

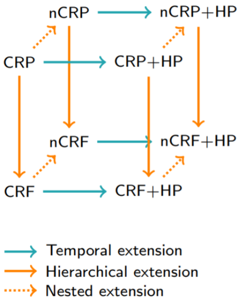


Figure 13: [Kapoor et al., 2018]

BUT!

Dirichlet prior is a choice

- Dirichlet-based priors are an arbitrary choice
  - Other priors are as fit [Welling, 2006]
  - The choice of the prior matters [Wallach et al., 2009]
  - Few variations proposed [Wallach et al., 2010, Pitman and Yor, 1997]
- DP exhibits "rich-get-richer" property
  - Why linear dependence?
  - Why this assumption at all? [Wallach et al., 2010]

Powered Dirichlet process

- Powered Chinese Restaurant Process:

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

- ◇ r<0: "rich-get-poorer"
- ◇ r=0: "rich-get-no-richer" (Uniform Process)
- ◇ 0<r<1: "rich-get-less-richer"
- ◇ r=1: "rich-get-richer" (Dirichlet Process)
- ◇ $r = \frac{\log(N_k - \beta)}{\log(N_k)}$: "rich-get-richer" (Pitman-Yor Process)
- ◇ r>1: "rich-get-more-richer"

PDP impact

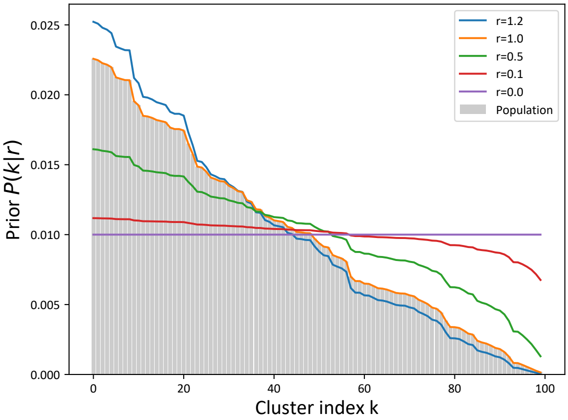
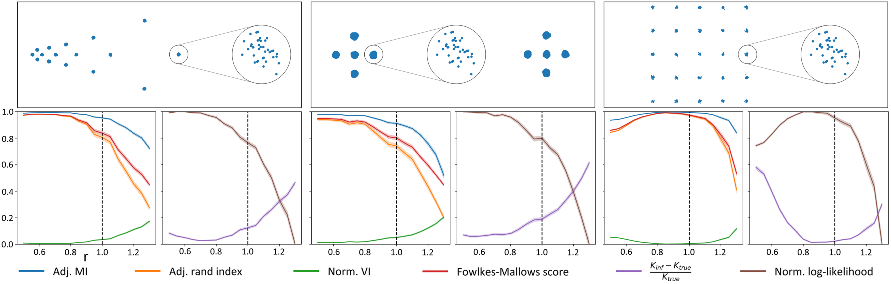


Figure 14: Prior probability for each of 100 clusters whose population is known (grey bars) w.r.t. r

Results

- Use as prior for IGMM
- DP not always the best prior



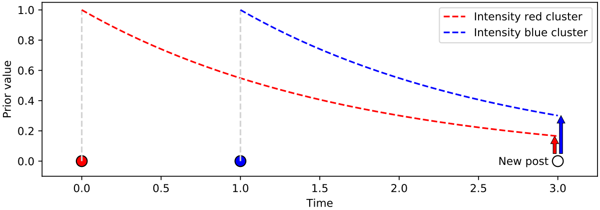
PDP into DHP

- Powered priors: controlling the informativeness of the prior
  - ◇ PDP: strength of the "rich-get-richer" hypothesis
  - ◇ PDHP: strength of the temporal dependence hypothesis
- PDHP [Poux-Médard et al., 2021]:

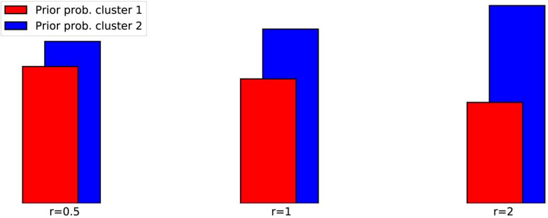
$$\underbrace{P(c|t,\mathcal{H},r)}_{\text{PDHP prior}} = \begin{cases} \frac{\lambda_c(t)^r}{\alpha_0+\sum_k \lambda_k(t)^r} & \text{if c = 1, ..., K} \\ \frac{\alpha_0}{\alpha_0+\sum_k \lambda_k(t)^r} & \text{if c = K+1} \end{cases}$$

- Generalization:
  - ◇ Uniform process: $r = 0$ (only textual information)
  - ◇ Dirichlet-Hawkes process: $r = 1$ (temporal and textual information)
  - ◇ Deterministic Hawkes process: $r \to \infty$ (only temporal information)

Effect of r



$\downarrow$



Changes induced by PDHP

$$P(\text{cluster}|\text{text, time}) \propto \underbrace{P(\text{text}|\text{cluster})}_{\text{Textual likelihood}} \times \underbrace{P(\text{cluster}|\text{time, r, history})}_{\text{PDHP temporal prior}}$$

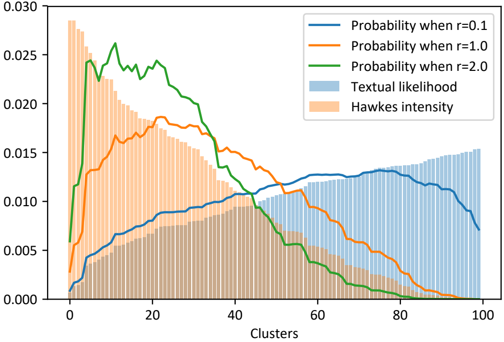
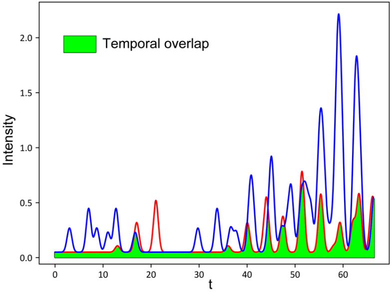


Figure 15: [Poux-Médard et al., 2021]

Why is it relevant - Overlaps

- Often, a piece of information is more informative than the other:
  - ◇ Twitter: short texts (few textual information) but informative cascade dynamics (helpful temporal information)
- Happens often because of overlaps:



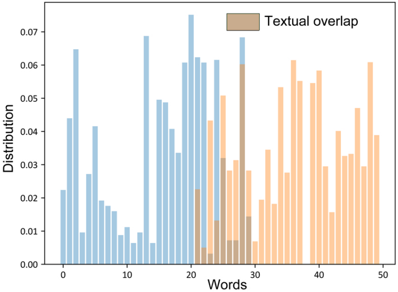


Figure 16: [Poux-Médard et al., 2021]

Results for various overlaps

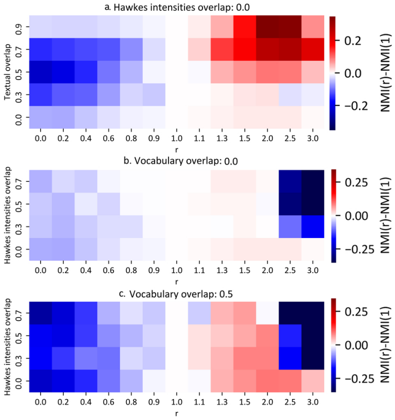
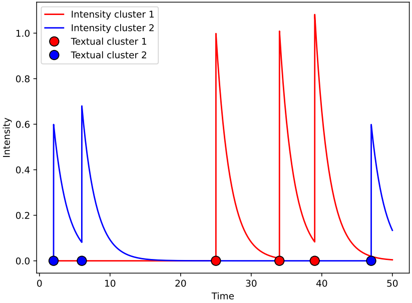


Figure 17: [Poux-Médard et al., 2021]

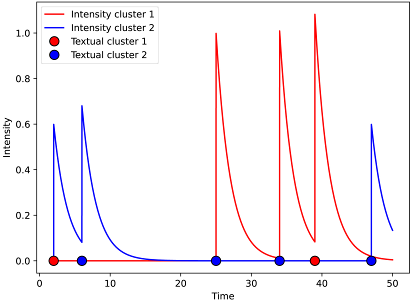
- PDHP adapts to various situations better than DHP:
  - Large textual overlap
  - Large temporal overlap
  - No overlap
- Up to +0.3 NMI in our case

Why is it relevant - Decorrelations

- Decorrelations:
  - Ex: influent journal publishing on a topic does not have same dynamics as less influent one on the same topic



$\rightarrow$



Results for various decorrelations

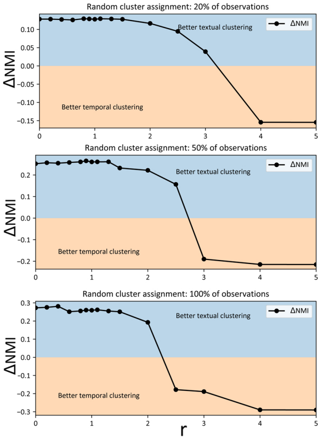


Figure 18: [Poux-Médard et al., 2021]

- PDHP retrieves either temporal or textual clusters
  - Small r: good textual clusters
  - Large r: good temporal clusters

Reddit r/news - Typical output

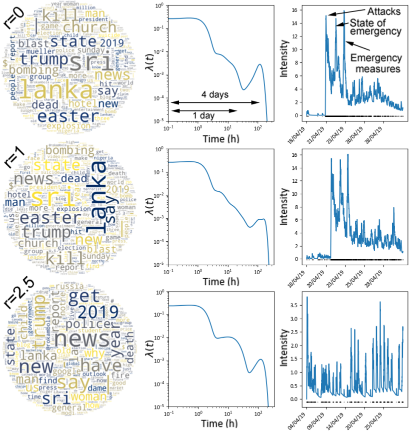


Figure 19: [Poux-Médard et al., 2021]

- Real world data: r/news
- Different clusters and dynamics for different r
  - Small r: similar vocabulary
  - Large r: specific dynamics

Reddit r/news, r/TodayILearned, r/AskScience - Some metrics

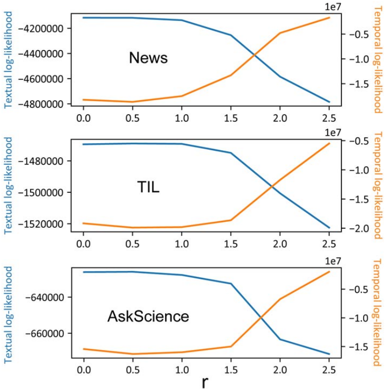


Figure 20: Textual and temporal likelihood vs r [Poux-Médard et al., 2021]

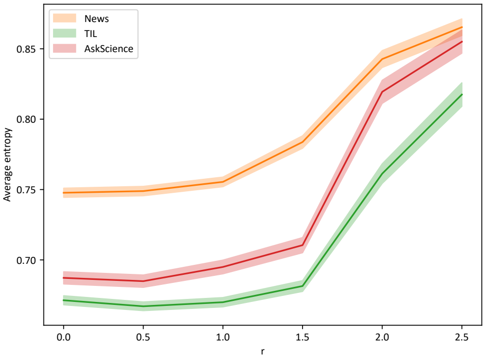


Figure 21: Entropy of textual clusters: sharper textual clusters for low r [Poux-Médard et al., 2021]

Structure matters!

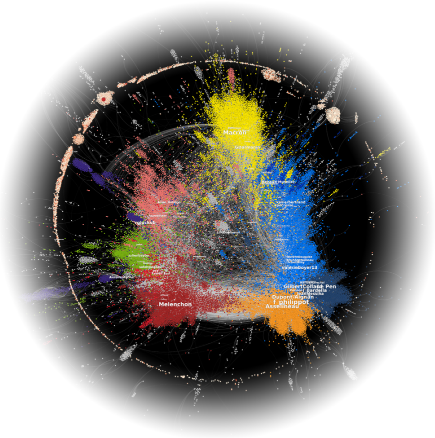


Figure 22: A sample from the Twitter structure (Politoscope [Gaumont et al., 2018])

Why (P)DHP is incomplete

- DHP prior accounts for time but not structure
  - ◇ Infers aggregated dynamics
  - ◇ Misses the structural aspect: discussions are not the same among different groups

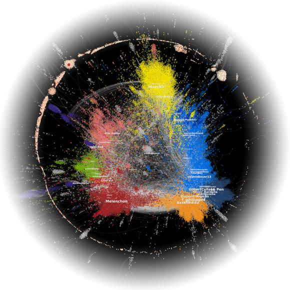
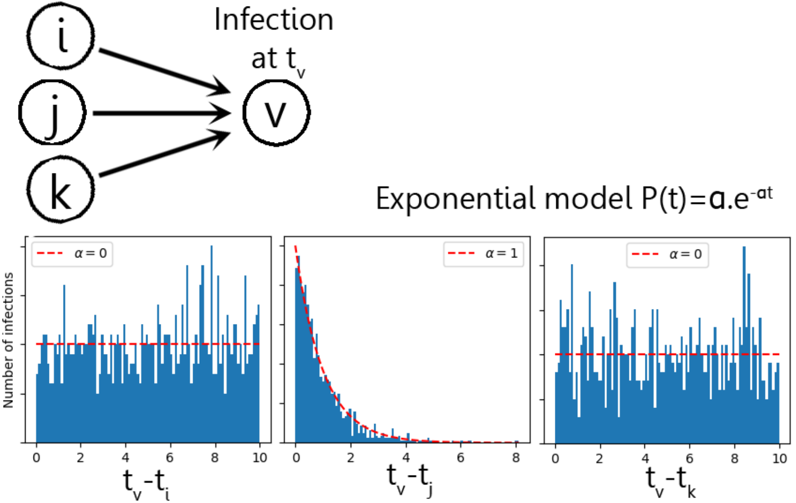


Figure 23: A sample from the Twitter structure (Politoscope [Gaumont et al., 2018])

Network inference



Network inference – Literature

- Several works on network inference using survival analysis:
  - ◇ NetRate [Gomez-Rodriguez et al., 2011]
  - ◇ InfoPath [Gomez-Rodriguez et al., 2013a]
  - ◇ KernelCascade [Du et al., 2012]
  - ◇ MoNet [Wang et al., 2012]
  - ◇ TopicCascade [Du et al., 2013]
- They are all special cases of [Gomez-Rodriguez et al., 2013b]
  - ◇ Bridges the gap between survival analysis and point processes
  - ◇ Formulates each of previous models as a counting point process

Point process

- Network inference naturally embeds into point processes literature
  → We can derive a temporal *and* structural Bayesian prior

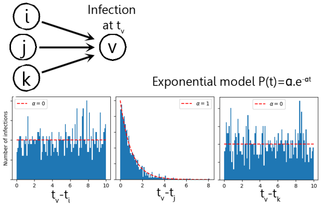


Figure 24: Survival process

Both are

point

processes

$<\approx>$

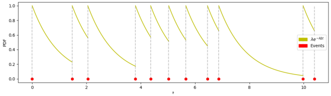


Figure 25: Hawkes process

Temporal and structural prior

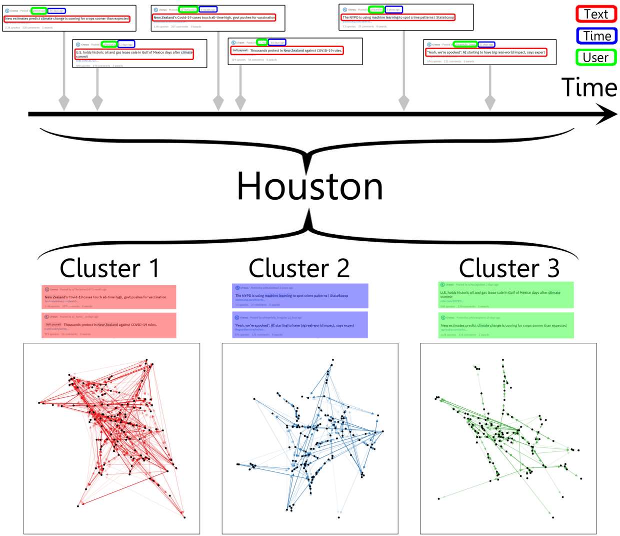
- Houston: **H**eterogeneous **O**nline **U**ser-**T**opic **N**etwork inference
- Prior on cluster membership C_i of observation i observed on node u at time t given history $\mathcal{H}$ and cluster-dependent networks A:

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

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

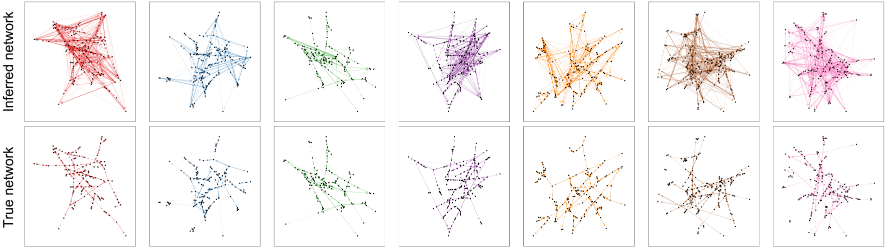
$$= \begin{cases} \dfrac{\text{Strength of incoming edges of cluster/subnetwork k at time t}}{\text{Normalizing term}} & \text{if k = 1, ..., K} \\ \dfrac{\text{Probability of a new cluster/subnetwork k+1 at time t}}{\text{Normalizing term}} & \text{if k = K+1} \end{cases}$$

Task



Results – Synthetic

- We simulate the spread of documents drawn from 5 topics, each with its own vocabulary and subnetwork

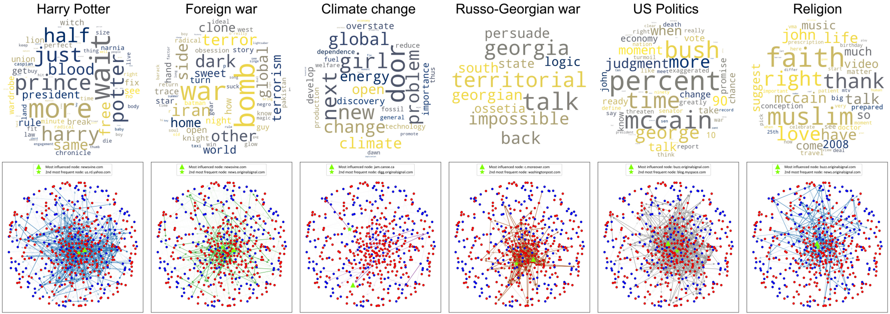


Numerical results

		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

Results – Real world

- Memetracker data (2009)



Conclusion

- Dirichlet and Hawkes process have an old and separate history
  - ◇ Only recently (2015) they have been brought together
  - ◇ Their reunion launched a new branch of inductive machine learning
- The number of extensions based on Dirichlet-Point-Processes might be enormous, because we touched core concepts of machine learning
  - ◇ Dirichlet processes (PDP): could be used to redefine hierarchical DP, nested DP, or any models built on them (LDA, SBMs, among others)
  - ◇ Point processes (Poisson, Hawkes, Survival/Counting, etc.): the new possibility to merge them with DP could lead to a potentially infinite number of different Dirichlet-Point-Process priors.
- We presented 2 of such extensions:
  - ◇ PDP+HP $\rightarrow$ PDHP (flexible temporal prior)
  - ◇ DP+Survival $\rightarrow$ Houston (temporal+structural prior)

Thanks for your attention!

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

$\times$

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

$=$

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

Bibliography I

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

[Blei and Frazier, 2010] Blei, D. M. and Frazier, P. (2010).
Distance dependent chinese restaurant processes.
In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 87–94, Madison, WI, USA. Omnipress.

[Blei and Lafferty, 2006] Blei, D. M. and Lafferty, J. D. (2006).
Dynamic topic models.
In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, USA. Association for Computing Machinery.

Bibliography II

[Du et al., 2015] 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 et al., 2012] Du, N., Song, L., Smola, A., and Yuan, M. (2012).
Learning networks of heterogeneous influence.
NIPS, 4:2780–2788.

[Du et al., 2013] Du, N., Song, L., Woo, H., and Zha, H. (2013).
Uncover topic-sensitive information diffusion networks.
In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 31 of *JMLR Workshop and Conference Proceedings*, pages 229–237. JMLR.org.

Bibliography III

[Gaumont et al., 2018] Gaumont, N., Panahi, M., and Chavalarias, D. (2018).
Reconstruction of the socio-semantic dynamics of political activist twitter networksâĂŤmethod and application to the 2017 french presidential election.
PLOS ONE, 13(9):1–38.

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

[Gomez-Rodriguez et al., 2013a] Gomez-Rodriguez, M., Leskovec, J., and Schoelkopf, B. (2013a).
Structure and dynamics of information pathways in online media.
WSDM.

Bibliography IV

[Gomez-Rodriguez et al., 2013b] Gomez-Rodriguez, M., Leskovec, J., and Schölkopf, B. (2013b).
Modeling information propagation with survival theory.
In *ICML*, volume 28, pages III–666–III–674.

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

[Pitman and Yor, 1997] 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.

Bibliography V

[Poux-Médard et al., 2021] Poux-Médard, G., Velcin, J., and Loudcher, S. (2021).
Powered hawkes-dirichlet process: Challenging textual clustering using a flexible temporal prior.
ICDM.

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

[Teh et al., 2006] 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 et al., 2010] Wallach, H., Jensen, S., Dicker, L., and Heller, K. (2010).
An alternative prior process for nonparametric bayesian clustering.
In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 892–899. JMLR.

Bibliography VI

[Wallach et al., 2009] Wallach, H., Mimno, D., and McCallum, A. (2009).
Rethinking lda: Why priors matter.
In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

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

[Wang and McCallum, 2006] Wang, X. and McCallum, A. (2006).
Topics over time: A non-markov continuous-time model of topical trends.
In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 424–433, New York, NY, USA. Association for Computing Machinery.

Bibliography VII

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

[Zheng et al., 2021] Zheng, P., Yuan, S., and Wu, X. (2021).
Using dirichlet marked hawkes processes for insider threat detection.
Digital Threats: Research and Practice, 3(1).