Dirichlet-Survival Process: Scalable Inference of Topic-Dependent Diffusion Networks

Gaël Poux-Médard, Julien Velcin, Sabine Loudcher

Dirichlet-Survival Process: Scalable Inference of Topic-Dependent Diffusion Networks

Gaël Poux-Médard[1 [0000−0002−0103−8778]], Julien Velcin[1 [0000−0002−2262−045X]], and Sabine Loudcher[1 [0000−0002−0494−0169]]

[1] Université de Lyon, Lyon 2, ERIC UR 3083, 5 avenue Pierre Mendès France, F69676 Bron Cedex, France
gael.poux-medard@univ-lyon2.fr
julien.velcin@univ-lyon2.fr
sabine.loudcher@univ-lyon2.fr

Abstract. Information spread on networks can be efficiently modeled by considering three features: documents' content, time of publication relative to other publications, and position of the spreader in the network. Most previous works model up to two of those jointly, or rely on heavily parametric approaches. Building on recent Dirichlet-Point processes literature, we introduce the Houston (Hidden Online User-Topic Network) model, that jointly considers all those features in a non-parametric unsupervised framework. It infers dynamic topic-dependent underlying diffusion networks in a continuous-time setting along with said topics. It is unsupervised; it considers an unlabeled stream of triplets shaped as *(time of publication, information's content, spreading entity)* as input data. Online inference is conducted using a sequential Monte-Carlo algorithm that scales linearly with the size of the dataset. Our approach yields consequent improvements over existing baselines on both cluster recovery and subnetworks inference tasks.

Keywords: Spreading process · Network Inference · Clustering · Bayesian Nonparametrics

1 Introduction

1.1 Overview of the contribution

Over the last decades, information spread patterns have become more and more complicated. The volume of data that flows on social networks keeps increasing every day that passes, and results in complex diffusion processes that can be described by many factors. However, recent advances suggest that documents complex diffusion processes can be efficiently modeled considering only three variables: their publication date (when), the publisher (who) and their semantic content (what). The idea of considering these three factors is not novel. However, most of the models that tackle diffusion problems tend to consider up to two of these, but seldom the three parameters.

We introduce the Houston model, that tackles the problem by jointly inferring clusters of textual documents spreading online *and* the subnetworks they spread on. Our method builds on recent Dirichlet-Point processes advances [9, 18, 23, 24]. To the best of our knowledge, it is the first model that considers semantic content, publication dynamics and the network of spreading documents in an online, non-parametric and unsupervised way.

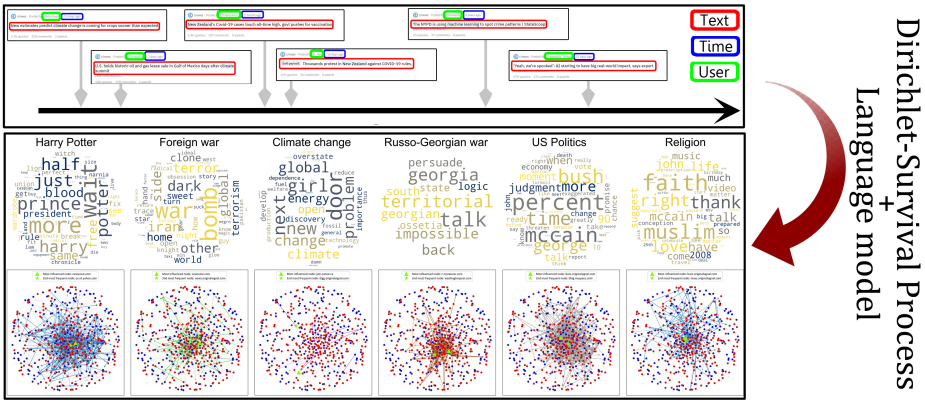

Fig. 1. From a stream of textual documents, we model the underlying topic-dependent diffusion subnetworks. Inference is unsupervised, non-parametric and conducted online, meaning data is processed sequentially. Results in the bottom row come from the application of our method to the Memetracker dataset [17]. Nodes colors represent traditional medias (red) and blog (blue).

1.2 Related works

It has been underlined on several occasions that efficiently modeling information diffusion involves accounting for the network's structure [16, 22], publication times [8, 12] and documents' content [10, 15]. Some approaches consider sequentially all three factors. Typically, they first infer topics based on documents content, and only then they use this information to infer the latent diffusion subnetworks [7, 10, 15, 26, 28, 29]. The work the closest to ours [4] is, to our knowledge, the only one that jointly models documents' content, dynamics and structure. It develops an unsupervised topic-dependent network inference method. The approach breaks down the topic-aware diffusion into two factors: each node is assumed to have a given sensitivity to a topic, and a certain authority on them. Given this assumption, the authors develop a parametric prior on the probability for a diffusion cascade to belong to a given topic. The textual content (or side information) is then accounted for using a homogeneous Poisson textual model [19], combined with the above prior. The model is optimized using

an EM algorithm. However, the optimization algorithm is not designed for online optimization –data cannot be added sequentially–, and topics optimization is parametric –the number of topics must be provided.

2 Model

2.1 Background

To answer these limitations, we build a Dirichlet-Survival process that can be used as a non-parametric Bayesian prior for online inference. The Dirichlet-Survival prior is created by merging Dirichlet processes with Point processes. The method has been explored by combining Hawkes processes to several variants of Dirichlet processes (hierarchical [18], mixed membership [27], powered [23], multivariate [24]). However, no work considered the combination with other point processes than the Hawkes process. Our approach using Survival analysis explores this new connection; it allows us to design an optimization algorithm (Sequential Monte Carlo) for online non-parametric topics-aware diffusion subnetworks inference (the number of topics/subnetworks does not have to be chosen in advance).

In [14], the authors show that a large part of the literature on underlying diffusion network inference [8, 12–14, 21, 28] can be expressed as special cases of a counting point process. The method allows to infer dynamic underlying diffusion networks using convex optimization tools.

2.2 Dirichlet process and Survival analysis

Dirichlet process The Dirichlet process is used as a non-parametric prior distribution over clusters in many clustering algorithms. It can be written as follows:

$$P(s_i = k | \{s_m\}_{m=1,...,n-1}, \alpha_0) = \begin{cases} \frac{N_k}{\alpha_0 + \sum_k^K N_k} & \text{if k = 1, ..., K} \\ \frac{\alpha_0}{\alpha_0 + \sum_k^K N_k} & \text{if k = K+1} \end{cases} \quad (1)$$

where s_i is a variable that represents the cluster of the i^{th} observation, $N_k = |\{s_i | s_i = k\}_{i=1,...,n-1}|$ the population of cluster k, K the total number of non-empty clusters and α_0 a concentration hyper-parameter. The choice of $K + 1$ means a new cluster is opened and K in increased by 1. Note that references [23, 24] use the powered version of this process [25].

Network inference model The edges of topic-dependent networks are inferred using the NetRate model [12], which is part of a broad literature on underlying spreading networks inference [10, 12–14, 28]. In particular in [14], the authors demonstrate that all these models can be expressed as special cases of a counting point process. These processes take a collection of independent timestamped diffusion cascades $\vec{c} = \{(u_i^c, t_i^c)\}_i$ as input, where u_i^c is the node on which the

i^{th} event occurred and t_i^c the time at which it happened in cascade c. The process is entirely characterized by a hazard function $H(t_i^c|t_j^c, \alpha_{u_j^c,u_i^c})$, which is the instantaneous infection rate of u_i^c at time t_i^c by u_j^c previously infected at time t_j^c, given it infection did not happen before t_i. In this paper, we express the hazard function as a constant $H(t|t_i, \alpha) = \alpha$, implying by definition that the probability of an event *non* happening before a time t given t_i decays exponentially as $e^{-\alpha(t-t_i)}$. The associated convex likelihood of α can be found in [12] (Eq.7).

2.3 Dirichlet-Survival process

In [9] the authors define the Dirichlet-Hawkes process by replacing the integer counts in Eq.1 by the intensity of a Hawkes process. It can be interpreted as replacing integers counts in Dirichlet Processes by non-integer time-dependent counts, encoded by the intensity of the point process. Here, we consider the hazard rate of the NetRate model instead to account for networks structure. Each node is associated to its own temporal point process, and counts are replaced by the number of times any neighbour has been infected, weighted according to time and to edges strength. Using the methodology introduced in [9] and substituting the Hawkes process by the hazard rate of a survival model [14], we make a yet unexplored bridge between Dirichlet processes and Survival analysis. We remind that [14] reformulates the work of [8, 12, 13, 28] in terms of Survival analysis and associated counting processes; we settled on using NetRate here, but any of these models would fit as well in our approach. The point process nature of survival analysis discussed in [14] makes this extension sound with respect to previous works on Dirichlet-Point processes [9, 18, 23, 27].

Let $\mathbf{A}^{(k)}$ be the adjacency matrix of the subnetwork associated to cluster k, whose entries are $\alpha_{i,j}^{(k)}$. We define $(u_j^c, t_j^c)^{(k)}$ as an event of cascade c observed on node u_j at t_j attributed to subnetwork $A^{(k)}$. We write the history of events in cascade c attributed to the subnetwork k as $\mathcal{H}_{i,c}^{(k)} = \{(u_j^c, t_j^c)^{(k)}\}_{j:t_j<t_i}$. We note $\mathcal{H}_{\mathbf{i,c}} = \{\mathcal{H}_{i,c}^{(k)}\}_k$ and $\mathbf{A} = \{\mathbf{A}^{(k)}\}_k$. We consider a new event from cascade c observed on node u_i^c at time t_i^c. At this point, the new event is not yet associated to any subnetwork. We write the Dirichlet-Survival prior probability for the new event to belong to subnetwork k:

$$P(s_i = k|\mathcal{H}_{\mathbf{i,c}}, \mathbf{A}, \boldsymbol{\lambda_0}) = \begin{cases} \frac{\lambda_0^{(k)} + \sum_{\mathcal{H}_{i,c}^{(k)}} H(t_i^c|t_j^c, \alpha_{u_j,u_i}^{(k)})}{\lambda_0^{(K+1)} + \sum_k^K \lambda_0^{(k)} + \sum_{\mathcal{H}_{i,c}^{(k)}} H(t_i^c|t_j^c, \alpha_{u_j,u_i}^{(k)})} & \text{if k = 1, ..., K} \\ \frac{\lambda_0^{(K+1)}}{\lambda_0^{(K+1)} + \sum_k^K \lambda_0^{(k)} + \sum_{\mathcal{H}_{i,c}^{(k)}} H(t_i^c|t_j^c, \alpha_{u_j,u_i}^{(k)})} & \text{if k = K+1} \end{cases} \tag{2}$$

We introduced a new parameter $\boldsymbol{\lambda_0} = \{\lambda_0^{(k)}\}_{k=1,...,K+1}$, which translates the probability for a new observation not to have been triggered by any neighbour. It represents the probability that an event of cluster k is exogenous [15, 20].

The Dirichlet-Survival prior is coupled to a sequential language model. For simplicity, we consider the bag-of-words Dirichlet-Multinomial model, as in [9,

18, 23]; note that more refined sequential language models are also fit to our approach (Dynamic Topic Model [6], online LDA [3], online PLSA [5], etc.).

The input data is a stream of events. Each event takes the form of a triplet (u_i^c, t_i^c, v_i^c), where c is the cascade an event has been observed in, u_i^c is the node corresponding to the event, t_i^c is its publication time, and v_i^c represents its textual content (e.G. words in a tweet or in a news article). By combining the Dirichlet-Survival prior to the textual likelihood, we get the posterior distribution of the i^{th} observation belonging to cluster (or subnetwork) k as:

$$P(s_i|v_i^c, \mathbf{N}, \mathcal{H}_{\mathbf{i,c}}, \mathbf{A}, \theta_0, \lambda_{\mathbf{0}}) \propto \underbrace{P(v_i^c|s_i, \mathbf{N}, \theta_0)}_{\text{Dirichlet-Multinomial}} \times \underbrace{P(s_i|\mathcal{H}_{\mathbf{i,c}}, \mathbf{A}, \lambda_{\mathbf{0}})}_{\text{Dirichlet-Survival prior (Eq.2)}} \quad (3)$$

where $\vec{N}$ contains the words counts within each cluster, v_i^c contains the words count in document i, and θ_0 the concentration parameter of the model.

Finally, inference is conducted using a Sequential Monte Carlo algorithm similar to [9, 18, 23]. We perform several parallel runs on the same data stream. Within each run, each new observation in the stream is assigned to a cluster according to Eq. 3. The adjacency matrix $\mathbf{A}$ is then updated by optimizing the convex likelihood associated to the NetRate point process (Eq. 7 in [12]). Finally, we compute the likelihood of the language model for each run; runs that have a likelihood lesser than a threshold are discarded and replaced by more likely ones. The process is repeated until the end of the data stream. According to this algorithm, Eq. 1, and introducing a cutoff on the exponential hazard function (observations older than a time t_{old} are ignored), the optimization runs in $\mathcal{O}(N_{obs}N_{runs}(N_{nodes} + K))$ where N_{part} is the number of particles, N_{nodes} is the maximum network size and K the number of clusters (typically $N_{runs} \ll K \ll N_{nodes}$). Inference hence scales linearly with the size of the dataset.

We point out that the Dirichlet-Survival process is not about refining complex diffusion models such as [4, 7, 26]. Instead, it introduces a different angle for tackling content-aware diffusion problems. This new angle allows for unsupervised, non-parametric and online inference.

3 Experiments

3.1 Data and experimental setup

All data, codes and results are available in open access [1]. We consider 3 different network types of 500 nodes each: power-law (**PL**) [2], random Erdös-Renye (**ER**) [11] and a real network of hyperlinks between political blogs (**Blogs**) [1]. From each network, we sample 5 subnetworks of 250 nodes and assign random weights α between 0 and 1 to their edges. Each of the generated subnetworks is used to propagate one given cluster of information. We then simulate infection cascades on each subnetwork according to the exponential NetRate model. Finally, we associate 5 words drawn from a vocabulary of size 100 to each so-generated

[1] `https://github.com/GaelPouxMedard/HOUsToN`

event according to its associated subnetwork (or cluster). We generate a total of 55,000 events $\{(u_i^c, t_i^c, v_i^c)\}_{i,c}$ for each network.

Our hyperparameters are $\theta_0 = 0.1$ and $\lambda_0^{(k)} = 0.001\ \forall k$. The SMC algorithm considers 4 parallel runs. We consider a constant hazard rate $H(t_i|t_j, \alpha_{j,i}) = \alpha_{j,i}$, so the probability of a new event *not* happening decays exponentially with time.

Table 1. Results on clusters (NMI, ARI) and edges (AUC-ROC, F1, MAE) retrieval.

		Houston	NRxDM	DHP	NetRate
PL	NMI	**0.809**	0.669	0.449	-
	ARI	**0.688**	0.330	0.063	-
	AUC-ROC	**0.807**	0.719	-	0.731
	F1	**0.199**	0.106	-	0.005
	MAE	**0.267**	0.338	-	0.460
ER	NMI	**0.787**	0.711	0.638	-
	ARI	**0.631**	0.488	0.411	-
	AUC-ROC	**0.849**	0.800	-	0.659
	F1	**0.263**	0.176	-	0.005
	MAE	**0.229**	0.278	-	0.481
Blogs	NMI	**0.750**	0.668	0.372	-
	ARI	**0.609**	0.365	0.023	-
	AUC-ROC	0.701	0.613	-	**0.710**
	F1	**0.168**	0.087	-	0.005
	MAE	**0.374**	0.444	-	0.499

3.2 Results

We compare to 3 similar baselines used as ablation tests: **Dirichlet-Hawkes process (DHP)** [9] clusters textual data by using temporal dynamics, and does not consider structure; **NetRate** [12] infers a dynamic network based on observed cascades without considering their content; **NetRate x Dirichlet-Multinomial (NRxDM)** first uses textual information to infer clusters, and only then infers the underlying subnetwork for each cluster, in the same fashion as [10, 15]. When applicable, we evaluate on a classification task (scores NMI and ARI with respect to the clusters used for data generation) and a network inference task (AUC-ROC, F1 and MAE on the true edges, same metrics as in [12]).

We see in Table 1 that Houston consistently outperforms methods that do not consider jointly text, time and structure of the network. To summarize, NRxDM only considers textual information to build clusters, making the network inference miss a great deal of temporal and structural information. DHP considers textual information and temporal dynamics, but misses the structural information. NetRate does not consider textual data and infers the network based on temporal dynamics only. Houston bridges the gap between these models, by making a joint use of textual, temporal and structural information.

As an illustration of what Dirichlet-Processes can yield on real-world data, we draft its application to the Memetracker dataset [17] in Fig. 1 (bottom). We retrieve the diffusion network associated to meme clusters and observe diverse spreading dynamics. Topics spread in distinct parts of the global network, and mostly do so through a reduced set of densely connected nodes, as shown in [13].

4 Conclusion

In this paper, we propose the Dirichlet-Survival process as an alternative way to jointly model textual, temporal and structural information in spreading processes. Ablation tests demonstrate the relevance of the proposed approach. As a prior, the Dirichlet-Survival process can add a dynamic network dimension to any sequential Bayesian model; it could be coupled to models that account for any type of clustering (e.g. images, time series, labels), or simply more refined language models. Its introduction opens new perspectives on traditional machine learning problems, including topic-dependent spreading processes on networks.

References

1. Adamic, L.A., Glance, N.: The political blogosphere and the 2004 u.s. election: Divided they blog. In: Proceedings of the 3rd International Workshop on Link Discovery. p. 36–43. LinkKDD '05, Association for Computing Machinery, New York, NY, USA (2005)
2. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. **74**, 47–97 (Jan 2002)
3. AlSumait, L., Barbará, D., Domeniconi, C.: On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking pp. 3–12 (2008). https://doi.org/10.1109/ICDM.2008.140
4. Barbieri, N., Manco, G., Ritacco, E.: Survival factorization on diffusion networks. Machine Learning and Knowledge Discovery in Databases pp. 684–700 (2017). https://doi.org/10.1007/978-3-319-71249-9_41
5. Bassiou, N.K., Kotropoulos, C.L.: Online plsa: Batch updating techniques including out-of-vocabulary words. IEEE Transactions on Neural Networks and Learning Systems **25**(11), 1953–1966 (2014). https://doi.org/10.1109/TNNLS.2014.2299806
6. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. p. 113–120. ICML '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1143844.1143859
7. Choudhari, J., Dasgupta, A., Bhattacharya, I., Bedathur, S.: Discovering topical interactions in text-based cascades using hidden markov hawkes processes pp. 923–928 (2018). https://doi.org/10.1109/ICDM.2018.00112
8. Du, N., Song, L., Smola, A., Yuan, M.: Learning networks of heterogeneous influence. NIPS **4**, 2780–2788 (01 2012)
9. Du, N., Farajtabar, M., Ahmed, A., Smola, A., Song, L.: Dirichlet-hawkes processes with applications to clustering continuous-time document streams. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015). https://doi.org/10.1145/2783258.2783411

10. Du, N., Song, L., Woo, H., Zha, H.: Uncover topic-sensitive information diffusion networks. In: Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS. JMLR Workshop and Conference Proceedings, vol. 31, pp. 229–237. JMLR.org (2013)
11. Erdős, P., Rényi, A.: On the evolution of random graphs. In: PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES. pp. 17–61 (1960)
12. Gomez-Rodriguez, M., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: ICML. p. 561–568 (2011)
13. Gomez-Rodriguez, M., Leskovec, J., Schoelkopf, B.: Structure and dynamics of information pathways in online media. WSDM (2013)
14. Gomez-Rodriguez, M., Leskovec, J., Schölkopf, B.: Modeling information propagation with survival theory. In: ICML. vol. 28, p. III–666–III–674 (2013)
15. He, X., Rekatsinas, T., Foulds, J.R., Getoor, L., Liu, Y.: Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In: ICML (2015)
16. Larremore, D., Carpenter, M., Ott, E., Restrepo, J.: Statistical properties of avalanches in networks. Physical Review E **85** (04 2012). https://doi.org/10.1103/PhysRevE.85.066131
17. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 497–506. KDD '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1557019.1557077
18. Mavroforakis, C., Valera, I., Gomez-Rodriguez, M.: Modeling the dynamics of learning activity on the web. In: Proceedings of the 26th International Conference on World Wide Web. p. 1421–1430. WWW '17 (2017)
19. Mei, Q., Fang, H., Zhai, C.: A study of poisson query generation model for information retrieval p. 319–326 (2007). https://doi.org/10.1145/1277741.1277797, https://doi.org/10.1145/1277741.1277797
20. Myers, S.A., Zhu, C., Leskovec, J.: Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 33–41. KDD '12, Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2339530.2339540
21. Nickel, M., Le, M.: Modeling sparse information diffusion at scale via lazy multivariate hawkes processes. In: Proceedings of the Web Conference 2021. p. 706–717. WWW '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3442381.3450094, https://doi.org/10.1145/3442381.3450094
22. Poux-Médard, G., Pastor-Satorras, R., Castellano, C.: Influential spreaders for recurrent epidemics on networks. Phys. Rev. Research **2**, 023332 (Jun 2020). https://doi.org/10.1103/PhysRevResearch.2.023332
23. Poux-Médard, G., Velcin, J., Loudcher, S.: Powered hawkes-dirichlet process: Challenging textual clustering using a flexible temporal prior. 2021 IEEE International Conference on Data Mining (ICDM) pp. 509–518 (2021)
24. Poux-Médard, G., Velcin, J., Loudcher, S.: Multivariate powered dirichlet-hawkes process. ECIR (2023)
25. Poux-Médard, G., Velcin, J., Loudcher, S.: Powered dirichlet process for controlling the importance of "rich-get-richer" prior assumptions in bayesian clustering. ArXiv (2021)

26. Suny, P., Li, J., Mao, Y., Zhang, R., Wang, L.: Inferring multiplex diffusion network via multivariate marked hawkes process. ArXiv **abs/1809.07688** (2018)
27. Tan, X., Rao, V.A., Neville, J.: The indian buffet hawkes process to model evolving latent influences. In: UAI (2018)
28. Wang, L., Ermon, S., Hopcroft, J.E.: Feature-enhanced probabilistic models for diffusion network inference. In: Machine Learning and Knowledge Discovery in Databases. pp. 499–514. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
29. Yang, S.H., Zha, H.: Mixture of mutually exciting processes for viral diffusion. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 28, pp. 1–9 (2013)