A framework for information dissemination in social networks using Hawkes processes

Julio Cesar Louzada Pinto, Tijani Chahed, Eitan C Altman

A framework for information dissemination in social networks using Hawkes processes

J. C. Louzada Pinto[a,*], T. Chahed[a], E. Altman[b]

[a]*Institut Mines-Telecom, Telecom SudParis, UMR CNRS 5157, France*
[b]*INRIA Sophia Antipolis Mediterranée, 06902 Sophia-Antipolis Cedex, France*

Abstract

We define in this paper a general Hawkes-based framework to model information diffusion in social networks. The proposed framework takes into consideration the hidden interactions between users as well as the interactions between contents and social networks, and can also accommodate dynamic social networks and various temporal effects of the diffusion, which provides a complete analysis of the hidden influences in social networks. This framework can be combined with topic modeling, for which modified collapsed Gibbs sampling and variational Bayes techniques are derived. We provide an estimation algorithm based on nonnegative tensor factorization techniques, which together with a dimensionality reduction argument are able to discover, in addition, the latent community structure of the social network. At last, we provide numerical examples from real-life networks: a Game of Thrones and a MemeTracker datasets.

Key words: Information diffusion, social networks, Hawkes processes, nonnegative tensor factorization, topic models

1. Introduction

Information diffusion/dissemination in social networks refers to users broadcasting (sharing, posting, tweeting, retweeting, liking, etc.) information to

*Corresponding author.

Email addresses: julio.louzada__11_pinto@telecom-sudparis.eu (J. C. Louzada Pinto), tijani.chahed@telecom-sudparis.eu (T. Chahed), eitan.altman@inria.fr (E. Altman)

others in the network. By tweeting, for example, users broadcast information to the network, which is then transmitted to their followers. These sequences of broadcasts by users are called information cascades, and have been studied extensively over the past years; see for example [1, 2, 3, 4]. The large amount of recent work on this subject reflects the strategic real-life implications which may be brought by the knowledge of such cascades: one can discover the hidden impact of users and contents on this diffusion, and highlight various characteristics of not only the social networks in question but also of the influential users and their contents [5, 6, 7, 8].

Information diffusion cascades are complex objects, for which there is no consensus on the standard way to study them; for example: Kempe *et al.* in their seminal paper [2] develop a framework based on submodular functions to detect the optimal seed group in order to diffuse a fixed content in a social network, based on the so-called independent cascade propagation model [9, 10], which is a well known information diffusion model. In [11], Myers and Leskovec study the variation of the probability in retransmiting information due to previous exposure to different types of information; they found that, for Twitter, these retransmission probabilities are indeed very different when compared to results stemming from independent cascade models; however, their approach does not take into consideration the time between broadcasts of information and the topology of the network. And in [4], Gomez-Rodriguez *et al.* study the network inference problem from information cascades using survival theory; however, again, the authors to not take into consideration the underlying network structure.

Among the works dealing with information diffusion, there has been a steady increase of interest in point-processes-based models [12, 13, 14, 15]. Point processes take into consideration the broadcast times of users, whereas a lot of information cascade models consider time to be discrete, i.e., time only evolves when events occur; point processes are counting processes and have thus a discrete state space, which makes them able to fully capture real-life features, such as the number of posts, without increasing the mathematical complexity of the models; and the closed formula for the likelihood of these point processes ([16] p. 232) gives us easy, simple and direct methods for the estimation of important model parameters. For instance, Myers *et al.* study in [12] the influence of externalities from other nodes on information cascades in networks; they use a point process approach, from which the time instances of infection are essential for the estimation of parameters, but the topological properties of the network are of secondary concern in their work.

One point process has been particularly useful in the modeling of these continuous-time models: the Hawkes process [17, 18]. Hawkes processes are self-exciting point processes and are perfect candidates for counting events on information cascades, where users transmit their information to their neighbors in a social network. The use of self-exciting processes here enlightens the necessity of a theory that can model the interaction between people having a conversation or exchanging messages: imagine two people messaging each other through SMS. Normally each one would have its own rhythm of messaging, but due to the self-excitation among these people, they will text and respond faster than they would normally do when generating SMS messages without response. For example, Yang and Zha study in [19] the propagation of memes (see definition in [20] p. 192) in social networks with linear Hawkes processes and couple the point process with a language model in order to estimate the memes. They provide a variational Bayes algorithm for the coupled estimation of the language model, the influence of users and their intrinsic diffusion rates; however, they do not take into consideration the influence that memes may have on one another; moreover, they propose the estimation of the entire social network, not taking into consideration the eventual lack of communication between users.

Hawkes processes have already been successfully employed to study earthquakes [21], neuronal activities [22], high-frequency finance [23], social sciences [19, 24, 25] and many other fields, with a vast and diversified literature.

This paper aims to provide a framework for information diffusion models in social networks using Hawkes processes. The framework encompasses:

- modeling and estimating *user-user* and *topic-topic* interactions,
- modeling and estimating *multiple* social networks and their interactions,
- being combined with topic models [26, 27, 28], for which modified collapsed Gibbs sampling [29, 30, 31] and variational Bayes techniques [32, 33] are derived,
- estimating *different temporal effects* of the users diffusion, such as seasonality and non-homogeneity,
- using and estimating *dynamic/temporal* social networks [34, 35], and

- retrieving the *community structure* of the underlying users influence in social networks, due to a dimensionality reduction during the parameters estimation (see [36] for another example of such methodology).

We also provide numerical examples of our framework for four different datasets: the first two datasets are synthetic datasets generated by Ogata's thinning method [37] for Hawkes processes, the third is a Game of Thrones[1] dataset consisting of the speakers information for the first episode for the first season, from which we retrieve the hidden influence graph of characters, and the last one is a MemeTracker dataset, with different topics and world news for the $5,000$ most active sites from 4 million sites from March 2011 to February 2012[2], where we use the 5 most broadcasted memes to estimate the hidden websites influence graph.

The rest of the paper is organized as follows. Section 2 describes our model for information diffusion using Hawkes processes. Section 3 details the estimation procedure of the hidden influences, with the modified collapsed Gibbs sampling and variational Bayes techniques for the author-topic model [27] as an example. Section 4 discusses some additional topics for our Hawkes diffusion framework. In Section 5 numerical experiments are performed for the beforementioned datsets, and Section 6 concludes the paper.

2. Models

A multivariate linear Hawkes process (see [17, 18] for more details) is a self-exciting orderly point process $X_t,\ t \in [0, \tau]$ with the intensity $\lambda_t = \lim_{\delta \searrow 0} \frac{\mathbb{E}[X_{t+\delta} - X_t | \mathcal{F}_t]}{\delta}$ satisfying

$$\lambda_t = \mu + \int_0^t \phi(t-s) dX_s,$$

where $\mathcal{F}_t = \sigma(X_s,\ s \leq t)$ is the filtration generated by X, μ is an intrinsic Poissonian rate and ϕ is a causal kernel that is responsible for the self-exciting part.

The intensity λ_t can be divided into two distinct parts: the intrinsic Poissonian rate μ, which models the base intensity of the Hawkes process,

[1] `http://en.wikipedia.org/wiki/Game_of_Thrones`.

[2] Data available at `http://snap.stanford.edu/netinf`.

and does not take into account the past of the process, and the self-exciting part $\int_0^t \phi(t-s)dX_s$, which models the interactions of the present with past events. The μ coefficient can, for example, model how some user tweets something, after learning about it in class or at work, after listening to the radio or watching television.

The orderly property of the Hawkes process means that X cannot have two events/jumps at the same time ([16] p. 232), and by the standard theory of point processes ([16] p. 233) we have that an orderly point process is completely characterized by its intensity, which in this case is also a stochastic process (as consequence, the orderly property of Hawkes processes allows the estimation of its parameters by maximum likelihood techniques, as performed in section 3); and the self-excitatory property of the Hawkes process means that future jumps become more probable to occur when X jumps.

We model the social network as a graph $G = (V, E)$, where V is the set of users and E is the set of links between users (Facebook friends, Twitter followers, etc.), such that this information is coded by a directed and unweighted inward adjacency matrix A such that $A_{i,j} = 1$ if user j can influence user i, i.e., if $(j, i) \in E$ or $j \rightsquigarrow i$, and $A_{i,j} = 0$ otherwise.

The broadcasting of messages can be performed in various ways, depending on the application and the social network in question: measuring tweets, retweets or likes, checking the history of a conversation in a chat room, etc. However, they all have one thing in common: messages are broadcasted by the $N = \sharp V$ users in the social network, and users that can receive these broadcasts may be directly influenced by them (this procedure of course allows the indirect influence of users, e.g. two-hops influence, three-hops influence, however these influences are not estimated in our model and are seen as a consequence of the direct influence).

Here is a concrete example of this mechanism: in Facebook, users can post various messages in their wall, where their friends can check on these messages regularly. Assuming that we are modeling the Facebook social network, our graph G is the friendship graph of Facebook, and the Hawkes processes X counts the number of messages that each user posted, regarding each different subject (assuming that we can in some sort categorize the messages' subjects). Thus, when friends check on each other, they increase the likelihood of taking some action, e.g. they may respond to the messages, they may like them, they may send them to other people, or they may even post something else entirely. This mechanism illustrates the *influence* of people and contents in the social network, where a user influences others if

her messages increase the likelihood of other users broadcasting their own messages by, for example, taking some of the beforementioned actions.

Throughout this paper we adopt two kinds of message categories: the first kind assumes that messages are of K predefined topics (economics, religion, culture, politics, sports, music, etc.) and that each message is represented by exactly one of these topics. The second kind assumes a "fuzzy" setup with K topics, where this time the topics are not known beforehand and messages are a mixture of these unknown (latent) topics; Barack Obama may for example tweet something that has 40% of its content about politics, 50% of its content about economics and 10% of its content related to something else.

2.1. User-user and topic-topic interactions with predefined topics

We first focus on the case where messages are about one of K predefined contents (economics, religion, sports, etc.). We assume that we have N users in the social network and that they can influence each other to broadcast, and that these influences are independent of the broadcasted content. On the other hand, the topic to be broadcasted is influenced by the topics already broadcasted beforehand.

This is the case, for example, if one wants to separate the influence effects of users and topics: posts about politics can influence posts about fashion, economics, religion, etc., and people can influence other people simply because they are friends, famous or charismatic. In this model we assume that the influence of a specific user when posting something is given by two different components, the user-user component and the topic-topic component.

These influences are coded by two matrices: the user-user influence matrix J and and the topic-topic influence matrix B, such that $J_{i,j} \geq 0$ is the influence of user i over user j and[3] $B_{c,k} \geq 0$ is the influence of topic c over topic k. In real-life social networks, people who tend to promote conversations and trends in Twitter or Facebook or any other social network would have larger values for the matrix J, whereas topics that are trendy or "important" would have larger values for the matrix B. A maximum likelihhod estimation procedure for these matrices is carried out in subsection 3.1

We model the number of messages broadcasted by users as a linear Hawkes process $X_t \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$, where $X_t^{i,k}$ is the cumulative number of messages of topic k broadcasted by user i until time $t \in [0, \tau]$ in the social network. In

[3]We assume that B is normalized such that $\sum_k B_{c,k} = 1$ for all c, such as in [38].

other words, this X_t is a $\mathbb{R}^{N\times K}$ point process with intensity

$$\lambda_t^{i,k} = \mu^{i,k} + \sum_c \sum_{j \rightsquigarrow i} B_{c,k} J_{i,j} \int_0^{t-} \phi(t-s) dX_s^{j,c} = \mu^{i,k} + \sum_c \sum_{j \rightsquigarrow i} B_{c,k} J_{i,j} (\phi * dX)_t^{j,c},$$

where $\mu^{i,k} \geq 0$ is the intrinsic rate of broadcasting of user i about topic k, $\phi(t) \geq 0$ is the temporal influence kernel that measures the temporal shape of influences coming from past broadcasts - which satisfies $||\phi||_1 = \int_0^\infty \phi(t)dt < \infty$ - and

$$(\phi * dX)_t = \int_0^{t-} \phi(t-s) dX_s \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$$

is the convolution matrix of the temporal kernel ϕ and the jumps dX. This allows one to use $N^2 + K^2$ parameters instead of N^2K^2 for the full fledged model without this influence factorization.

As said before, not all users can communicate with each other. Hence one must take into consideration the inward adjacency matrix A given by the underlying structure on the social network. This is done by the relationship

$$A_{i,j} = 0 \Rightarrow J_{i,j} = 0. \tag{1}$$

Remark: Two standard time-decaying functions are $\phi(t) = \omega e^{-\omega t}.\mathbb{I}_{\{t>0\}}$ a light-tailed exponential kernel [39, 40, 19] and $\phi(t) = (b-1)(a+t)^{-b}.\mathbb{I}_{\{t>0\}}$ a heavy-tailed power-law kernel [25].

2.2. User-topic interactions and global influence in the social network

A different model arises when users do not influence other individually, but they influence the social network as a whole. This means that instead of having an influence matrix $J \in \mathcal{M}_{N\times N}(\mathbb{R}_+)$ that measures the user-user interactions, we have now an influence matrix $\tilde{J} \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$ such that $\tilde{J}_{i,k} \geq 0$ is the influence of user i over the whole social network, when he broadcasts something about topic k.

Hence, the associated Hawkes process $X_t^{i,k}$, which measures the cumulative number of messages broadcasted by user i about topic k until time $t \in [0, \tau]$, has intensity

$$\lambda_t^{i,k} = \mu^{i,k} + \sum_c \sum_{j \rightsquigarrow i} B_{c,k} \tilde{J}_{j,c} \int_0^{t-} \phi(t-s) dX_s^{j,c}.$$

Think about Barack Obama: it is natural that posts or tweets about economics or politics coming from Obama are going to have a much bigger impact than posts about sports or fashion.

2.3. User-user and topic-topic interactions with "fuzzy" topic label

Up until now we have dealt with information dissemination models having K predefined topics and in which each broadcasted message was assumed to belong to one, and only one, of these topics. We consider now a different point of view regarding the broadcasted messages: each message now is a mixture over K undiscovered/latent topics. These topics are distributions over words and each message broadcasted at time $t_s \in [0, \tau]$ generates the *message's empirical distribution of topics* random variable Z^{t_s} such that

$$Z_k^{t_s} = \frac{1}{N_s} \sum_{w=1}^{N_s} z_k^{s,w}, \tag{2}$$

where N_s is the number of words in the message broadcasted at time t_s and $z^{s,w}$ are discrete random variables modeling the topic of word w, i.e., $z_k^{s,w} = 1$ if and only if word w in message t_s is about topic k, and 0 otherwise.

In this model users receive messages that are mixtures of topics and each user reacts to the topics in a different manner. These user-topic interactions are characterized by the matrix $H \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$, such that[4] $H_{i,k}$ measures the influence of topic k over user i.

We define thus the Hawkes processes X_t^i as the cumulative number of messages broadcasted by user i in the social network until time $t \in [0, \tau]$, with intensity

$$\begin{aligned}\lambda_t^i &= \mu^i + \sum_{j \rightsquigarrow i} J_{i,j} \sum_{c,k} B_{c,k} H_{i,k} \int_0^{t-} \phi(t-s) Z_c^s dX_s^j \\ &= \mu^i + \sum_{j \rightsquigarrow i} J_{i,j} \sum_{c,k} B_{c,k} H_{i,k} (\phi *_Z dX)_t^{j,c},\end{aligned}$$

where $\mu^i \geq 0$ represents the intrinsic dissemination rate of user i and

$$(\phi *_Z dX)_t^{j,c} = \int_0^{t-} \phi(t-s) Z_c^s dX_s^j$$

is the (j, c) entry of the weighted convolution of the temporal kernel ϕ and the jumps dX, where the weights are the topic empirical proportions of each message broadcasted by user j.

[4]We also assume that $\sum_k H_{i,k} = 1$, following [41].

Again, not all users can communicate among themselves, hence one must take into consideration Eqn. (1).

In order to fully exploit the random variables Z^{t_s} we use topic models [26, 27, 28], as for example the latent Dirichlet allocation [26] (see[41] for such a methodology) or the author-topic model [27], which will be presented later.

Remark: One can also easily extend the model in subsection 2.2 to the "fuzzy" diffusion framework, following these ideas.

2.4. User-user and topic-topic interactions with predefined topics in multiple social networks

We now turn to the case where we have several "interconnected" social networks. The $m^{th} \in \{1, 2, \cdots, M\}$ social network is defined as a communication graph $G^m = (V^m, E^m)$, where V^m is the set of users and E^m is the edge set, i.e., the set with all the possible communication links between users of the m^{th} social network. Again, we assume these graphs to be directed and unweighted, and coded by inward adjacency matrices A^m such that $A^m_{i,j} = 1$ if user j is able to influence user i in social network m, or $A^m_{i,j} = 0$ otherwise.

One can think about Facebook and Twitter users: there are users in Facebook that do not necessarily follow the same people on Facebook and on Twitter, and vice-versa. Let us say that Facebook is social network 1 and Twitter is social network 2; $A^1_{i,j} = 1$ means that user i follows user j in Facebook and receives the news published by user j in his or her timeline. This does not necessarily imply that $A^2_{i,j} = 1$, i.e., user i also follows user j on Twitter.

Assuming that we have M different social networks, each one with its own adjacency matrix A^m, we model the influence of broadcasts using, similarly to the model in subsection 2.1, three matrices $J \in \mathcal{M}_{N\times N}(\mathbb{R}_+)$, $B \in \mathcal{M}_{K\times K}(\mathbb{R}_+)$ and $S \in \mathcal{M}_{M\times M}(\mathbb{R}_+)$, such that $J_{i,j} \geq 0$ is the influence of user i over user j, $B_{c,k} \geq 0$ is the influence of topic c over topic k and[5] $S_{m,n}$ is the influence that a generic user of social network m has over a generic user of social network n. The network-network influence matrix S measures thus how broadcasts made on one social network influence broadcasts made on the others.

Let $X^{i,k,n}_t$ be the cumulative number of messages broadcasted by user i

[5] As before, we assume that $\sum_n S_{m,n} = 1$.

about content k at social network n until time $t \in [0, \tau]$. The intensity for this process is thus

$$\lambda_t^{i,k,n} = \mu^{i,k,n} + \sum_{m,c,j \rightsquigarrow i} S_{m,n} J_{i,j} B_{c,k} \int_0^t \phi^n(t-s) dX_s^{j,c,m},$$

where J is again the user-user influence matrix, B is the topic-topic influence matrix and μ is the intrinsic rate of dissemination on different social networks.

In view of Eqn. (1), if there exists an edge $j \rightsquigarrow i$ in some social network, then user i can be influenced by user j. Our new constraint becomes

$$\sum_m A_{i,j}^m = 0 \Rightarrow J_{i,j} = 0.$$

One can notice in the definition of the intensity of this model that each social network m has its own[6] temporal kernel function ϕ^m. Each temporal kernel ϕ^m represents how users and contents in each social network are affected by ancient messages, and are considered a *timescale* parameter[7]. Let us take for comparison Twitter and Flickr: in Twitter users chat, discuss, post comments and retweet, while Flickr is a photo-sharing social network that allows users to upload photos and post comments. This means that the conversation and interaction mechanisms in both social networks are different, since they serve different purposes. It is thus natural to assume that users in both social networks react differently to the information received; these different reactions are in part measured by the different temporal kernels $(\phi^m)_{m \in \{1, \cdots, M\}}$.

2.5. Network dependent user-user and topic-topic interactions in multiple social networks

A second (and more complex) extension to the single social network information diffusion model is to assume that the different broadcasting mechanisms in each social network imply different influences on users and topics.

[6]The temporal kernel functions could take more complicated forms, such as $\phi^{k,m}$, where each topic in a social network would have an idiosyncratic temporal kernel function. This enlightens the versatility of this Hawkes framework, allowing one to adapt the system parameters to any desired situation.

[7]Take for example the exponential kernel $\phi(t) = \omega e^{-\omega t}.\mathbb{I}_{\{t>0\}}$: the larger the ω, the larger is the influence of recent broadcasts. This may imply users responding faster to immediate messages.

It means that the user-user and topic-topic influences are now specific to each social network, i.e., user j broadcasting a message about content c on a social network m influences user i *in this same social network* when he broadcasts some message about content k. These network-dependent influences are measured by the user-user influence matrices $(J^m)_{m\in\{1,\cdots,M\}}$ and topic-topic influence matrices $(B^m)_{m\in\{1,\cdots,M\}}$.

Remark: Viewed as high-dimensional objects, J and B are three-dimensional tensors.

We can define, again, $X_t^{i,k,n}$ to be the cumulative number of messages broadcasted by user i about content k in social network n until time $t \in [0, \tau]$. The intensity for this process is then

$$\lambda_t^{i,k,n} = \mu^{i,k,n} + \sum_{m,c,j \overset{m}{\rightsquigarrow} i} S_{m,n} J_{i,j}^m B_{c,k}^m \int_0^t \phi^n(t-s) dX_s^{j,c,m},$$

where $j \overset{m}{\rightsquigarrow} i$ means that user j can influence user i in social network m, i.e., $A_{i,j}^m = 1$.

Since now users only influence themselves in the same social network, the adjacency matrix constraint in Eqn. (1) becomes

$$A_{i,j}^m = 0 \Rightarrow J_{i,j}^m = 0.$$

Remark: One can easily extend the model to social network-social network specific influences of the form $J_{i,j}^{m,n}$ and $B_{c,k}^{m,n}$, for which the above extension is a particular case $J_{i,j}^{m,n} = J_{i,j}^m S_{m,n}$ and $B_{c,k}^{m,n} = B_{c,k}^m S_{m,n}$.

Remark: One can also easily extend the user-topic information dissemination model in a single social network and the "fuzzy" diffusion model in a single social network to take into account multiple social networks, following these ideas.

3. Maximum likelihood estimation and multiplicative updates

One of the strong points about point processes (and Hawkes processes for that matter) is the analytic form of the likelihood of their realization (see [42] or [16] p. 232), where Hawkes-based models for information diffusion used extensively this property in order to derive convex-optimization-based maximum likelihood estimates for the system parameters [15, 19, 43].

Another technique for the maximum likelihood estimation of the Hawkes process X was derived in [39, 40], where the authors slice the information

time period $[0, \tau]$ into T small bins of size $\delta > 0$ in order to create suitable tensors for the intensity and the Hawkes jumps, and show that maximizing an approximation of the log-likelihood is equivalent to solving a nonnegative tensor factorization (NTF) problem [44, 45, 46, 47].

Since we deal with real-life social networks, the number of parameters to be estimated is large and convex optimization techniques that estimate each parameter separately are too demanding in terms of complexity. That is why we adopt the estimation framework of [39, 40], for which multiplicative updates can be derived (see [38, 41] for the same methodology).

Let us take a $\delta > 0$ that is smaller than the minimum elapsed time between broadcasts in $[0, \tau]$ and divide $[0, \tau]$ into $T = \lceil \frac{\tau}{\delta} \rceil$ time bins such that we do not have more[8] than one broadcast in each bin, in order to preserve the orderliness property of X.

Let Y, $\overline{\lambda}$ and $\overline{\phi}$ be tensors such that

$$Y_t = \frac{dX_{(t-1)\delta}}{\delta} = \frac{X_{t\delta} - X_{(t-1)\delta}}{\delta}, \quad \overline{\lambda}_t = \lambda_{(t-1)\delta} \text{ and}$$

$$\overline{\phi}_t^m = \begin{cases} (\phi^m * dX)_{(t-1)\delta} & \text{for predefined topics model} \\ (\phi^m *_Z dX)_{(t-1)\delta} & \text{for "fuzzy" diffusion model,} \end{cases}$$

i.e., Y contains the jumps of X_t at each time bin $((t-1)\delta, t\delta]$.

We begin by showing that maximizing the Riemann-sum approximation of the log-likelihood of X is equivalent to minimizing the Kullback-Leibler (KL) divergence between Y and $\overline{\lambda}$.

Lemma 1. *If $\int_0^\tau \log(\lambda_t^{i,k,m}) dX_t^{i,k,m}$ and $\int_0^\tau \lambda_t^{i,k,m} dt$ are approximated by their respective Riemann sums, then maximizing the approximated log-likelihood of X in $[0, \tau]$ is equivalent to minimizing*

$$D_{KL}(Y|\overline{\lambda}) = \sum_{i,k,m,t} d_{KL}(Y_t^{i,k,m}|\overline{\lambda}_t^{i,k,m}), \tag{3}$$

where $d_{KL}(y|x) = y \log(\frac{y}{x}) - y + x$ is the Kullback-Leibler divergence between x and y.

[8] In practice, this orderliness constraint is not satisfied in order to decrease the complexity of the multiplicative updates.

Proof. Let us place ourselves, without loss of generality, in an information diffusion model with predefined topics[9] and let t_n be the broadcast times in $[0, \tau]$, such that user i_n broadcasted a message about topic k_n in social network m_n at time t_n. We have that the log-likelihood of X is given by (see for example [42] or [16] p. 232)

$$\begin{aligned}\mathcal{L} &= \log\left(\prod_{0\leq t_n\leq\tau} \lambda_{t_n}^{i_n,k_n,m_n}\right) - \sum_{i,k,m}\int_0^\tau \lambda_t^{i,k,m}dt \\ &= \sum_{i,k,m}\left(\int_0^\tau \log\lambda_t^{i,k,m}dX_t^{i,k,m} - \int_0^\tau \lambda_t^{i,k,m}dt\right).\end{aligned}$$

Approximating the integrals in $\mathcal{L}$ by their Riemann sums we get

$$\mathcal{L} \sim \sum_{i,k,m}\sum_t\left(\log\lambda_{(t-1)\delta}^{i,k,m}(X_{t\delta}^{i,k,m} - X_{(t-1)\delta}^{i,k,m}) - \delta\lambda_{(t-1)\delta}^{i,k,m}\right),$$

thus maximizing the approximation of $\mathcal{L}$ is equivalent to minimizing

$$-\mathcal{L}/\delta \sim \sum_{i,k,m}\sum_t\left(\overline{\lambda}_t^{i,k,m} - Y_t^{i,k,m}\log\overline{\lambda}_t^{i,k,m}\right).$$

With Y fixed, this is equivalent to minimizing

$$D_{KL}(Y|\overline{\lambda}) = \sum_{i,k,m,t} d_{KL}(Y_t^{i,k,m}|\overline{\lambda}_t^{i,k,m}).$$

□

Using lemma 1, we have that the maximization of the approximated log-likelihood of X is equivalent to a nonnegative tensor factorization problem with cost function $D_{KL}(Y|\overline{\lambda})$, where Y are the normalized jumps of X and $\overline{\lambda}$ is a tensor representing the intensity of X.

[9]For "fuzzy" diffusion models, we consider the conditional log-likelihood with respect to Z, which is (see for example [16] p. 251)

$$\mathcal{L}(X|Z) = \log\left(\prod_{0\leq t_n\leq\tau}\lambda_{t_n}^{i_n,m_n}\right) - \sum_{i,m}\int_0^\tau\lambda_t^{i,m}dt = \sum_{i,m}\left(\int_0^\tau\log\lambda_t^{i,m}dX_t^{i,m} - \int_0^\tau\lambda_t^{i,m}dt\right).$$

This nonnegative tensor factorization problem stemming from the minimization of the cost function $D_{KL}(Y|\overline{\lambda})$ has already been studied at length in [44, 48, 46, 47], where authors derive convergent multiplicative updates [49, 45].

These multiplicative updates are interesting for several reasons: they are simple to implement (they are basically matrix products and entrywise operations), can be performed in a distributed fashion and have a low complexity on the data, thus being adequate to work on real-life social networks of millions (or even hundreds of millions) of nodes.

These NTF techniques are based on the multiplicative updates given by the following lemma:

Lemma 2. *Let Y be a nonnegative tensor of dimension M, S a nonnegative tensor of dimension $s_S + L$ and H a nonnegative tensor of dimension $h_H + L$ such that $s_S + h_H \geq M$. Define SH, a nonnegative tensor of dimension M, such that*

$$(SH)_{j_1,\cdots,j_M} = \sum_{l_1,\cdots,l_L} S_{i_{s_1},\cdots,i_{s_S},l_1,\cdots,l_L} H_{i_{h_1},\cdots,i_{h_H},l_1,\cdots,l_L},$$

where we have that

- $\{i_{s_1},\cdots,i_{s_S}\} \cup \{i_{h_1},\cdots,i_{h_H}\} = \{j_1,j_2,\cdots,j_M\}$ *(we can still have $\{i_{s_1},\cdots,i_{s_S}\} \cap \{i_{h_1},\cdots,i_{h_H}\} \neq \emptyset$) and*
- $\{j_1,\cdots,j_M\} \cap \{l_1,\cdots,l_L\} = \emptyset$.

Define the cost function

$$D_{KL}(Y|SH) = \sum_{j_1,\cdots,j_M} d_{KL}(Y_{j_1,\cdots,j_M}|(SH)_{j_1,\cdots,j_M}),$$

where $d_{KL}(y|x) = y\log(\frac{y}{x}) - y + x$ is the Kullback-Leibler divergence between x and y.

The multiplicative updates for $D_{KL}(Y|SH)$ of the form

$$Z^{n+1} \leftarrow Z^n \odot \frac{\nabla_Z^- D_{KL}(Y|SH)_{|Z^n}}{\nabla_Z^+ D_{KL}(Y|SH)_{|Z^n}}, \tag{4}$$

with

- *the variables* $Z \in \{S, H\}$, $\nabla_Z^{+/-} D_{KL}(Y|SH)$ *the positive/negative part of* $\nabla_Z D_{KL}(Y|SH)$,
- $A \odot B$ *the entrywise product between two tensors* A *and* B, *and*
- $\frac{A}{B}$ *the entrywise division between two tensors* A *and* B,

satisfy

$$D_{KL}(Y|S^{n+1}H) \leq D_{KL}(Y|S^nH) \quad \textit{and} \quad D_{KL}(Y|SH^{n+1}) \leq D_{KL}(Y|SH^n),$$

i.e., the multiplicative updates produce nonincreasing values for the cost function $D_{KL}(Y|SH)$.

The proof of lemma 2 is based on bounding $D_{KL}(Y|SH)$ by above using an auxiliary function, due to the convexity of d_{KL}. The result when Y, S and H are matrices is well explained in [49, 45] and the general case is demonstrated in A.

Unfortunately, the cost function (3) is not convex on the ensemble of tensors, which means that we cannot expect to retrieve the global minimum of $D_{KL}(Y|\overline{\lambda})$, i.e., the global maximum of the Hawkes likelihood. Nevertheless, it is convex (due to the convexity of the Kullback-Leibler divergence) on each tensor, given that the other is fixed. So, estimating each tensor given the rest fixed in a cyclic way produces nonincreasing values for Eqn. (3), as in [49, 45], thus converging to a local maximum of the approximated log-likelihood.

When $\delta \to 0$, the Riemann sums converge to their respective integrals, and minimizing the cost function in Eqn. (3) becomes equivalent to maximizing the likelihood of X.

As all information diffusion models of our Hawkes-based framework can be estimated using the same techniques based on lemmas 1 and 2, we have thus created a unified information dissemination framework using Hawkes processes.

Similarly to nonnegative matrix factorization (NMF) problems [48, 45], the multiplicative updates in lemma 2 can be sometimes written in a concise matrix form. We give next two examples of such cases: the models of subsections 2.1 and 2.3.

3.1. Estimation of model in subsection 2.1

In order to proceed to the estimation procedure, one needs first to handle the user-user interaction with care: due to the overwhelming number of user-user interaction parameters $J_{i,j}$ in real-life social networks (where we have

millions or even hundreds of millions of users), we factorize J into FG, such that $F \in \mathcal{M}_{N\times d}(\mathbb{R}_+)$ is a $N\times d$ matrix and $G \in \mathcal{M}_{d\times N}(\mathbb{R}_+)$ is a $d\times N$ matrix, with $d \ll N$. This method is similar to clustering the hidden influence graph into different communities (see [36]).

One can also notice that by performing a dimensionality reduction $J = FG$ during the estimation, we not only estimate the influence that users have over one another but we also acquire information on the communities of the underlying social network, since we are able to factorize the hidden influence graph J.

This is a very difficult problem, since the cyclic multiplicative updates destroy this relationship, and the only other way to do so is to estimate each coordinate separately. Since $A_{i,j} \in \{0, 1\}$, we can circumvent this problem using a convex relaxation of this constraint of the form[10] $\eta\langle 1 - A, FG\rangle$ and $\eta \geq 0$ a penalization parameter.

We have the following penalization $\eta\langle 1 - A, FG\rangle$, with derivatives

$$\nabla_F \eta\langle 1 - A, FG\rangle = \eta(1 - A)G^T \quad \text{and} \quad \nabla_G \eta\langle 1 - A, FG\rangle = \eta F^T(1 - A).$$

Unfortunately, since F and G act as a product, there is a potential identifiability issue of the form $FG = F\Gamma\Gamma^{-1}G = \tilde{F}\tilde{G}$ where Γ is any scaled permutation and the pair $\tilde{F} = F\Gamma$, $\tilde{G} = \Gamma^{-1}G$ is also a valid factorization of J (see [44, 50, 39]). We deal with this issue by normalizing the rows of G to sum up to 1 (see [44, 39]). This normalization step involves the resolution of a nonlinear system for each row of G to find the associated Lagrange multipliers.

Our constraint thus becomes $G1 = 1$, for which the Karush-Kuhn-Tucker conditions are written in matrix form as $\overline{\eta}_G = \sum_{i=1}^{d} \eta_{G,i} e_i^T 1$, with $(e_i)_{i\in\{1,\cdots,d\}}$ the standard basis vectors and $\eta_{G,i} \in \mathbb{R}$ the Lagrange multipliers solution of the nonlinear equation $G1 = 1$ after the update.

Let us recall that in this particular model we have the Hawkes parameters $J = FG, B$ and μ. With an abuse of notation, let us define the $N\times T$ matrices Y^k, $\overline{\lambda}^k$ and $\overline{\phi}^k$ such as $Y^k_{i,t} = Y^{i,k}_t$, $\overline{\lambda}^k_{i,t} = \overline{\lambda}^{i,k}_t$ and $\overline{\phi}^k_{i,t} = \overline{\phi}^{i,k}_t$, and the $d \times T$ matrices ρ^k and $\overline{\rho}^k = \sum_{k'=1}^{K} B_{k',k}\rho^{k'}$ such that $\rho^k_{i,t} = \sum_j G_{i,j}\overline{\phi}^k_{j,t}$.

[10] From now on we denote by 1 any vector or matrix with entries equal to 1. The dimension of 1 will be clear in the context.

If we proceed using lemma 2, following [45], we have

$$\partial_{F_{n,m}} D_{KL}(Y|\overline{\lambda}) = \sum_{t,k} \big(1 - \frac{Y_t^{n,k}}{\overline{\lambda}_t^{n,k}}\big)(G\overline{\phi}_t B)_{m,k}$$

$$= \sum_k \Big(\sum_t \big(1 - \frac{Y_t^{n,k}}{\overline{\lambda}_t^{n,k}}\big)(G\overline{\phi}_t B)_{m,k} \Big) = \sum_k \Big(\big(1 - \frac{Y^k}{\overline{\lambda}^k}\big)(\overline{\rho}^k)^T \Big)_{n,m}.$$

Taking the positive and negative parts of $\nabla_F D_{KL}(Y|\overline{\lambda})$ we have that

$$\nabla_F^+ D_{KL} = \sum_k 1(\overline{\rho}^k)^T \quad \text{and} \quad \nabla_F^- D_{KL} = \sum_k [\frac{Y^k}{\overline{\lambda}^k}](\overline{\rho}^k)^T.$$

Since the constraint $\nabla_F \eta \langle 1 - A, FG \rangle = \eta(1 - A)G^T$ is nonnegative, we have that it belongs to the positive part of $\nabla_F D_{KL}(Y|\overline{\lambda})$, which gives the multiplicative updates

$$F \leftarrow F \odot \frac{\sum_{k=1}^K [\frac{Y^k}{\overline{\lambda}^k}](\overline{\rho}^k)^T}{\sum_{k=1}^K 1(\overline{\rho}^k)^T + \eta(1 - A)G^T}.$$

We can proceed accordingly with G, B and μ in a cyclical manner, as already mentioned.

Remark: One can read [38] for the details of estimation in this information diffusion model.

3.2. Estimation of model in subsection 2.3

We focus in this subsection on the "fuzzy" diffusion model with user-user and topic-topic interactions in a single social network. The estimation procedure in "fuzzy" diffusions follows the same ideas as in the preceding subsection, with a minor difference: one also needs to estimate the topic model parameters. The topic model parameters are estimated during the Hawkes parameters estimation, and are influenced by these. At the same time, the topic model parameters also influence the Hawkes parameters estimation through the random variables Z.

The estimation step is thus performed in two steps: first one estimates the topic parameters with the initial values for the Hawkes parameters F, G, B, H and μ, using the conditional log-likelihood of X given Z and lemma 2. Then, one estimates the topic model parameters with the Hawkes parameters fixed. One may just continue the reiteration of the Hawkes parameters estimation, and topic model parameters estimation until convergence.

3.2.1. Hawkes parameters estimation

We perform briefly here an estimation procedure for the Hawkes parameters, given the empirical topic proportions Z fixed.

As before, we have by lemma 1 that maximizing the conditional log-likelihood of X given Z is equivalent to minimizing

$$D_{KL}(Y|\overline{\lambda}) = \sum_{i,t} d_{KL}(Y_t^i|\mu^i + \sum_k H_{i,k} \sum_l F_{i,l}(G\overline{\phi}_t B)_{l,k}).$$

We can, again, use lemma 2 to derive multiplicative updates, as

$$\begin{aligned}\partial_{F_{n,m}} D_{KL}(Y|\overline{\lambda}) &= \sum_t \big(1 - \frac{Y_t^n}{\overline{\lambda}_t^n}\big) \sum_k H_{n,k}(G\overline{\phi}_t B)_{m,k} \\ &= \sum_k \Bigg(\sum_t H_{n,k}\big(1 - \frac{Y_t^n}{\overline{\lambda}_t^n}\big)(G\overline{\phi}_t B)_{m,k} \Bigg).\end{aligned}$$

Define now F^n to be the n^{th} row of F, i.e., the $1 \times d$ vector $F^n = (F_{n1}, \cdots, F_{nd})$, we have thus

$$F^n \leftarrow F^n \odot \frac{\sum_{k=1}^K H_{n,k}\big([\frac{Y^n}{\overline{\lambda}^n}](\overline{\rho}^k)^T\big)}{\sum_{k=1}^K H_{n,k} 1 (\overline{\rho}^k)^T + \eta(1-A)G^T},$$

where $Y^n = (Y_1^n, Y_2^n, \cdots, Y_T^n)$ and $\overline{\lambda}^n = (\overline{\lambda}_1^n, \overline{\lambda}_2^n, \cdots, \overline{\lambda}_T^n)$ are $1 \times T$ row vectors, $\rho = (\mathbb{I} \otimes G)\overline{\phi}$ is an auxiliary $dK \times T$ and $\overline{\rho}^k = \sum_{k'=1}^K B_{k',k}\rho^{k'}$.

We can apply the same principle to G, B, H and μ, yielding:

For G

$$G \leftarrow G \odot \frac{\sum_{k=1}^K (F^k)^T\big([\frac{Y}{\overline{\lambda}}](\overline{\Phi}^k)^T\big)}{\sum_{k=1}^K (F^k)^T 1 (\overline{\Phi}^k)^T + \eta F^T(1-A) + \overline{\eta}_G},$$

where Y and λ are $N \times T$ matrices, F^k is a $N \times d$ matrix such that $F_{i,l}^k = F_{i,l}H_{i,k}$, $\overline{\Phi}^k$ is an auxiliary $N \times T$ matrix such that $\overline{\Phi}_{i,t}^k = \sum_{k'} B_{k'k}\overline{\phi}_t^{i,k'}$, $\eta F^T(1-A)$ is the constraint from Eqn. (1) and $\overline{\eta}_G$ is the constraint $G1 = 1$.

For B

$$B^k \leftarrow B^k \odot \frac{\sum_{i=1}^N H_{i,k}\zeta^i[\frac{Y^i}{\overline{\lambda}^i}]}{\sum_{i=1}^N H_{i,k}\zeta^i 1},$$

where B^k is the k^{th} column of B, $Y^i = (Y_1^i, \cdots, Y_T^i)$ and $\overline{\lambda}^i = (\overline{\lambda}_1^i, \cdots, \overline{\lambda}_T^i)$ are column vectors and ζ^i is a $K \times T$ matrix such that $\zeta_{k,t}^i = \sum_j J_{ij}\overline{\phi}_t^{j,k}$.

For H

$$H^n \leftarrow H^n \odot \frac{[\frac{Y^n}{\overline{\lambda}^n}](\Psi^n)^T}{1(\Psi^n)^T},$$

where H^n is the n^{th} row of H, $Y^n = (Y_1^n, \cdots, Y_T^n)$ and $\overline{\lambda}^n = (\overline{\lambda}_1^n, \cdots, \overline{\lambda}_T^n)$ are row vectors and Ψ^n is a $K \times T$ matrix such that $\Psi_{k,t}^n = (J\overline{\phi}_t B)_{n,k}$.

For μ

$$\mu = \mu \odot \frac{[\frac{Y}{\overline{\lambda}}]1}{1^T 1} = \mu \odot \frac{[\frac{Y}{\overline{\lambda}}]1}{T}.$$

3.2.2. The author-topic model

One of the fundamental pillars of the "fuzzy" diffusion models are the topic models used to estimate the unknown/latent K topics for each broadcasted message, which are statistical models for discovering compound "topics", composed of multiple ideas, that appear in documents [26, 27, 28].

We focus in this subsection on a particular topic model, the author-topic (AT) model [27]; any other topic model whose estimation procedure adapts well in our framework can however be used. One such example is the latent Dirichlet allocation (LDA) topic model of [26], used similarly in [41].

The author-topic model is a generative model that works as follows: a group of authors $a_t \in V$ broadcasts a message at time t_t. For each word w in the message, an author is chosen uniformly at random in a_t, which is coded by the random variable $x^{t,w}$. Then, a topic $z^{t,w}$ is chosen from a discrete random variable[11] $\theta^{x^{t,w}}$, and finally the word w is generated from the chosen topic, following the distribution $\mathbb{P}(w = W^v | z^{t,w} = k) = \beta_{k,v}$, where β is the topic-word distribution[12] and W^v is the v^{th} entry in the dictionary $\{1, 2, \cdots, W\}$ (see figure 1).

The random variables $x^{t,w}$ represent the author associated with the word w in the message broadcasted at time t_t, sampled uniformly from $a_t \subset \{1, 2, \cdots, N\}$. In our case, $\sharp a_t = 1$ for all messages broadcasted at times $t_t \leq \tau$, i.e., each message has only one author and this author is the user $i_t \in \{1, 2, \cdots, N\}$ that posted the message at time t_t.

The LDA model, on the other hand, models each message broadcasted at time t_t with an independent topic random variable θ^t, which is itself sampled

[11] The discrete random variables θ^a, $a \in \{1, 2, \cdots, N\}$ are sampled from a Dirichlet random variable with parameter α.

[12] The rows β_k, $k \in \{1, \cdots, K\}$ are generated by a Dirichlet random variables with parameter η.

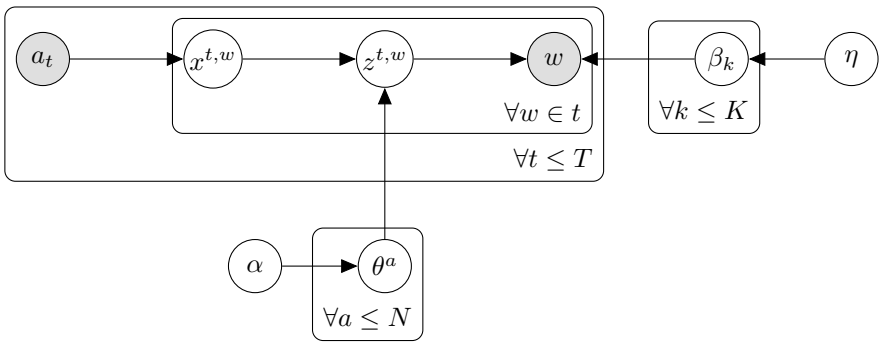

Figure 1: Graphical model for author-topic language model [27].

from a Dirichlet distribution with parameter α. Thus, for each word w in the message broadcasted at time t_t, we sample independent word-topic discrete random variables $z^{t,w}$ using the message-topic random variable θ^t, and select the word w following the distribution $\mathbb{P}(w = W^v | z^{t,w} = k) = \beta_{k,v}$, where β is the topic-word distribution and W^v is the v^{th} entry in the dictionary W (see figure 2).

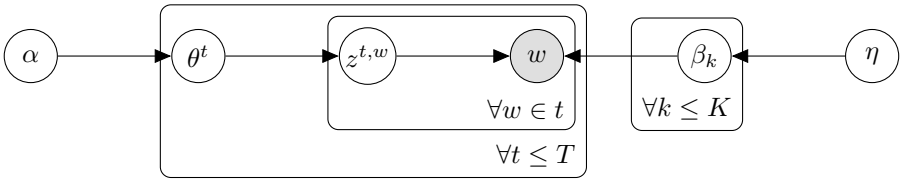

Figure 2: Graphical model for latent Dirichlet allocation [26].

The motivation for using the AT language model instead of the LDA language model is the fact that in the LDA model, messages are a mixture of topics, independent of other messages. This means that we do not take into consideration the authors inclination to post messages on their topics of expertise or interest. Think for example of Barack Obama: he is more likely to tweet about topics related to economics or politics than topics related to fashion or sports. The AT model takes that individuality into consideration when discovering the latent topics.

A potential weakness of the AT model is that it does not consider messages individually. Messages are thus generated only by a mixture of the authors topic distributions. The LDA model on the other hand is in a sense its complete opposite - it allows each message to have its own message-specific

topic mixture. One could also provide less "extreme" topic models that lie between these two.

There are two very common ways of estimating the author-topic parameters β, θ and z used in this generative model: variational methods [26, 32, 33] and Gibbs sampling methods [29, 30, 31].

3.2.3. Modified collapsed Gibbs sampling

We take advantage of the relationship between the topic model and the information diffusion, given by the random variables Z^{t_s}, and derive a more data-driven methodology for the author-topic model and the content dissemination itself. Another example of such methodology is [19].

We derive hence a modified collapsed Gibbs sampling, using the sampling equation of [27]. Gibbs sampling is a member of the Markov-chain Monte Carlo (MCMC) framework (see [51, 29]), and it is a particular instance of the Metropolis-Hastings algorithm. In Bayesian estimation, MCMC algorithms aim to construct a Markov chain that has the posterior distribution as its unique stationary distribution.

In the standard author-topic generative model, one wants to sample from the posterior[13] of (z, x), i.e., $\mathbb{P}(z, x, \theta, \beta | w, \alpha, \eta)$, however this expression is unknown. On the other hand, since the discrete distribution and the Dirichlet distribution are conjugate, one can analytically integrate θ and β out of the posterior.

Let $z^{s,i}$ be the topic and $x^{s,i}$ be the author of word w^i, belonging to message[14] s, and $z^{-(s,i)}$ be the topics and $x^{-(s,i)}$ be the authors of all other words except the word w^i; we have (following [52, 30, 31]) that

$$\begin{aligned} &\mathbb{P}(z^{s,i} = k, x^{(s,i)} = a | z^{-(s,i)}, x^{-(s,i)}, w, \alpha, \eta) \\ &\propto \frac{n_{w^i,k}^{-w^i,WT} + \eta_{w^i}}{n_{\cdot,k}^{-w^i,WT} + \sum_j \eta_j} \cdot \frac{n_{a,k}^{-w^i,TA} + \alpha_k}{n_{a,\cdot}^{-w^i,TA} + \sum_{k'} \alpha_{k'}}, \end{aligned} \tag{5}$$

where

- $n_{w^i,k}^{-w^i,WT}$ is the number of instances of word w^i assigned to topic k, in exception of word w^i in message s,

[13] Since $\sharp a_s = 1$ for every message s, from now on we omit every dependence of the probabilities on the authors. For example, $\mathbb{P}(z, x, \theta, \beta | w, a, \alpha, \eta) = \mathbb{P}(z, x, \theta, \beta | w, \alpha, \eta)$.

[14] From now on, we denote message s the message broadcasted at time t_s.

- $n_{\cdot,k}^{-w^i,WT}$ is the total number of words, in exception of word w^i in message s, that are assigned to topic k,
- $n_{a,k}^{-w^i,TA}$ is the number of words of author a assigned to topic k, in exception of word w^i in message s,
- $n_{a,\cdot}^{-w^i,TA}$ is the total number of words of author a, in exception of word w^i in message s.

Since $\sharp a_s = 1$ for all message s, we have that

$$\mathbb{P}(z^{s,i}, x^{s,i} | z^{-(s,i)}, x^{-(s,i)}, X, w, \alpha, \eta) = \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, X, w, \alpha, \eta)$$

and by Bayes rule (since the intensity λ_t depends only on z through Z)

$$\begin{aligned}
&\mathbb{P}(z^{s,i}, x^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, X, \alpha, \eta) = \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, X, \alpha, \eta) \\
&\propto \mathbb{P}(X | z^{s,i}, z^{-(s,i)}, x^{-(s,i)}, w, \alpha, \eta) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, \alpha, \eta) \\
&= \mathbb{P}(X | z^{s,i}, z^{-(s,i)}) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, \alpha, \eta) \\
&= \mathbb{P}(X | Z) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, \alpha, \eta) \\
&= L(X|Z) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, \alpha, \eta),
\end{aligned}$$

where $L(X|Z)$ is the conditional likelihood of X given Z (see [16] p. 251), as

$$L(X|Z) = \big[\prod_{n=1}^{X(\tau)} \lambda_{t_n}^{i_n} \big] \exp(-\sum_i \int_0^\tau \lambda_u^i du), \tag{6}$$

where i_n is the user that broadcasted the message at time t_n and $X(\tau)$ is the total number of jumps of X in $[0, \tau]$, i.e., the total number of messages broadcasted in $[0, \tau]$.

Remark: Let t_s be the time of broadcast of the message s containing word w^i, i_s be the user that broadcasted the message at time t_s. Looking at the likelihood $L(X|Z)$ more closely, one can see that some terms do not depend on $z^{s,i}$, and are casted out during the normalization process; these are the terms not containing Z^{t_s}.

Remark: After achieving the stationary regime for z one can compute the estimators for θ and β as

$$\hat{\theta}_{a,k} = \frac{n^{TA}(a,k) + \alpha_k}{\sum_{k'} n^{TA}(a,k') + \alpha_{k'}} \quad \text{and} \quad \hat{\beta}_{k,j} = \frac{n^{WT}(k,j) + \eta_j}{\sum_{j'} n^{WT}(k,j') + \eta_{j'}},$$

where $n^{TA}(a,k)$ is the number of times a word of author a is of topic k and $n^{WT}(k,j)$ is the number of times the j-word of the vocabulary was associated with topic k (see [27]).

3.2.4. Modified variational Bayes

One alternative to Gibbs sampling is a variational method, where one replaces the sampling part by an optimization procedure. We derive here a modified variational Bayes estimation, using B.

Following B we introduce the free Dirichlet variables $\gamma^a = (\gamma_1^a, \cdots, \gamma_K^a)$, $\gamma_k^a \geq 0$, for the author topic distributions θ^a and the free discrete variables $\psi^{a,i}$ for $z^{l,i}$, where $\sum_k \psi_k^{l,i} = 1$ and $\psi_k^{l,i} \geq 0$, i.e., $\psi^{l,i}$ evolves in the probability simplex. We can thus retrieve the random variables θ^a, $z^{l,i}$ as $\theta^a \sim Dirichlet(\gamma^a)$ and $z^{l,i} \sim Discrete(\psi^{l,i})$

Our approach, again, makes use of the Hawkes process X to modify the true posterior and introduces a dependence between the dynamics of X and the author-topic model. To include the Hawkes process X into our posterior, we use Bayes rule as (again, since $\sharp a_s = 1$)

$$\begin{aligned}\mathbb{P}(\theta, z, w, x, X|\alpha, \beta) &= \mathbb{P}(X|\theta, z, w, x, \alpha, \beta).\mathbb{P}(\theta, z, w, x|\alpha, \beta) \\ &= \mathbb{P}(X|Z).\mathbb{P}(\theta, z, w|\alpha, \beta) = L(X|Z).\mathbb{P}(\theta, z, w|\alpha, \beta),\end{aligned} \tag{7}$$

where $L(X|Z)$ is the conditional likelihood of X given Z, as in [41].

Applying the same methods as in appendix $A.3$ in [26] andB, we have

$$\log \mathbb{P}(X, w|\alpha, \beta) = L(\gamma, \psi; \alpha, \beta) + d_{KL}(q(\theta, z|\gamma, \psi)|\mathbb{P}(\theta, z|X, w, \alpha, \beta)),$$

where

$$\begin{aligned}L(\gamma, \psi; \alpha, \beta) &= \mathbb{E}_q[\log \mathbb{P}(\theta, z, w, X|\alpha, \beta)] - \mathbb{E}_q[q(\theta, z)] \\ &= \mathbb{E}_q[\log L(X|Z)] + \mathbb{E}_q[\log \mathbb{P}(\theta, z, w|\alpha, \beta)] - \mathbb{E}_q[q(\theta, z)]\end{aligned}$$

by Eqn. (7) and, by Eqn. (6),

$$\mathbb{E}_q[\log L(X|Z)] = \mathbb{E}_q[\sum_{n=1}^{X(\tau)} \log(\lambda_{t_n}^{i_n})] - \mathbb{E}_q[\sum_i \int_0^\tau \lambda_t^i dt]. \tag{8}$$

However, one cannot compute analytically Eqn. (8), but one can derive a lower bound: let i_l be the user that broadcasted message l. Due to the concavity of the logarithm and the fact that $\mathbb{E}_q[Z^{t_l}] = \frac{1}{N_l} \sum_i \mathbb{E}_q[z^{l,i}] =$

$\frac{1}{N_l}\sum_i \psi^{l,i} = \tilde{\psi}^l$, one can introduce nonnegative branching variables u such that $u^t_{i,0} + \sum_{t_l<t,c} u^t_{i,c,l}\tilde{\psi}^l_c = 1$ and bound $\mathbb{E}_q[\log(\lambda^i_t)]$ in the following way:

$$\begin{aligned}
&\mathbb{E}_q[\log(\lambda^i_t)] = \mathbb{E}_q[\log(\mu^i + \sum_{t_l<t}\sum_{c,k} J_{i,i_l}B_{c,k}H_{i,k}\phi(t-t_l)Z^{t_l}_c)] \\
&\geq \sum_{t_l<t,c} u^t_{i,c,l}\mathbb{E}_q[Z^{t_l}_c]\log(J_{i,i_l}\sum_k B_{c,k}H_{i,k}\phi(t-t_l)) \\
&+ u^t_{i,0}\log(\mu^i) - u^t_{i,0}\log(u^t_{i,0}) - \sum_{t_l<t,c} u^t_{i,c,l}\mathbb{E}_q[Z^{t_l}_c]\log(u^t_{i,c,l}) \\
&= u^t_{i,0}\log(\mu^i) + \sum_{t_l<t,c} u^t_{i,c,l}\tilde{\psi}^l_c\log(J_{i,i_l}\sum_k B_{c,k}H_{i,k}\phi(t-t_l)) \\
&- u^t_{i,0}\log(u^t_{i,0}) - \sum_{t_l<t,c} u^t_{i,c,l}\tilde{\psi}^l_c\log(u^t_{i,c,l}). \qquad (9)
\end{aligned}$$

We can find the u that makes this bound the tightest possible by maximizing it under the constraint $u^t_{i,0} + \sum_{t_l<t,c} u^t_{i,c,l}\tilde{\psi}^l_c = 1$, which gives us

$$u^t_{i,0} = \frac{\mu^i}{\mu^i + \sum_{t_l<t,c}\tilde{\psi}^l_c J_{i,i_l}\sum_k B_{c,k}H_{i,k}\phi(t-t_l)}$$

and

$$u^t_{i,c,l} = \frac{J_{i,i_l}\sum_k B_{c,k}H_{i,k}\phi(t-t_l)}{\mu^i + \sum_{t_l<t,c}\tilde{\psi}^l_c J_{i,i_l}\sum_k B_{c,k}H_{i,k}\phi(t-t_l)}.$$

We also trivially have

$$\mathbb{E}_q[\sum_i \int_0^\tau \lambda^i_t dt] = \tau\sum_i \mu^i + \sum_{i,c,k,t_l<\tau}\Phi(\tau-t_l)J_{i,i_l}B_{c,k}H_{i,k}\tilde{\psi}^l_c,$$

where $\Phi(t) = \int_0^t \phi(s)ds$ is the primitive of $\phi(t)$.

Plugging Eqns. (9) and (10) into $L(\gamma,\psi;\alpha,\beta)$ and deriving with respect to $\psi^{l,w}_c$ we find

$$\partial_{\psi^{l,w}_c} L(\gamma,\psi;\alpha,\beta) = \varphi^{l,w}_c + AT^{l,w}_c + LM^{s,w} = 0,$$

where

$$\begin{aligned}
\varphi^{l,w}_c = \frac{1}{N_l}\Bigg(&- \sum_i \Phi(\tau-t_l)J_{i,i_l}\sum_k B_{c,k}H_{i,k} \\
&+ \sum_{t_s>t_l} u^s_{i_s,c,l}\log(\frac{J_{i_s,i_l}\sum_k B_{c,k}H_{i_s,k}\phi(t_s-t_l)}{u^s_{i_s,c,l}})\Bigg),
\end{aligned}$$

with i_n the user that broadcasted message n, $AT_c^{l,w}$ a term stemming from the standard variational approach for the author-topic model responsible for $\psi_c^{l,w}$ in Eqn. (12), $v_w \in \{1, 2, \cdots, W\}$ the unique index such that $w_v = 1$ and $LM^{s,w}$ is a Lagrange multiplier for the constraint $\sum_c \psi_c^{l,w} = 1$.

Following appendix $A.3$ in [26] it is then straightforward to get

$$\psi_k^{l,w} \propto \beta_{k,v_w} \exp(\varphi_k^{l,w} + \Psi(\gamma_k^{i_l}) - \Psi(\sum_{k'} \gamma_{k'}^{i_l})).$$

Since $L(X|Z)$ does not depend on θ (and by consequence on γ) nor β, we have that the updates for γ and β are the same as fromB.

Remark: A great deal of importance is given to the hyperparameters η and α. They are responsible for "smoothing" the Dirichlet random variables β and θ, and giving them a predetermined shape (see [53, 54, 55] for a more throughout discussion).

4. Additional remarks

- *Introduction of seasonality in the intrinsic intensity μ:* it may be desirable to introduce periods in which people behave differently and thus broadcast messages differently; take Twitter, for example, where users probably have a higher intrinsic rate during the evening compared to the late night or early morning.

  To do so, let us define periods $\tau_n \in [0, \tau]$ such that $\tau_i \cap \tau_j = \emptyset$ and $\bigcup_n \tau_n = [0, \tau]$. An example: let τ_1 be all the periods $[0, 6h]$ for every day in $[0, \tau]$, $\tau_2 = (6h, 12h]$, $\tau_2 = (12h, 18h]$ and $\tau_2 = (18h, 24h]$..

  Let 1^{τ_n} be the $T \times 1$ vector such that $1_t^{\tau_n} = \mathbb{I}_{\{\delta(t-1) \in \tau_n\}}$ and μ_{τ_n} be the intrinsic rate associated with the period τ_n. Thus, $1 = \sum_n 1^{\tau_n}$ and we can apply our NTF procedure for each μ_{τ_n} separately. For example, the model in subsection 2.3 has the following updates for the time periods τ_n

  $$\mu_{\tau_n} \leftarrow \mu_{\tau_n} \odot \frac{[\frac{Y}{\overline{\lambda}}]1^{\tau_n}}{\langle 1, 1^{\tau_n} \rangle}.$$

  One could also incorporate a nonparametric estimation of the intrinsic rate μ as in Lewis and Mohler [56].

Algorithm 1 Estimation procedure

Input: jumps dX, step size δ, temporal kernels $(\phi^m)_{m\in\{1,\cdots,M\}}$
1**:** Discretize $[0,\tau]$ into T bins of size δ
2**:** Calculate normalized jumps $Y = \frac{dX}{\delta}$, convolution tensors $\overline{\phi}$ and discretized intensities $\overline{\lambda}$
3**:** Initialize Hawkes matrices set $\mathcal{X}$ *(for example, in a user-user topic-topic model with predefined topics* $\mathcal{X} = \{F, G, B, S, \mu\}$*)*
while Matrices in $\mathcal{X}$ have not converged **do**
  if In a "fuzzy" diffusion model **then**
    4**:** With all Hawkes matrices $\mathcal{X}$ fixed, run round of topic model estimation
  end if
  for matrix x **in** $\mathcal{X}$ **do**
    5**:** With all other matrices fixed *(and topic model parameters as well, if in a "fuzzy" diffusion model)*, update x as

$$x^{n+1} \leftarrow x^n \odot \frac{\nabla_x^- D_{KL}(Y|\overline{\lambda})_{|x^n}}{\nabla_x^+ D_{KL}(Y|\overline{\lambda})_{|x^n}},$$

  end for
end while
Output: Hawkes matrices $\mathcal{X}$ and topic model parameters

- *Estimation of the temporal kernel:* As already mentioned before, the temporal kernel represents the temporal interactions between the ancient and future broadcasted messages on the social network in question. It is a timescale parameter and it may be advantageous to retrieve it from data, instead of being an input of the model.

  A temporal kernel estimation step can be introduced in algorithm 1 without difficulty, where expectation-maximization algorithms [57, 56] for parametric kernels, e.g., the exponential and power-law kernels, or even nonparametric algorithms [58] can be implemented.

- *Introduction of dynamic/temporal networks [34, 35]:* In many cases, links in social networks are severed or acquired, which means that the social network in question can be a dynamic object, instead of a static one.

Let us consider the model in subsection 2.1, for simplicity. In this case, one can define P increasing periods of time $(\tau_p)_{p\in\{1,\cdots,P\}}$ such that $\tau_p \in [0,\tau]$, $\tau_{p'} \cap \tau_p = \emptyset$, $\bigcup_p \tau_p = [0,\tau]$ and $\sup \tau_p = \inf \tau_{p+1}$. We may also have that the adjacency matrices A^p satisfy $A^p \neq A^{p'}$ if $p \neq p'$, i.e., each period of time represents a change in the underlying network structure.

Let 1^{τ_p} be the $N \times T$ matrix such that $1^{\tau_p}_{i,t} = \mathbb{I}_{\{\delta(t-1)\geq \inf \tau_p\}}$, $p_t = \{p \in \{1,\cdots,P\} \mid t \in \tau_p\}$ be the unique index p such that $t \in \tau_p$, $\mathcal{P}_t = \bigcup_{s\leq t} p_s$ are all time period indices until time t and $j \overset{p}{\rightsquigarrow} i$ means that user j can influence user i on the time period τ_p. Thus the intensity is given by

$$\begin{aligned}\lambda^{i,k}_t &= \mu^{i,k} + \sum_c \int_0^{-t} \sum_{j \overset{p_s}{\rightsquigarrow} i} J^{p_s}_{i,j} \phi(t-s) dX^{j,c}_s \\ &= \mu^{i,k} + \sum_c \sum_{p\in\mathcal{P}_t} \int_{\inf \tau_p}^{t\wedge \sup \tau_p} \sum_{j \overset{p}{\rightsquigarrow} i} J^p_{i,j} \phi(t-s) dX^{j,c}_s,\end{aligned}$$

which in tensor form is given by (as a product of matrices)

$$\overline{\lambda}_t = \mu + \sum_{p\in\mathcal{P}_t} J^p \overline{\phi}_{p,t} B, \tag{10}$$

where $\overline{\phi}^{j,c}_{p,t} = \int_{\inf \tau_p}^{(t-1)\delta\wedge \sup \tau_p} \phi(t-s) dX^{j,c}_s$ if $(t-1)\delta \in \tau_p$ and 0 otherwise.

Define the $N\times T$ matrices $Y^{p,k}$ and $\overline{\lambda}^{p,k}$ such that $Y^{p,k}_{i,t} = Y^{i,k}_t . \mathbb{I}_{\{(t-1)\delta\geq \inf \tau_p\}}$ and $\overline{\lambda}^{p,k}_{i,t} = \overline{\lambda}^{i,k}_t . \mathbb{I}_{\{(t-1)\delta\geq \inf \tau_p\}}$, and the $d \times T$ matrices $\rho^{p,k}$ and $\overline{\rho}^{p,k} = \sum_{k'} B_{k',k} \rho^{p,k'}$ such that $\rho^{p,k}_{i,t} = \sum_j G^p_{i,j} \overline{\phi}^{p,k}_{j,t}$. Using lemma 2 we have that the multiplicative updates for F take the form

$$F^p \leftarrow F^p \odot \frac{\sum_{k=1}^K [\frac{Y^{p,k}}{\overline{\lambda}^{p,k}}] (\overline{\rho}^{p,k})^T}{\sum_{k=1}^K 1^{\tau_p} (\overline{\rho}^{p,k})^T + \eta^p (1 - A^p)(G^p)^T},$$

with similar multiplicative updates for G, B and μ.

One may also be interested in finding the ensemble $(J^p)_{p\in\{1,\cdots,P\}}$ the smoothest as possible (if one sees the temporal function $p \mapsto J^p$ as the way the adjacency matrix J evolves through time), one may also apply L^1 or L^2 regularization techniques during the estimation of F

and G. If, for example, we use a cost function of the form $g(F,G) = \eta \sum_{0\leq p<P} ||J^{p+1} - J^p||^2 = \eta \sum_{0\leq p<P} ||F^{p+1}G^{p+1} - F^pG^p||^2$ as regularizing function, we have that the derivatives $\partial_{F^p} g$ and $\partial_{G^p} g$ behave as in lemma 2 (being divided into positive and negative parts) and it is thus easy to incorporate this step into the cyclical estimation algorithm 1.

- *Alternative estimation methods:* The problem of nonnegative tensor factorization (or more specifically nonnegative matrix factorization) has been studied for a long time now, with a vast and varied research literature. The NTD multiplicative updates used in this chapter are simply *one* of the existing methods for NTD estimation. The reasons for the use of multiplicative NTD updates are: they are easy to implement, can be implemented in a distributed fashion, have a low (even linear) complexity on the data, provide an easy way to introduce penalizations and constraints, and they provide a mathematically solid and unified estimation framework for the Hawkes-based information diffusion models.

  Other methods are: projected gradient and alternate least-square algorithms [59, 60, 61, 62], fixed-point alternating least-squares algorithms [63, 64], quasi-Newton algorithms [61, 65], multilayer techniques and hierarchical methods [61, 66, 67], etc. The reader has an excellent review of these methods in [68].

5. Numerical examples

In this section we describe some numerical examples of this information diffusion framework, or more specifically, examples of the model in subsection 2.1.

We have four different datasets, two simulated with the thinning algorithm[15] developed by Ogata in [37] and two real-life datasets:

- The first example is a synthetic dataset of a 2-clique uniformly random network with $N = 100$ (each complete clique having 50 nodes), $K = 10$

[15]The thinning algorithm simulates a standard Poisson process P_t with intensity $M > \sum_{i,k} \lambda_t^{i,k}$ for all $t \in [0,\tau]$ and selects from each jump of P_t the Hawkes jumps of $X_t^{i,k}$ with probability $\frac{\lambda_t^{i,k}}{M}$, or no jump at all with probability $\frac{M - \sum_{i,k} \lambda_t^{i,k}}{M}$.

and an exponential temporal kernel; corresponding figures are 3, 4, 5 and 6.

We used $d = 51$ for our factorization $J = FG$, with a linear penalization (as in subsection 3.1) with constants $\eta_F = \eta_G = 10^3$. We did not use cross-validation techniques to find optimal penalization parameters η_F and η_G, since the algorithm is robust enough with respect to them.

Figure 3 is the heatmap of $J = FG$, where the left heatmap is the estimated $J = FG$ and the right heatmap is the true value for J. One can clearly see that our algorithm retrieves quite well the structure behind the true J, i.e., two distinct cliques.

Figure 4 is the heatmap of the squared difference of the true J and its estimation $\tilde{J}$, i.e., for each true entry $J_{i,j}$ and estimated entry $\tilde{J}_{i,j}$ we have plotted the differences $(J_{i,j} - \tilde{J}_{i,j})^2$ and $(J_{i,j} - \tilde{J}_{i,j})^2 / J_{i,j}^2$ (when $J_{i,j}$ is nonzero).

Figure 5 refers to the squared difference of B and its estimation and figure 6 refers to the squared difference of the true μ and its estimation, as in figure 4.

- The second example is again a synthetic dataset, simulated for a 2-clique uniformly random network with $N = 20$ and $K = 1$ and exponential temporal kernel; corresponding figures are7 and 8.

  We compare our estimation choosing $d = 10$ with the estimation algorithm in [19] (with the simplification of $K = 1$ and no language model), which models memes propagation in a social network using a Hawkes model similar to ours (identical to ours when $K = 1$), but making use of an auxiliary language model for the memes labeling and not using the factorization $J = FG$ as in subsection 3.1; one can see that our algorithm (on the left of figure 7) outperforms the algorithm of [19] not only in the estimation[16] of μ, but also in the estimation of J, retrieving

[16] One can clearly see that our algorithm is able to detect the different sets of values for μ, although with a high variance. This is understandable, because a linear Hawkes process is equivalent to a Poisson cluster process (see [69]), where immigrants arrive following a Poisson process with rate μ. This means that the algorithm estimates a rate μ of a Poisson process, which is known to have (optimal) variance μ itself (see [70]), hence a larger rate implies a larger variance. Of course the estimation improves when $\tau \to \infty$, since for J fixed this is equivalent to a maximum likelihood estimator (MLE), which is consistent and

the community structure when the algorithm in [19] did not. Moreover, the algorithm of [19] needs an ad-hoc parameter ρ to control the sparsity of the network, which is not needed in our case.

- The third example is a Game of Thrones[17] (GOT) dataset with the dialogues of the pilot episode, with their respective timestamps and characters. We assumed that every GOT character could influence all the other characters to speak during the episode (by measuring their temporal influence with a temporal kernel, as before), and we used the tools of subsection 3.1 to estimate the characters hidden influence matrix J using $K = 1$, i.e., we are only concerned with the characters' influence on each other without any topic distinction. The heatmap of J is plotted in figure 9, and shows that our estimation algorithm indeed performs a community detection procedure, by dividing the influence graph into the two most famous families *Stark* and *Lannister*.

  Although this example is not a social network in the sense of Twitter or Facebook, it has nevertheless all of its characteristics: users (the GOT characters) that interact with each other following broadcast of messages (their lines during the episode). Moreover, since we do know the intrinsic hidden structure of allegiances in the series, i.e. Starks vs. Lannisters, we can measure if our model recovers this structure through matrix J, which is indeed the case.

- The last example is a MemeTracker dataset, with different topics and world news for the $5,000$ most active websites from 4 million sites from March 2011 to February 2012[18]. We used the 5 most broadcasted memes, i.e., $K = 5$, leading to the websites influence graph in figure 10. This graph was plotted with the websites having the 10% largest outdegrees[19] and shows the influence of websites on one another. The thicker the edge lines, the larger the influence, and the larger the website's name, the larger the overall influence of the website (the sum of its influences).

asymptotically normal (see [42]); we used in this example a rather small τ for performance reasons.

[17] `http://en.wikipedia.org/wiki/Game_of_Thrones`.

[18] Data available at `http://snap.stanford.edu/netinf`.

[19] We have chosen the 9^{th} decile of nodes regarding the distribution $(\sum_i J_{i,j})_{j \in V}$.

This example is in fact a simple illustration of our Hawkes framework since we cannot validate our results: the parameters J and B do not actually exist in real-life social networks because they are the *hidden influences* of users (in this case websites) and topics (in this case the 5 most broadcasted memes). *As a consequence, our framework helps gain qualitative knowledge on real-life social networks, which seems to be a hard enough task in social sciences.*

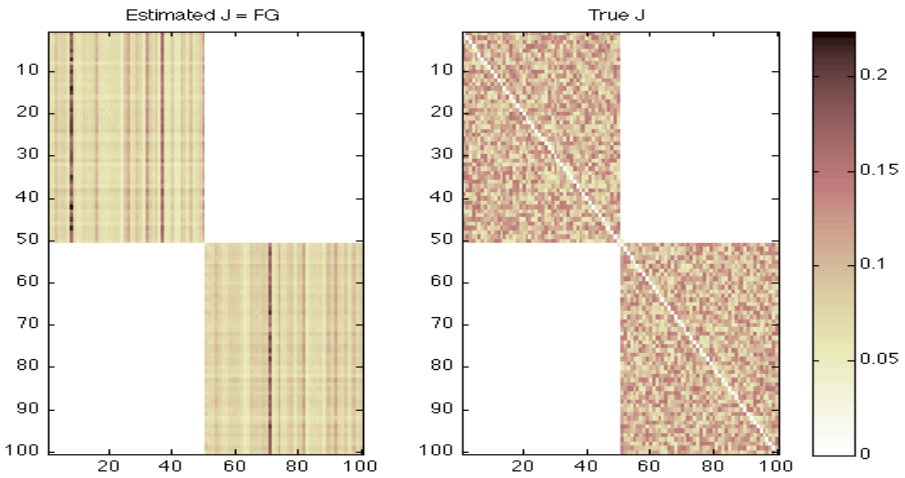

Figure 3: Heatmap of $J = FG$ for 2-clique network of 100 nodes.

6. Conclusion

We presented in this paper a general framework to model information diffusion in social networks based on the theory of self-exciting point processes - linear multivariate Hawkes processes. Hawkes processes were already successfully introduced in a multitude of domains, such as neuroscience, finance, seismology, and even social sciences, and present themselves as a natural way to model information cascades in social networks.

The framework developed here exploits the real broadcasting times of users - a feature that comes with no mathematical overhead since we do so in the theory of point processes - which guarantees a more realistic view of the information diffusion cascades.

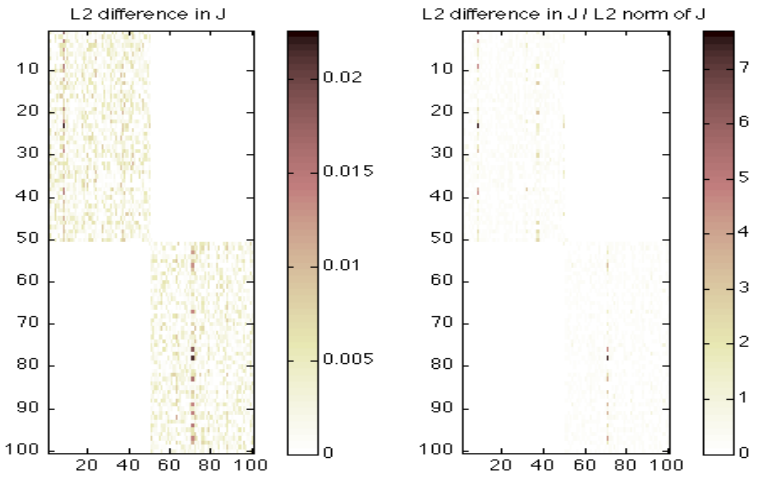

Figure 4: Heatmap of L^2 differences (absolute and relative) between entries of true J and estimated J.

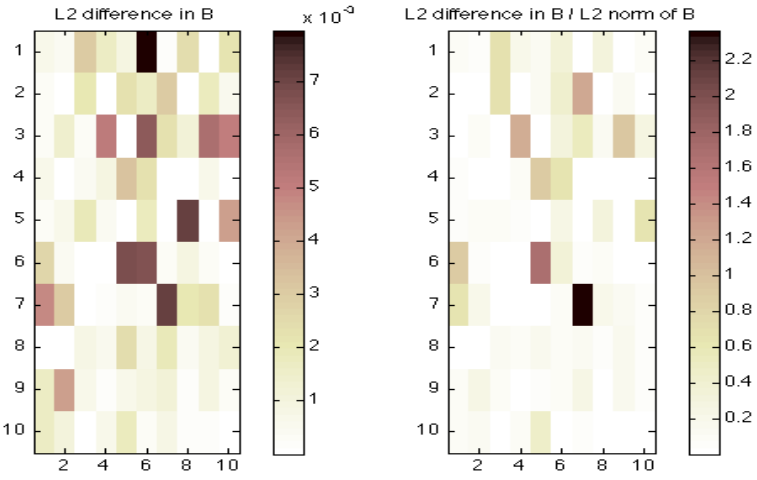

Figure 5: Heatmap of L^2 differences (absolute and relative) between entries of true B and estimated B.

Our framework can take into consideration every influence between users and contents in social networks, under a variety of assumptions, which provides a clear view of hidden influences in social networks.

This framework also allows one to use predefined topics (labeled data) and

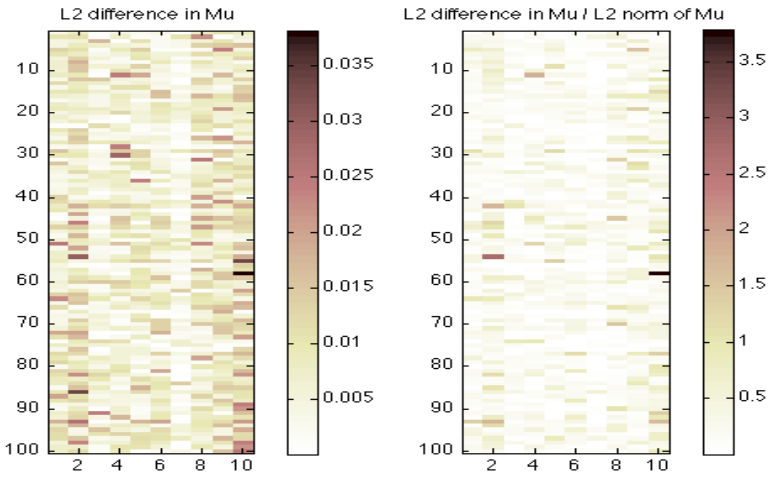

Figure 6: Heatmap of L^2 differences (absolute and relative) between entries of true μ and estimated μ.

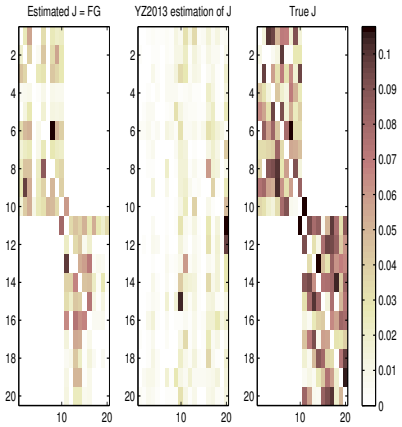

Figure 7: Left: our proposed estimation. Center: estimation following [19]. Right: true J.

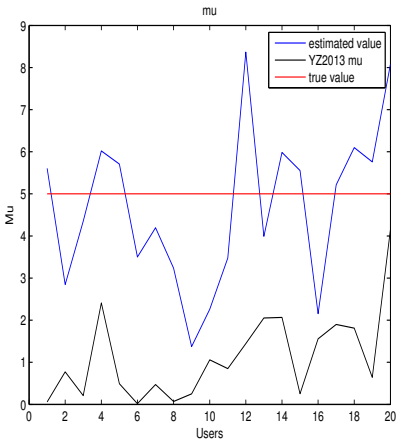

Figure 8: Plot of μ for comparison with [19].

unknown topics (unlabeled data). The overhead of introducing topic models into the Hawkes models is minimal and allows a much more data-driven way of discovering the hidden influences on social networks, for which modified collapsed Gibbs sampling and variational Bayes techniques are derived;

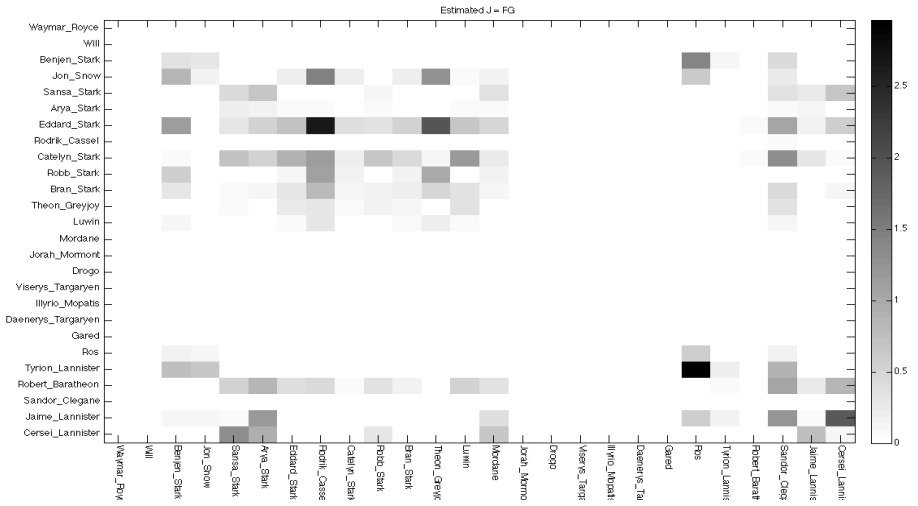

Figure 9: Game of Thrones influence heatmap.

moreover, the generality of these topic models also simplifies the extension of our framework to any kind of data modeling, such as hierarchical topic models, semi-supervised topic models, nonparametric topic models, spatial topic models, etc. It further allows dynamic social networks and/or various temporal effects of users intrinsic rate of diffusion to be investigated and discovered.

Our estimation algorithms depend on the temporal kernel of the underlying Hawkes process, and a variety of kernels can be used to model different kinds of temporal interactions, for example: close-range interactions with the exponential kernel and long-range interactions with power law kernels. The estimation algorithms remain robust and fast, and parametric estimation algorithms (or with nonparametric ones) for these temporal kernels may also be coupled in our estimation procedure.

The multiplicative updates stemming from the nonnegative tensor factorization are also appealing: the multiplicative updates derived from the optimization problem are easy to implement, even in a distributed fashion - they are basically matrix products and entrywise operations - with a low complexity on the data, allowing one to perform estimations in real-life social networks, especially if some of the parameters are already known beforehand.

One can also notice that by performing a dimensionality reduction during our nonnegative tensor factorization estimation, we not only estimate the

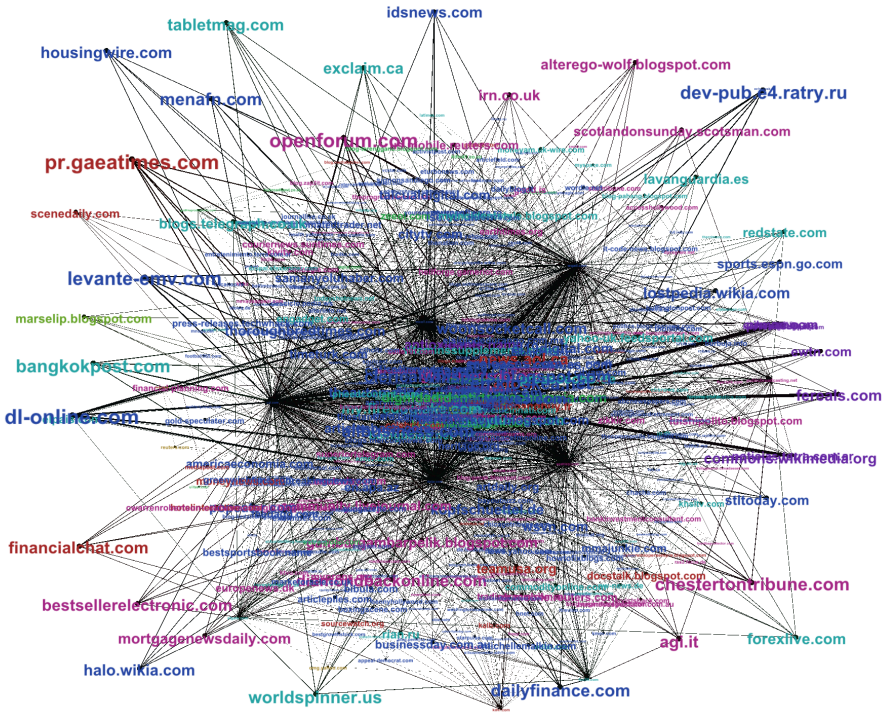

Figure 10: Sites influence graph.

influence that users have on one another but we also acquire information on the communities of the underlying social networks, since we are able to factorize the hidden influence graphs. Here, we used heavily the self-exciting model to retrieve the hidden influence graphs, which is different from other graphs generated by different methods; for example, one could weight the communication graph with the number of messages from one user to its neighbors, but by doing so one looses the temporal character. Moreover, the graphs found by performing this kind of technique are under the assumption that messages influence directly other users, which may not be the case. In our Hawkes framework, the influence is a byproduct of the interaction of users and information, and therefore their influence is probabilistic - it may or may not occur at each broadcast.

Acknowledgment

The authors would like to thank Xavier Bost for the Game of Thrones dataset.

A. Proof of lemma 2

We prove the result only for the tensor S, the calculations for the tensor H are equivalent. Let

$$D_{KL}(Y|SH) = \sum_{j_1,\cdots,j_M} d_{KL}(Y_{j_1,\cdots,j_M}|(SH)_{j_1,\cdots,j_M}),$$

where $d_{KL}(x|y) = x\log(\frac{x}{y}) - x + y$ is the Kullback-Leibler divergence between x and y.

In order to find suitable multiplicative updates for this cost function we proceed in the same manner as in [45, 49], i.e., we find an auxiliary function G such that $G(S, \tilde{S}) \geq D(Y|SH)$ for all nonnegative tensor S and $G(S, S) = D(Y|SH)$, with the NTF updates $S^n,\ n \in \{0, 1, 2, \cdots\}$ of the form

$$S^{n+1} = argmin_{X\geq 0} G(X, S^n). \tag{11}$$

We have thus

$$D(Y|S^{n+1}H) \leq G(S^{n+1}, S^n) = \min_{\tilde{S}\geq 0} G(\tilde{S}, S^n) \leq G(S^n, S^n) = D(Y|S^nH).$$

Let $\mathcal{J} = \{j_1, \cdots, j_M\}$, $\mathcal{S} = \{i_{s_1}, \cdots, i_{s_S}\}$, $\mathcal{H} = \{i_{h_1}, \cdots, i_{h_H}\}$ and $\mathcal{L} = \{l_1, \cdots, l_L\}$ be the index sets for the tensor summations such that $\mathcal{S}\cup\mathcal{H} = \mathcal{J}$ and $\mathcal{J}\cap\mathcal{L} = \emptyset$, and define function G as

$$G(S, \tilde{S}) = \sum_{\mathcal{J}}\sum_{\mathcal{L}} \frac{S_{\mathcal{S},\mathcal{L}}H_{\mathcal{H},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}}} d_{KL}(Y_{\mathcal{J}}|\tilde{Y}_{\mathcal{J}}\frac{S_{\mathcal{S},\mathcal{L}}}{\tilde{S}_{\mathcal{S},\mathcal{L}}})$$

where $\tilde{Y}_{\mathcal{J}} = \sum_{\mathcal{L}} \tilde{S}_{\mathcal{S},\mathcal{L}}H_{\mathcal{H},\mathcal{L}}$ (if $\tilde{S} = S$ then $\tilde{Y}_{\mathcal{J}} = \sum_{\mathcal{L}} S_{\mathcal{S},\mathcal{L}}H_{\mathcal{H},\mathcal{L}} = (SH)_{\mathcal{J}}$).

We easily have that $G(S, S) = D(Y|SH)$. Moreover, by the convexity of $d_{KL}(x|y)$ in y and $\sum_{\mathcal{L}} \frac{\tilde{S}_{\mathcal{S},\mathcal{L}}H_{\mathcal{H},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}}} = 1$, we have that

$$\begin{aligned} G(S, \tilde{S}) &\geq \sum_{\mathcal{J}} d_{KL}(Y_{\mathcal{J}}|\sum_{\mathcal{L}} \frac{\tilde{S}_{\mathcal{S},\mathcal{L}}H_{\mathcal{H},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}}}\tilde{Y}_{\mathcal{J}}\frac{S_{\mathcal{S},\mathcal{L}}}{\tilde{S}_{\mathcal{S},\mathcal{L}}}) \\ &= \sum_{\mathcal{J}} d_{KL}(Y_{\mathcal{J}}|\sum_{\mathcal{L}} S_{\mathcal{S},\mathcal{L}}H_{\mathcal{H},\mathcal{L}}) = D(Y|SH), \end{aligned}$$

thus G is indeed an auxiliary function.

Now, we calculate the multiplicative updates for this auxiliary function as in Eqn. (11). Taking the gradient $\nabla_S G(S^{n+1}, S^n) = 0$ gives us

$$\begin{aligned}
\partial_{S_{\mathcal{S},\mathcal{L}}} G(S^{n+1}, S^n) &= \sum_{\mathcal{J}\setminus\mathcal{S}} \left(1 - \frac{Y_{\mathcal{J}} S^n_{\mathcal{S},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}} S^{n+1}_{\mathcal{S},\mathcal{L}}}\right) H_{\mathcal{H},\mathcal{L}} \\
&= \sum_{\mathcal{J}\setminus\mathcal{S}} H_{\mathcal{H},\mathcal{L}} - \left(\sum_{\mathcal{J}\setminus\mathcal{S}} \frac{Y_{\mathcal{J}}}{\tilde{Y}_{\mathcal{J}}} H_{\mathcal{H},\mathcal{L}}\right) \frac{S^n_{\mathcal{S},\mathcal{L}}}{S^{n+1}_{\mathcal{S},\mathcal{L}}} \\
&= \partial^+_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H) - \partial^-_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H) \frac{S^n_{\mathcal{S},\mathcal{L}}}{S^{n+1}_{\mathcal{S},\mathcal{L}}} = 0,
\end{aligned}$$

which easily implies $S^{n+1}_{\mathcal{S},\mathcal{L}} = S^n_{\mathcal{S},\mathcal{L}} \times \frac{\partial^-_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H)}{\partial^+_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H)}$, the multiplicative updates of lemma 2.

B. Variational Bayes for author-topic model

This appendix is dedicated to the variational Bayes updates for the author-topic model. This approach is similar to the one used in LDA [26] and is a particular case of [33], so the calculations will be omitted. Let us recall that in our framework, every message has only one author, thus the author latent variables $x^{t,w}$ do not interfere.

For every user $a \in V$ we define free variational Dirichlet variables $\gamma^a = (\gamma^a_1, \cdots, \gamma^a_K)$, $\gamma^a_k \geq 0$, for the author-topic latent random variables θ^a and for every word i in the message broadcasted at time t_s we define free variational discrete variables $\psi^{s,i}$ for the word-topic latent random variables $z^{s,i}$, where $\sum_k \psi^{s,i}_k = 1$ and $\psi^{s,i}_k \geq 0$. We can thus retrieve the random variables θ^a, $z^{s,i}$ as $\theta^a \sim Dirichlet(\gamma^a)$ and $z^{s,i} \sim Discrete(\psi^{s,i})$.

Applying the same methods as in appendix $A.3$ in [26], we have the updates for the free variational parameters

$$\begin{aligned}
\gamma^a_k &= \alpha_k + \frac{\sum_{s\in A_a} \sum_{w=1}^{N_s} \psi^{s,w}_j}{\sharp A_a}, \\
\psi^{s,w}_k &\propto \beta_{k,v_w} \exp(\Psi(\gamma^{a_s}_k) - \Psi'(\sum_{k'} \gamma^{a_s}_{k'})) \text{ and} \\
\beta_{k,j} &\propto \sum_s \sum_{i=1}^{N_s} \psi^{s,i}_k w^{s,i}_j,
\end{aligned} \tag{12}$$

where a_s is the author of message s, $A_a = \{s \mid a_s = a\}$, v_w is the index for word w at the dictionary and $\sum_j \beta_{k,j} = 1$.

If we consider $\beta_k \sim Dirichlet(\eta)$ and use a free variational parameter ρ_k for each β_k, we get (see [26])

$$\rho_{k,j} = \eta_j + \sum_{s=1} \sum_{i=1}^{N_s} \psi_k^{s,i} w_j^{s,i}.$$

For the hyperparameters α and η, one can proceed as in [26] to find a Newton-Raphson algorithm to find the optimal values.

References

[1] S. Bikhchandani, D. Hirshleifer, I. Welch, A theory of fads, fashion, custom and cultural change as information cascades, Journal of Political Economy 100 (1992) 992–1026.

[2] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the spread of influence through a social network, In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (2003) 137–146.

[3] D. Centola, The spread of behavior in an online social network experiment, Science 329 (5996).

[4] M. Gomez-Rodriguez, J. Leskovec, B. Schölkopf, Modeling information propagation with survival theory, in: ICML, 2013.

[5] T. M. Snowsill, N. Fyson, T. D. Bie, N. Cristianini, Refining causality: who copied from whom?, ACM SIGKDD 2011 (2011) 466–474.

[6] R. J. Ypma, A. M. Bataille, A. Stegeman, G. Koch, J. Wallinga, W. M. van Ballegooijen, Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data, Procedings of the Royal Society B 279 (2012) 444–450.

[7] M. Gomez-Rodriguez, J. Leskovec, A. Krause, Inferring networks of diffusion and influence, ACM Transactions on Knowledge Discovery from Data (TKDD) 5 (4) (2012) 21.

[8] K. Zhou, H. Zha, L. Song, Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes, Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS) (2013) 641–649.

[9] J. Goldenberg, B. Libai, E. Muller, Talk of the network: A complex systems look at the underlying process of word-of-mouth, Marketing Letters 12 (3) (2001) 211–223.

[10] J. Goldenberg, B. Libai, E. Muller, Using complex systems analysis to advance marketing theory development, Academy of Marketing Science Review (2001) 19.

[11] S. Myers, J. Leskovec, Clash of the contagions: Cooperation and competition in information diffusion, IEEE International Conference On Data Mining (ICDM) (2012) 10.

[12] S. Myers, J. Leskovec, C. Zhu, Information diffusion and external influence in networks, KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2012) 33–41.

[13] P. O. Perry, P. J. Wolfe, Point process modelling for directed interaction networks, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (5) (2013) 821–849.

[14] C. Blundell, J. Beck, K. A. Heller, Modelling reciprocating relationships with Hawkes processes, Advances in Neural Information Processing Systems (NIPS) (2012) 2600–2608.

[15] T. Iwata, A. Shah, Z. Ghahramani, Discovering latent influence in online social activities via shared cascade Poisson processes, In Proceedings of the nternational conference on Knowledge discovery and data mining (KDD) (2013) 266–274.

[16] D. J. Daley, D. Vere-Jones, An introduction to the theory of point processes, Springer series in Statistics, Springer, 2005.

[17] A. G. Hawkes, Spectra of some self-exciting and mutually exciting point processes, Biometrika 58 (1971) 83–90.

[18] T. Liniger, Multivariate Hawkes processes, ETH Doctoral Dissertation (2009) 265.

[19] S.-H. Yang, H. Zha, Mixture of mutually exciting processes for viral diffusion, Proceedings of the 30th International Conference on Machine Learning (ICML) (2013) 1–9.

[20] R. Dawkins, The Selfish Gene, 2nd Edition, Oxford University Press, 1989.

[21] Y. Ogata, Space-time point-process models for earthquake occurrences, Annals of the Institute of Statistical Mathematics 50 (2) (1998) 379–402.

[22] G. N. Borisyuk, R. M. Borisyuk, A. B. Kirillov, E. I. Kovalenko, V. I. Kryukov, A new statistical method for identifying interconnections between neuronal network elements, Biological Cybernetics 52 (1985) 301–306.

[23] E. Bacry, S. Delattre, M. Hoffmann, J.-F. Muzy, Modelling microstructure noise with mutually exciting point processes, arXiv:1101.3422v1.

[24] L. Li, H. Zha, Dyadic event attribution in social networks with mixtures of Hawkes processes, Proceedings of 22nd ACM International Conference on Information and Knowledge Management (CIKM) (2013) 1667–1672.

[25] R. Crane, D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, Proceedings of the National Academy of Sciences 105 (41) (2008) 15649–15653.

[26] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation, Journal of machine Learning research 3 (2009) 993–1022.

[27] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth, The author-topic model for authors and documents, Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI '04) (2004) 487–494.

[28] D. M. Blei, J. D. Lafferty, Topic models, Notes.

[29] W. R. Gilks, S. Richardson, D. Spiegelhalter, Markov Chain Monte Carlo in Pratice, Interdisciplinary Statistics, Chapman & Hall/CRC, 1999.

[30] W. M. Darling, A theoretical and practical implementation tutorial on topic modeling and gibbs sampling, Technical report (2011) 1–10.

[31] T. Griffiths, Gibbs sampling in the generative model of latent Dirichlet allocation, Technical report (2002) 1–3.

[32] M. D. Hoffman, D. M. Blei, F. Bach, Online learning for latent Dirichlet allocation, Advances in Neural and Information Processing Systems (NIPS) 13 (2010) 856–864.

[33] G. Heinrich, M. Goesele, Variational Bayes for generic topic models, KI 2009: Advances in Artificial Intelligence Lecture Notes in Computer Science 5803 (2009) 161–168.

[34] P. Holme, J. Saramäki, Temporal networks, Physics Reports 519 (3) (2012) 97–125.

[35] P. Holme, J. Saramäki, (eds.), Temporal Networks, Springer, Berlin, 2013.

[36] V. Krishnamurthy, A. d'Aspremont, Convex algorithms for nonnegative matrix factorization, ArXiv: 1207.0318.

[37] Y. Ogata, On lewis simulation method for point processes, IEEE Transactions on Information Theory 27 (1) (1981) 23–31.

[38] J. C. Louzada Pinto, T. Chahed, Modeling user and topic interactions in social networks using Hawkes processes, 8th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS '14) (2014) 58–65.

[39] S. A. Pasha, V. Solo, Hawkes-Laguerre dynamic index models for point processes, Proceedings of 52nd IEEE Conference on Decision and Control (CDC).

[40] S. A. Pasha, V. Solo, Hawkes-Laguerre reduced rank model for point process, Proceedings of 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2013) 6098–6102.

[41] J. C. Louzada Pinto, T. Chahed, Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes

processes, The 10th International Conference on Signal Image Technology and Internet Based Systems (SITIS '14).

[42] Y. Ogata, The asymptotic behaviour of maximum likelihood estimators for stationary point processes, Annals of the Institute of Statistical Mathematics 30:Part A (1978) 243–261.

[43] K. Zhou, H. Zha, L. Song, Learning triggering kernels for multi-dimensional Hawkes processes, Proceedings of the 30th International Conference on Machine Learning (ICML), 2013.

[44] D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[45] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the β-divergence, Neural Computation 23 (9) (2011) 2421–2456.

[46] Y.-D. Kim, S. Choi, Nonnegative tucker decomposition, Proceedings of 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[47] A. Cichocki, R. Zdunek, A.-H. Phan, S. Amari, Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, John Wiley & Sons, Ltd, 2009.

[48] D. D. Lee, H. S. Seung, Algorithms for non-negative matrix factorization, Advances in Neural and Information Processing Systems (NIPS) 13 (2001) 556–562.

[49] R. Kompass, A generalized divergence measure fon nonnegative matrix factorization, Neural Computation 19 (3) (2007) 780–791.

[50] H. Minc, Nonnegative Matrices, John Wiley & Sons, New York, NY, USA, 1988.

[51] C. M. Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, Springer-Verlag, 2006.

[52] K. R. Canini, L. Shi, T. L. Griffiths, Online inference of topics with latent Dirichlet allocation, In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) (2009) 65–72.

[53] A. Asuncion, M. Welling, P. Smyth, Y. W. Teh, On smoothing and inference for topic models, In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI) (2009) 27–34.

[54] H. M. Wallach, D. Mimno, A. McCallum, Rethinking lda: Why priors matter, In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS).

[55] T. P. Minka, Estimating a dirichlet distribution, Notes.

[56] E. Lewis, G. O. Mohler, A nonparametric EM algorithm for multiscale Hawkes processes, Journal of Nonparametric Statistics (2011) 1–16.

[57] P. F. Halpin, An EM algorithm for Hawkes process, Proceedings of the 77th Annual Meeting of the Psychometric Society.

[58] E. Bacry, J.-F. Muzy, Second order statistics characterization of Hawkes processes and non-parametric estimation, ArXiv: 1401.0903v1.

[59] A.-H. Phan, A. Cichocki, R. Zdunek, T. Vu-Dinh, Novel alternating least squares algorithms for nonnegative matrix and tensor factorizations 6443 (2010) 262–269.

[60] A. Cichocki, R. Zdunek, Regularized alternating least squares algorithms for non-negative matrix/tensor factorizations, in: Advances in Neural Networks ISNN 2007, Vol. 4493 of Lecture Notes in Computer Science, 2007, pp. 793–802.

[61] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, S.-I. Amari, Novel multilayer nonnegative tensor factorization with sparsity constraints, Springer LNCS 4432 (2007) 271–280.

[62] C.-J. Lin, Projected gradient methods for nonnegative matrix factorization, Neural Computation 19 (10) (2007) 2756–2779.

[63] A. Cichocki, R. Zdunek, S.-I. Amari, Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization, Springer LNCS 4666 (2007) 169–176.

[64] J. Marot, S. Bourennane, Fast tensor signal filtering using fixed point algorithm, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2008) 921–924.

[65] R. Zdunek, A. Cichocki, Nonnegative matrix factorization with constrained second-order optimization, Signal Processing 87 (8) (2007) 1904–1916.

[66] A.-H. Phan, A. Cichocki, Extended HALS algorithm for nonnegative tucker decomposition and its applications for multi-way analysis and classification, Neurocomputing 74 (2011) 1956–1969.

[67] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, Z. He, Extended SMART algorithms for non-negative matrix factorization, Springer LNAI 4029 (2006) 548–562.

[68] A. Cichocki, R. Zdunek, S. Amari, Nonnegative matrix and tensor factorization, IEEE Signal Processing Magazine 25 (1) (2008) 142–145.

[69] A. G. Hawkes, D. Oakes, A cluster process representation of a self-exciting point process, J. Appl. Prob. 11 (1974) 493–503.

[70] E. J. Dudewicz, S. N. Mishra, Modern Mathematical Statistics, Wiley, 1988.