A Factorization Approach for Survival Analysis on Diffusion Networks

Giuseppe Manco, Ettore Ritacco and Nicola Barbieri

Abstract—In this paper we propose a survival factorization framework that models information cascades by tying together social influence patterns, topical structure and temporal dynamics. This is achieved through the introduction of a latent space which encodes: (a) the relevance of an information cascade on a topic; (b) the topical authoritativeness and the susceptibility of each individual involved in the information cascade, and (c) temporal topical patterns. By exploiting the cumulative properties of the survival function and of the likelihood of the model on a given adoption log, which records the observed activation times of users and side-information for each cascade, we show that the inference phase is linear in the number of users and in the number of adoptions. The evaluation on both synthetic and real-world data shows the effectiveness of the model in detecting the interplay between topics and social influence patterns, which ultimately provides high accuracy in predicting users activation times.

Index Terms—Social Influence, Information Diffusion, Social Network Analysis, Community Detection

1 Introduction

Social network platforms provide sharing and reposting functionalities that facilitate the diffusion of information through the network, by enabling users to simultaneously share information with their social peers and triggering a cascade of adoptions. An information cascade is a social process for adoptions, where the decision of each individual depends on the decision of people who have adopted the same content earlier. Such cascades have been identified in settings such as blogging, e-mail, product recommendation, and social Web platforms. The availability of large-scale, time-resolved cascade data on the social Web allows the study of interesting questions, such as: (i) How does information spread on networks? (ii) How far and fast does information flow? (iii) What is the network structure upon that allows the diffusion of information? (iv) How does the network structure affect information flow (and viceversa)? (v) How does the content being propagated affect the structure and shape of information cascades?

In this work we are mainly interested on the latter research question. Understanding the structural, topical and temporal dynamics of information cascades can provide insights on the complex patterns that govern the information propagation process and it can be used to forecast future events. The problem of inferring the topical, temporal and network properties that characterize an observed set of information cascades is complicated by the fact that the diffusion network, transmission rates and the topical structure are hidden. In most cases of interests, we only observe users activation times (e.g. the time at which a user re-shares a tweet, purchases an item, etc.) and we are given some meta information about the cascade (e.g. hashtags associated with a tweet, textual description for an item).

To infer the diffusion network and the topical structure jointly, a natural approach is to model user's activation times as continuous random variables. Then, we can assume that those variables are generated by a stochastic process that depends on pairwise transmission rates $\lambda_{u,v}$ (which explains the influence exerted by user v on u) and on the topical distribution of each information cascade. A straightforward realization of such idea is to introduce topical pairwise transmission rates $\lambda^k_{u,v}$, where k denotes a topic index (see e.g. [30]). This approach has three main drawbacks: it introduces a large number of parameters (and hence prone to overfitting); the inference phase that does not scale well; it may produce poor estimates if the episodes of information propagation from v to u are limited.

To address these issues, in this paper we introduce a stochastic model that factorizes pairwise transmission rates in terms of general user authoritativeness and susceptibility on a set of topics of interest. According to such a principle, both the side-information and temporal dynamics observed on a given information cascade are explained by 3 low-dimensional latent factors that encode: (i) the topical authority of each user $A_{v,k}$, (ii) the topical susceptibility $S_{u,k}$ and (iii) the relevance of side information w (e.g. hashtag) on topic k, $\varphi_{w,k}$.

This framework draws from recent investigations [3], [5], [24], which explain information propagation in terms of authoritativeness and susceptibility. In particular, each user exerts a degree of influence or susceptibility according to a set of topics which also represent her interests. The participation to an information cascade, referred in the next as *adoption* or *activation*, can hence be explained as the effect of such degrees of peers' influence and user's susceptibility on the topic that best represents the content of the information being propagated.

Notably, the computational complexity of the proposed approach is linear in the number of users and in the size of the adoption log (number of users who participated in

- *G. Manco E. Ritacco are researchers of the Institute of High Performance Computing and Networking (ICAR-CNR), via P. Bucci cubi 7/11 C, 87036 Arcavacata di Rende (CS) - Italy.*
  Email: {giuseppe.manco, ettore.ritacco}@icar.cnr.it
- *N. Barbieri, email: nicolabarbieri1@gmail.com*

the information cascade). The inference phase is based on the above mentioned EM scheme for learning the model parameters. By exploiting the properties of the underlying propagation models, each iteration of the learning procedure can be efficiently computed in at most three scans of the adoption log.

The main contributions of this work can be summarized as follows.

- We review previous studies on information diffusion (Sec. 2) and briefly introduce a general framework for modeling information diffusion cascade data via survival analysis (Sec. 3.1).
- Next, we introduce a factorization model (Sec. 3.2) that expresses topical pairwise transmission rates in terms of user's authority, susceptibility on topics of interests, by coupling the topical content of a cascade and the observed users activation times.
- We devise an expectation maximization algorithm (Sec. 4) for learning the parameters of the model. By fully exploiting the cumulative properties of the data likelihood under the proposed model, each iteration of the learning procedure can be efficiently computed in at most three scans of the propagation log.
- We run an extensive evaluation (Sec. 5) on both synthetic and real-world data. We assess the capability of the model in detecting the interplay between the topical structure and temporal dynamics, which finally provides an accurate characterization of users' behavior.

2 Related work

Starting from seminal studies [16], [21], [28], [33], the research on information diffusion and influence propagation has been mainly focused on determining how information spreads across pairs of users, observing the social network structure and the adoption log. A recent line of research [14], [15] studies a different perspective, where the social network is not given as input, and the problem is how to uncover the hidden network structure starting from the log of users activity. This problem is addressed by assuming that infections follow a continuous-time independent cascade model: active nodes try independently to activate inactive peers and each node becomes active once the first parent infects it. For example, in NetRate [14], if node u succeeds in activating v, then the contagion of the latter happens after an incubation period sampled from a chosen distribution. The latter defines the conditional likelihood of transmission between each pair of nodes and it actually depends on the difference of their activation times. According to this propagation model, the likelihood of a propagation cascade can be formulated by applying standard survival analysis [22], in terms of survival (which models the probability that a node survives uninfected until a time T) and hazard functions (which models instantaneous infections). More recent extensions of the diffusion process based on survival analysis exploit non-parametric methods based on kernels [11], or more sophisticated modeling through Poisson and Hawkes processes [9], [19], [34].

A different line of research extends the diffusion process discussed above by considering enhancements based on features [29], or topics which characterize cascades [6], [10], [17], [18], [30]. These models assume that the speed of the diffusion process can depends on several factors, including connections between nodes as well as other features characterizing either the users or the cascades. In particular, in the context of network reconstruction, the strength of connections can depend on topical affinity between nodes [10], [17], [30].

Recent works have also focused on alternative ways of representing interactions between nodes, using latent-dimensional embedding techniques. In [8] authors propose a framework based on a *heat diffusion process* which projects each node into a latent space where the proximity between a pair of nodes reflects the proximity of their activations times in the observed cascades. The embedding space models how information diffuses and determines which users will be contaminated by a particular content, given the identity of the source node for the diffusion and the features that characterize the information being diffused.

The approaches described so far do not explicitly consider the diffusion process as a result of the interaction between influence and susceptibility. In [4], [6], the probability of activation is modeled as the effect of the influence of neighbor nodes within the cascades and/or the network. Furthermore, the approaches [3], [31] propose factorization techniques which associate two low-dimensional vectors to each node, representing influence and susceptibility. The propagation probability that one user forwards information depends on the product of her activated neighbors' influence vectors and her own susceptibility vector. The drawback of these approaches is that they only model cascades in a discrete-time scenario.

Table 1 compares the approach proposed in this work and some paradigmatic approaches mentioned above, by considering the following dimensions: modeling of time (continuous vs. discrete), whether they require as input the underlying network, complexity of the inference phase, modeling of side information, whether they are able to detect clustering structure. By denoting with N, M the number of nodes and cascades, we can see that all methods based on pairwise transmission rates suffer from the drawback of quadratic complexity in the learning phase. Thus, they do not scale to a large number of users and cascades. In the experiments in Section 5 we use NetRate as baseline to evaluate the capabilities of our approach.

By contrast, linear methods only model discrete time, and they do not necessarily model side information. To the best of our knowledge, our method is the only capable of combining the advantages of linear complexity and comprehensive modeling of temporal dynamics.

3 Modeling Information Diffusion

3.1 Background

Notation. A cascade represents the propagation of a piece of information (e.g. news, post, meme, etc.) over a set of nodes (e.g., users of the system). We can specify each cascade as the activation times of a set of nodes $\mathcal{V}$ with cardinality N (i.e., $|\mathcal{V}| = N$). Formally, $\mathbf{t}^c$ can be represented as a N-dimensional vector $\mathbf{t}^c = (t_1(c), \cdots, t_N(c))$, where $t_u(c) \in [0, T^c] \cup \{\infty\}$ represents the timestamp at which the

	Time	*Req. Network*	*Inference*	*Side Info*	*Clustering*
NetRate [14]	contin.	no	$O(N^2)$	no	no
MONET [29]	contin.	yes	$O(N^2)$	nodes	no
MMRate [30]	contin.	no	$O(N^2)$	no	cascades
CSDK [8]	contin.	no	$O(NM)$	cascades	no
LIS [31]	discrete	no	$O(N^2)$	no	no
AIR [4]	discrete	yes	$O(N)$	cascades	nodes
CCN [3]	discrete	yes	$O(N)$	no	nodes
CWN [6]	both	no	$O(N)$	no	nodes
Our method	contin.	no	$O(N)$	cascades	cascades

TABLE 1: Comparison of the proposed method to the state of the art.

node u becomes active on the cascade $\mathbf{t}^c$. For instance, if each cascade refers to the propagation of a meme, $t_u(c)$ will represent the timestamp at which user u re-posted meme c. Without loss of generality, we can assume that each cascade starts at timestamp 0; moreover, $t_u(c) = \infty$ encodes the fact that the node u has not been infected during the observation window $[0, T^c]$. Let $\mathcal{V}^+(c)$ denote the set of active nodes on the cascade c (i.e., $t_u(c) \neq \infty$), while $\mathcal{V}^-(c) = \mathcal{V} \setminus \mathcal{V}^+(c)$ denotes the set of inactive nodes. The term N_c denotes the size of $\mathcal{V}^+(c)$.

Let $\mathbf{w}^c$ denote side information on the cascade c. We represent such information as a bag-of-words $\mathbf{w}^c = \{w_1, \cdots, w_{len(c)}\}$, where each w_i is a word from a dictionary $\mathcal{W}$ and $len(c)$ is the number of words associated with the cascade $\mathbf{c}$.

Finally, let $\mathcal{C} = \{(\mathbf{t}^1, \mathbf{w}^1) \cdots (\mathbf{t}^M, \mathbf{w}^M)\}$ denote a collection of M cascades over $\mathcal{V}$.

Propagation model. In our setting, we assume that *(i)* an event can trigger further events in the future, within the same cascade; *(ii)* events in different cascades are independent from each other. That is, a node v can trigger the activation of a node u on cascade c if and only if $t_v(c) < t_u(c)$. Hence, each cascade $\mathbf{t}^c$ defines a directed-acyclic graph, where $par_u(c) = \{v \in \mathcal{V} : t_v(c) < t_u(c)\}$. In the following we will use the notation $v \prec_c u$ to represent that v is a potential influencer for the activation of u within the cascade c, i.e. $v \in par_u(c)$.

Similar to the Independent Cascade model [21], we assume that nodes' activations are binary (either active or inactive), *progressive* (an active node cannot turn inactive in the future) and all the parents try to infect their child nodes independently. Based on such assumptions, we can model each cascade by expressing the likelihood of activation times for active nodes and the likelihood that the adoption did not happen by time T^c for inactive nodes, according to a chosen propagation model.

Survival analysis for diffusion cascades. Let T denote a non-negative random variable representing the time of occurrence on an event. We can assume that for each pair of nodes (v, u) such that v triggered u's activation within the considered cascade c, there is a dependency between the respective activation times. Following [14], we formalize such dependency by introducing a conditional pairwise transmission likelihood $f(t_u(c)|t_v(c), \lambda_{v,u})$ which depends on the delay $\Delta^c_{u,v} = t_u(c) - t_v(c)$ between activation times and on the transmission rate $\lambda_{v,u}$. Then, the likelihood of observing the activation times within a cascade can be formulated by applying a survival analysis framework [14]:

$$\begin{aligned}\Pr(\mathbf{t}^c|\Theta) = &\prod_{u\in\mathcal{V}^-(c)}\prod_{v\in\mathcal{V}^+(c)} S(T^c - t_v(c);\lambda_{v,u})\cdot\\ &\prod_{u\in\mathcal{V}^+(c)}\prod_{v\prec_c u} S(\Delta^c_{u,v};\lambda_{v,u})\cdot\sum_{v'\prec_c u} h(\Delta^c_{u,v'};\lambda_{v',u}),\end{aligned} \tag{1}$$

where the *survival function* $S(t - t';\lambda) = \Pr(T \geq t|t', \lambda) = 1 - \int_{t'}^{t} f(x|t', \lambda)dx$ encodes the probability that an event does not occur by time t and the *hazard function* $h(t - t'|\lambda) = \frac{f(t|t',\lambda)}{S(t-t'|\lambda)}$ is the rate of instantaneous infection at time t. Similarly, let W denote a random variable over words in $\mathcal{W}$; we can consider $\mathbf{w}^c$ as a collection of $len(c)$ i.i.d draws from a distribution Φ over $\mathcal{W}$:

$$\Pr(\mathbf{w}^c|\Phi) = \prod_{w\in\mathbf{w}^c}\Pr(w|\Phi). \tag{2}$$

In the following we will build upon such a basic model and propose a factorization technique for jointly modeling temporal dynamics and side information that characterize each cascade.

3.2 Factorization Model

We start from the idea that the temporal dynamics, governing the activations of each node within observed cascades, depend on a set of *hidden* topics. The propagation of a piece of information depends inherently on its content and on pairwise transmission that are topic-dependent. The goal of our framework is to jointly factorize activation times and side information about each cascade to detect a finite set of K topics (where K is given as input), representing both a diffusion pattern and thematic information about the content.

This setting presents two challenges. First, in many practical scenarios we observe only node activations within a cascade, with no knowledge about what (or who) triggered them. Secondly, we observe side information and activation times of nodes within a set of cascades, but both the topical-structure and the relationships between topics and pairwise transmission rates are hidden.

To infer hidden topics and diffusion patterns we will introduce a generative process. As aforesaid, $\mathcal{C}$ is governed by a mixture of K underlying topics. Such a mixture is specified by introducing binary random variables $z_{c,k}$ which denote the membership of the cascade within each topic, with the constraint $\sum_{k=1}^{K} z_{c,k} = 1$. Let $\mathbf{Z}$ denote the overall $M \times K$ hidden topic assignments matrix. We characterize each topic k with the following 3 *non-negative* components:

- $A_{u,k}$, the authority degree of node u (i.e. tendency of triggering the activation of other nodes);
- $S_{u,k}$, the susceptibility degree of node u (i.e., tendency of being influenced by other nodes);
- $\varphi_{w,k}$, the relevance of word w.

Our factorization model is based on the assumption that the pairwise transmission rates within topic k can be factorized as a linear combination of users' authority and susceptibility components:

$$\lambda_{v,u,k} = A_{v,k} \cdot S_{u,k} \tag{3}$$

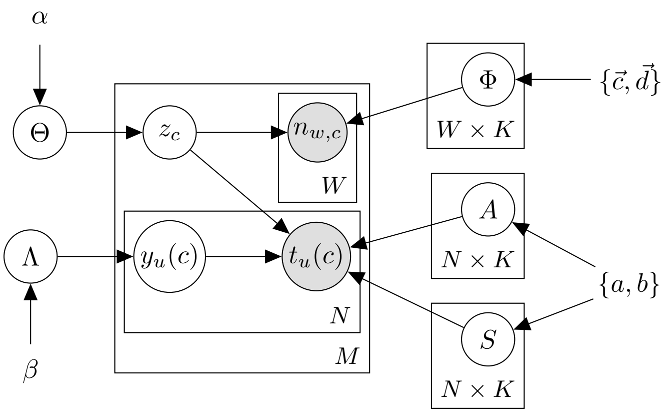


Fig. 1: Graphical model of Survival Factorization.

The generation of a cascade unfolds as follows. First, we pick a topic z_c which specifies a topical-diffusion pattern, by drawing upon a multinomial distribution over topics $\Theta = \{\pi_1, \ldots, \pi_k\}$. Then, we adopt a *Poisson language model* [25] to generate the side-information by drawing the number of occurrences of each term w in the cascade c, shorted as $n_{w,c}$ from a *Poisson* distribution governed by the parameter set $\mathbf{\Phi}_k = \{\varphi_{w,k}\}_{w \in \mathcal{W}}$. Finally, the observed activation times within a cascade are generated according to a survival model. A summary of the conditional dependencies between latent and observed variables in our model is given in Fig. 1 and discussed below.

The modeling of activation times for each node in the cascade assumes that the delay between the influencer v and the influenced u ($t_v(c) < t_u(c)$) is generated accordingly to a *Weibull* distribution, whose scale parameter is the transmission rate, while the shape ρ is fixed:

$$f(t_u(c)|t_v(c), \lambda_{v,u,k}) = \mathcal{W}eib(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho). \tag{4}$$

Here, $\mathcal{W}eib(t; \rho, \lambda) = \rho\lambda t^{\rho-1} e^{-\lambda t^\rho}$. Different choices of ρ correspond to different assumptions about the hazard: the hazard is rising if $\rho > 1$, constant if $\rho = 1$ (exponential model), and declining if $\rho < 1$. The corresponding survival and hazard functions are:

$$h(t; \lambda, \rho) = \rho\lambda t^{\rho-1}\,, \quad (5) \qquad S(t; \lambda, \rho) = e^{-\lambda t^\rho}\,. \quad (6)$$

As stated above, we only observe activation times but not who triggered the activation. To model the hidden influencer for the activation of each node u within a cascade, we introduce latent binary variables $y^c_{u,v}$, with the constraint $\sum_{v \in \mathcal{V}} y^c_{u,v} = 1$. Let $\mathbf{Y}$ denote a $M \times N \times N$ binary matrix, where $y^c_{u,v} = 1$ represents the fact that node v triggered the activation of node u in the cascade c. For each pair of users u and v, the prior probability that $y^c_{u,v} = 1$ is governed by a multinomial distribution Λ over all possible v's.

In its general formulation, the model includes hyperparameters $\alpha, \beta, a, b, \vec{c}, \vec{d}$, where α and β are devised as paramters of the Dirichlet distribution, and a, b and $\vec{c}, \vec{d}$ parameterize a Gamma distribution. These components can model prior assumptions about the most likely values for the parameters $\Theta, \Lambda, \Phi, \mathbf{A}$ and $\mathbf{S}$. For example, one could assume that the most active users are more likely to be trigger activations and choose a β value that produces a Λ skewed towards such users. In the rest of this section, we simplify the model and devise uniform prior assumptions: for example, with reference to Λ, each each v has equal chances of activating u. A more detailed treatment of the whole model in a full bayesian setting is outlined in section 4.1.

Given the status of the hidden variables $\mathbf{Z}$ and $\mathbf{Y}$, we can finally formalize the likelihood of observing the activation times within a cascade c:

$$\Pr(\mathbf{t}^c|\mathbf{Z}, \mathbf{Y}, \mathbf{A}, \mathbf{S}) = \prod_k \Pr(\mathbf{t}^c|\mathbf{Y}, \mathbf{A}_k, \mathbf{S}_k)^{z_{c,k}} \tag{7}$$

where:

$$\begin{aligned} &\Pr(\mathbf{t}^c|\mathbf{Y}, \mathbf{A}_k, \mathbf{S}_k) = \\ &\quad \prod_{u \in \mathcal{V}^-(c)} \prod_{v \in \mathcal{V}^+(c)} S(T^c - t_v(c); \lambda_{v,u,k}, \rho) \\ &\quad \cdot \prod_{u \in \mathcal{V}^+(c)} \prod_{v \prec_c u} h(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho)^{y^c_{u,v}} \cdot S(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho). \end{aligned} \tag{8}$$

Finally, the overall likelihood of all cascades is

$$\Pr(\{\mathbf{t}^1, \cdots, \mathbf{t}^M\}|\mathbf{Z}, \mathbf{Y}, \mathbf{A}, \mathbf{S}) = \prod_{c=1}^{M} \Pr(\mathbf{t}^c|\mathbf{Z}, \mathbf{Y}, \mathbf{A}, \mathbf{S})\,.$$

Compared to the modeling in eq. 1, the above model exhibits two main differences. First, cascade are characterized by a topic which also governs the propagation rate. Second, we explicitly model influencers by introducing the $\mathbf{Y}$ matrix. In fact, eq. 7 is a refined extension of eq. 1, since the latter can be obtained from the former by assuming $K = 1$ and marginalizing over $\mathbf{Y}$.

Likelihood of side-information. The probability of observing content $\mathbf{w}^c$ under topic k is given by the probability of observing the frequency count $n_{w,c}$ of each word. Within the *homogeneous Poisson model* [25], this frequency under topic k follows a Poisson distribution with parameter $\varphi_{w,k}$. The latter is the expected number of occurrences of w in a unit of time, and the time associated to the generation of side-information $\mathbf{w}^c$ is assumed to be $|\mathbf{w}^c| = len(c)$. Thus, according to this model, the likelihood of observing a bag-of-words $\mathbf{w}^c$ when the topic is k can be expressed as:

$$\Pr(\mathbf{w}^c|\mathbf{\Phi}_k) = \prod_w \frac{(|\mathbf{w}^c| \cdot \varphi_{w,k})^{n_{w,c}} \exp\{-|\mathbf{w}^c| \cdot \varphi_{w,k}\}}{n_{w,c}!}. \tag{9}$$

Since each cascade is generated independently from each other, the overall likelihood of side information over all cascades, given hidden topic-assignment $\mathbf{Z}$, can be expressed as:

$$\Pr(\{\mathbf{w}^1, \cdots, \mathbf{w}^M\}|\mathbf{\Phi}, \mathbf{Z}) = \prod_{c=1}^{M} \prod_k \Pr(\mathbf{w}^c|\mathbf{\Phi}_k)^{z_{c,k}}.$$

4 Inference and Parameter Estimation

Let $\mathbf{\Xi} = \{\mathbf{A}, \mathbf{S}, \mathbf{\Phi}, \mathbf{\Lambda}, \mathbf{\Theta}\}$ denote the status of parameters of the model. Given latent assignments $\mathbf{Z}$ and $\mathbf{Y}$, the conditional data likelihood is:

$$\begin{aligned} \Pr(\mathcal{C}|\mathbf{Z}, \mathbf{Y}, \Xi) = &\Pr(\{\mathbf{t}^1, \cdots, \mathbf{t}^M\}|\mathbf{Z}, \mathbf{Y}, \Xi) \\ &\cdot \Pr(\{\mathbf{w}^1, \cdots, \mathbf{w}^M\}|\mathbf{Z}, \Xi)\,. \end{aligned}$$

Thus, the optimal values for Ξ can be obtained by optimizing the likelihood:

$$\Pr(\mathcal{C}, \mathbf{\Xi}) = \sum_{\mathbf{Z},\mathbf{Y}} \Pr(\mathcal{C}|\mathbf{Z}, \mathbf{Y}, \Xi) \Pr(\mathbf{Z}, \mathbf{Y}, \mathbf{\Xi}) \,. \tag{10}$$

Exact inference is intractable, and we have to resort to heuristic optimization strategies. It turns out that the Expectation Maximization algorithm can be easily adapted for estimating the optimal parameters. The log-likelihood of the observed cascades can be written as

$$\begin{aligned}\mathcal{L}(\mathbf{\Xi}; \mathcal{C}) &= \log \sum_{\mathbf{Z},\mathbf{Y}} \Pr(\mathcal{C}|\mathbf{Z}, \mathbf{Y}, \Xi) \Pr(\mathbf{Z}, \mathbf{Y}|\Xi) \\ &\geq \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z}, \mathbf{Y}) \log \frac{\Pr(\mathcal{C}, \mathbf{Z}, \mathbf{Y}|\Xi)}{q(\mathbf{Z}, \mathbf{Y})} = \mathcal{Q}(q; \mathcal{C}, \mathbf{\Xi}),\end{aligned}$$

where q is an arbitrary instrumental distribution over the latent variables. It can be shown [7] that the lower bound is tight for the exact posterior, i.e.,

$$\arg\max_{q(\mathbf{Z},\mathbf{Y})} \mathcal{Q}(q; \mathcal{C}, \mathbf{\Xi}) = \Pr(\mathbf{Z}, \mathbf{Y}|\mathcal{C}, \mathbf{\Xi}).$$

Hence, the log-likelihood can be maximized iteratively in the usual EM setting by computing the variational approximations given by the following two steps:

E step: estimate the posterior $\Pr(\mathbf{Z}, \mathbf{Y}|\mathcal{C}, \mathbf{\Xi}^{(n-1)})$
M step: exploit the posterior to solve

$$\begin{aligned}\mathbf{\Xi}^{(n)} = \arg\max_{\mathbf{\Xi}} \sum_{\mathbf{Z},\mathbf{Y}} \Pr(\mathbf{Z}, \mathbf{Y}|\mathcal{C}, \mathbf{\Xi}^{(n-1)}) \\ \cdot \log \Pr(\mathcal{C}, \mathbf{Z}, \mathbf{Y}, \Xi)\end{aligned}$$

Both steps are tractable and the estimation produces closed formulas. The details of the derivations can be found in the appendix. In particular, for the E step the estimation of $\Pr(\mathbf{Z}, \mathbf{Y}|\mathcal{C}, \mathbf{\Xi}^{(n)})$ can be decomposed into the specific components, thus yielding

$$\Pr(z_{c,k}, y^c_{u,v}|\mathbf{t}^c, \mathbf{w}^c, \Xi) = \eta^k_{c,u,v} \cdot \gamma_{c,k}$$

where

$$\eta^k_{c,u,v} = \frac{h(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho)}{\sum_{v' \prec_c u} h(\Delta^c_{u,v'}; \lambda_{v',u,k}, \rho)}, \tag{11}$$

$$\gamma_{c,k} = \frac{\Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k) \Pr(\mathbf{w}^c|\Phi_k)\pi_k}{\sum_{k'} \Pr(\mathbf{t}^c|\mathbf{A}_{k'}, \mathbf{S}_{k'}) \Pr(\mathbf{w}^c|\Phi_{k'})\pi_{k'}}. \tag{12}$$

Here, $\gamma_{c,k}$ represents the posterior probability that cascade c is relative to topic k, and $\eta^k_{c,u,v}$ the posterior probability that the activation of u was triggered by v within topic k. The component $\Pr(\mathbf{w}^c|\Phi_k)$ is specified by equation 9, and $\Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k)$ is obtained by marginalizing $\Pr(\mathbf{t}^c|z_c, \mathbf{Y}^c, \mathbf{A}, \mathbf{S})$ in eq. 8 with respect to $\mathbf{Y}$.

For the M step, by plugging η and γ into the expected log-posterior we can solve the optimization step with regards to all the available parameters. In particular, optimal values for $\mathbf{\Theta}$ and $\mathbf{\Phi}$ can be obtained directly:

$$\pi_k = \frac{1}{M} \sum_c \gamma_{c,k} \tag{13}$$

$$\varphi_{w,k} = \frac{\sum_c \gamma_{c,k} n_{w,c}}{\sum_c \gamma_{c,k} |\mathbf{w}^c|} \tag{14}$$

Term	*Definition*	*Term*	*Definition*
$A_{c,u,k}$	$\sum_{v \prec_c u} A_{v,k}$	$S_{c,u,k}$	$\sum_{v \preceq_c u} S_{v,k}$
$\tilde{A}_{c,u,k}$	$\sum_{v \prec_c u} t_v(c) A_{v,k}$	$\tilde{S}_{c,u,k}$	$\sum_{v \preceq_c u} t_v(c) S_{v,k}$
$A_{c,k}$	$\sum_{v \in \mathcal{V}^+(c)} A_{v,k}$	$S_{c,k}$	$\sum_{v \in \mathcal{V}^+(c)} S_{v,k}$
$\tilde{A}_{c,k}$	$\sum_{v \in \mathcal{V}^+(c)} t_v(c) A_{v,k}$	$\tilde{S}_{c,k}$	$\sum_{v \in \mathcal{V}^+(c)} t_v(c) S_{v,k}$
$R_{c,u,k}$	$\sum_{\substack{v \in \mathcal{V}^+(c) \\ u \prec_c v}} (A_{c,v,k})^{-1}$	S_k	$\sum_v S_{v,k}$
$\tilde{R}_{c,v,k}$	$\sum_{\substack{u \in \mathcal{V}^+(c) \\ v \prec_c u}} t_u(c) \left(t_u(c) A_{c,u,k} - \tilde{A}_{c.u,k}\right)^{-1}$	$L_{c,k}$	$\sum_{v \in \mathcal{V}^+(c)} \log S_{v,k}$

TABLE 2: Counters on the cascades.

$$S_{u,k} = \frac{\sum_{c:u \in \mathcal{V}^+(c)} \gamma_{c,k}}{\bar{S}_{u,k}} \tag{17}$$

where

$$\begin{aligned}\bar{S}_{u,k} = &\sum_{c:u \in \mathcal{V}^+(c)} \gamma_{c,k} \left(t_u(c) A_{c,u,k} - \tilde{A}_{c,u,k}\right) \\ &+ \sum_{c:u \in \mathcal{V}^-(c)} \gamma_{c,k} \left(T^c A_{c,k} - \tilde{A}_{c,k}\right)\end{aligned}$$

$$A_{v,k} = \frac{A^{(n-1)}_{v,k} \sum_{c:v \in \mathcal{V}^+(c)} \gamma_{c,k} R_{c,v,k}}{\sum_{c:v \in \mathcal{V}^+(c)} \gamma_{c,k} \bar{A}_{v,k,c}} \tag{18}$$

where

$$\bar{A}_{v,k,c} = \tilde{S}_{c,k} - \tilde{S}_{c,v,k} + T^c(S_k - S_{c,k}) - t_v(c)(S_k - S_{c,v,k})$$

$$\begin{aligned}\log \Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k) = &L_{c,k} - (S_k - S_{c,k})(T^c A_{c,k} - \tilde{A}_{c,k}) \\ &+ \sum_{u \in \mathcal{V}^+(c)} \left\{\log A_{c,u,k} - S_{u,k} \left(t_u(c) A_{c,u,k} - \tilde{A}_{c,u,k}\right)\right\}\end{aligned} \tag{19}$$

Fig. 2: Optimized estimations for the exponential distribution. All equations rely on counters defined in table 2.

Concerning $\mathbf{A}$ and $\mathbf{S}$, the expected likelihood expresses an interdependency which can be resolved by block coordinate ascent optimization:

$$S_{u,k} = \frac{\sum_{c:u \in \mathcal{V}^+(c)} \gamma_{c,k}}{\sum_{c=1}^{M} \sum_{v \prec_c u} \gamma_{c,k} \cdot (\Delta^c_{u,v})^\rho \cdot A_{v,k}} \tag{15}$$

$$A_{v,k} = \frac{\sum_{c:v \in \mathcal{V}^+(c)} \sum_{\substack{u \in \mathcal{V}^+(c) \\ v \prec_c u}} \eta^k_{c;u,v} \cdot \gamma_{c,k}}{\sum_{c:v \in \mathcal{V}^+(c)} \sum_u \gamma_{c,k} \cdot (\Delta^c_{u,v})^\rho \cdot S_{u,k}} \tag{16}$$

We deliberately choose not to optimize the ρ parameter, and to investigate the case $\rho = 1$. In such a case, in fact, the above equations can be further simplified and are amenable an efficient implementation as described in the following section. Similar results also hold for the case $\rho = 2$.

Scaling up the estimation

When $\rho = 1$, the Weibull distribution simplifies to an exponential distribution. In such a case, we can introduce the counters described in table 2 and rewrite the update equations for $\mathbf{A}$ and $\mathbf{S}$ as shown in figure 2. Notice that, in this new formulation, $\eta^k_{u,v}$ does not need to be explicitly computed. In fact, it is decomposed within the update equation of $A_{v,k}$, by resorting to a previous value of the same variable and the $R_{c,u,k}$ counter (see appendix for details). Algorithm 1 describes the overall procedure for estimating the parameters.

Theorem 1. Algorithm 1 has complexity $O(\sum_c N_c \log N_c + nK(N + W + \sum_c N_c))$ time (where n is the total number

Algorithm 1 Optimized Survival Factorization EM

```
Require: $\mathcal{C}$, the number of latent features $K$
Ensure: matrices $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Phi}$
Randomly initialization for $\mathbf{A}$, $\mathbf{S}$, $\mathbf{\Phi}$;
Compute all counters of table 2;
$n \leftarrow 0$
while Increment in Likelihood is negligible do
    for all cascades $c$ and topic $k$ do
        Compute $\gamma_{c,k}$ exploiting $\log \Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k)$ as defined in Eq. 19;
    end for
    for all  topic $k$ do
        Update $\pi_k$ according to Eq.  13;
        for all users $u$ do
            Compute $S_{u,k}$ according to Eq. 17;
        end for
        Update all counters relative to $\mathbf{S}$ as defined in table 2;
        for all users $u$ do
            Compute $A_{u,k}$ according to Eq. 18;
        end for
        Update counters relative to $\mathbf{A}$ as defined in table 2;
        for all words $w$ do
            Compute $\phi_{w,k}$ according to Eq. 14;
        end for
    end for
    $n \leftarrow n + 1$
end while
```

of iterations) and $O(KN)$ space.

PROOF. See appendix. □

4.1 Full Bayesian Inference

With reference of the general model depicted in fig. 1, it is possible to devise a more general inference procedure, based on the marginal distribution $\Pr(\mathcal{C}|\mathcal{H})$, with $\mathcal{H} = \{\alpha, \beta, a, b, \vec{c}, \vec{d}\}$:

$$\Pr(\mathcal{C}|\mathcal{H}) = \int_{\Xi} \sum_{\mathbf{Z},\mathbf{Y}} \Pr(\mathcal{C}|\mathbf{Z}, \mathbf{Y}, \Xi) \Pr(\mathbf{Z}, \mathbf{Y}, \Xi|\mathcal{H}) \, \mathrm{d}\Xi$$

The marginal distribution is useful in several respects including model comparison via *Bayes factors*, or more general inference.

We choose to express priors for Θ and Λ by means of Dirichlet distributions (parameterized by α and β, respectively), Analogously, the priors relative to $\mathbf{A}, \mathbf{S}$ and $\mathbf{\Phi}$ are expressed by Gamma distributions (parameterized by a, b and $\vec{c}, \vec{d}$, respectively).

The conjugacy property (Gamma to both the Weibull and the Poisson distributions, and Dirichlet to multinomial distribution) allows us to approximate the above marginal by exploiting collapsed Gibbs Sampling [7], based on the Markov chain $\{\mathbf{Z}, \mathbf{Y}, \mathbf{A}, \mathbf{S}, \mathbf{\Phi}\}$. In fact, given a status of the Markov chain, we can express the full conditionals in closed form and hence device sampling based on such conditionals.

- Concerning $\mathbf{A}$, we have

$$A_{u,k}|\mathit{Rest} \sim \mathcal{G}\left(\sum_v \sum_c z_{c,k} y^c_{u,v} + a, \right. \\ \left. \sum_{c:u \in \mathcal{V}^+(c)} \sum_{v:u \prec_c v} z_{c,k} (\Delta^c_{v,u})^\rho S_{v,k} + b \right) \tag{20}$$

- Concerning $\mathbf{S}$,

$$S_{u,k}|\mathit{Rest} \sim \mathcal{G}\left(\sum_v \sum_c z_{c,k} y^c_{u,v} + a, \right. \\ \left. \sum_{c=1}^{M} z_{c,k} \sum_{c:v \prec_c u} A_{v,k} (\Delta^c_{u,v})^\rho + b \right) \tag{21}$$

- Concerning $\phi_{w,k}$ we have

$$\varphi_{w,k}|\mathit{Rest} \sim \mathcal{G}\left(\sum_c z_{c,k}(len(c) - 1) \right. \\ \left. + \sum_c z_{c,k} n_{w,c} + c_w, \sum_{\substack{c: z_c = k \\ w \in \mathbf{w}^c}} |\mathbf{w}|^c + d_w \right). \tag{22}$$

- Concerning $z_{c,k}$, we have

$$\Pr(z_{c,k}|\mathit{Rest}) \propto \Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k) \\ \cdot \Pr(\mathbf{w}^c|\mathbf{\Phi}_k) \cdot \left(\sum_{c' \neq c} z_{c',k} + \alpha_k\right) \tag{23}$$

- Finally, concerning $y^c_{u,v}$, we have

$$\Pr(y^c_{u,v}|z_{c,k}, \mathit{Rest}) \propto (\Delta^c_{u,v})^\rho A_{v,k} S_{u,k} \\ \cdot \left(\sum_{c' \neq c} z_{c',k} y^c_{u,v} + \beta_v\right). \tag{24}$$

Notice that, when sampling $\mathbf{A}$ and $\mathbf{S}$, the parameters of the Gamma distributions are amenable the same optimization discussed in the previous section, for the cases $\rho = 1, 2$. Thus, each sampling step can be computed linearly in the size of the cascades and the number of nodes.

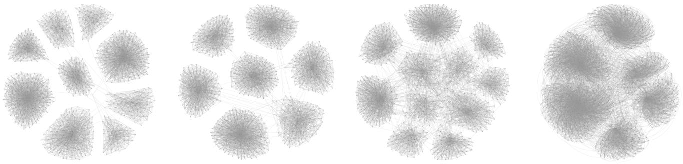
Finally, given a state of the Markov chain, it is possible to devise a procedure for estimating the optimal hyper parameters $\mathcal{H}$, e.g. by resorting to the techniques described in [26], [27].

5 EVALUATION

The following experimental evaluation aims at investigating the following aspects:

1) Determine the conditions upon which the proposed method can correctly detect authoritativeness and susceptibility from propagation logs;
2) Evaluate the proposed models under two different prediction scenarios: (i) given a partially observed cascade, predict which nodes are more likely to become active within a fixed time window and (ii) inferring the underlying propagation network among nodes;
3) Assess the adequacy of the model at fitting real-world data and at identifying topical diffusion patterns.

To perform such analyses we rely on both synthetic and real data, as reported below. The implementation we used in the experiments can be found at http://github.com/gmanco/SurvivalFactorization.



(a) $\mu = 0.001$ (b) $\mu = 0.01$ (c) $\mu = 0.05$ (d) $\mu = 0.1$

Fig. 3: Synthetic networks generated according to different values of μ.

5.1 Synthetic data

The first set of experiments is conducted in a controlled environment. We artificially generate the cascades by hypothesizing a diffusion process and measure the goodness-of-fit of the algorithm to the underlying process. The use of synthetic data allows us to specify a ground truth, i.e., predefined structures that we aim at inferring.

We base the generation on the assumption (studied, e.g., in [32]) that vertices are connected and the diffusion of information happens through the links of the underlying network. Thus, to generate synthesized data, we proceed in three steps. The first step is to generate networks with a known community structure by varying connectivity structure of the network. To this aim, we borrow the synthetic networks studied in [6].

The process of network generation is controlled by five parameters: (i) number of nodes (1k); (ii) average in-degree (10); (iii) maximum in-degree (150); (iv) min/max size of each community (50/750); (v) percentage μ of overlapping memberships $(0.001, 0.01, 0.05, 0.1)$. This last parameter affects the overall connectivity structure of the network, which ranges from well-separated (but still connected) components to strongly overlapping, as shown in fig. 3.

Given a network $G = (V, E)$, we next generate synthetic propagation cascades by simulating a propagation process which spreads over E. The process generates $|\mathcal{I}|$ propagation traces according to the following protocol. The degree of authoritativeness and susceptibility of each node in each community depend on its connectivity pattern. If the node u belongs to community k the values $A_{u,k}$ and $S_{u,k}$ are sampled from lognormal distributions with means $p \cdot \frac{indegree(u)}{\max_v indegree(v)} + (1-p) \cdot rand(0.1, 1)$ and $p \cdot (1 - \frac{outdegree(u)}{\max_v outdegree(v)}) + (1-p) \cdot rand(0.1, 1)$ respectively. For all the remaining communities $h \neq k$, the values for $A_{u,h}$ and $S_{u,h}$ are randomly sampled within a uniform range lower than $A_{u,k}$ $(S_{u,k})$ by an order of magnitude.

The propagation cascades are generated exploiting $\mathbf{A}$ and $\mathbf{S}$: for each cascade to generate, we randomly sample a topic k and a maximal propagation horizon T_{max}. Then, we sample an initial node v with probability proportional to $A_{v,k}$. From this node we start the subsequent diffusion process. Given an active node u and a neighbor v, we sample a hypothetical infection time $t_{u,v}$ using t_v and the rate $A_{u,k} \cdot S_{v,k}$. Node v then becomes active if there exist an influencer u such that $t_{u,v} < T_{max}$.

Finally, for each cascade we generate the content. For each topic k, we generate $\varphi_{w,k}$ randomly and then draw word-frequencies according to the Poisson model and to the topic of the cascade. The size of the content is fixed arbitrarily to 1024 words.

	S1	S2	S3	S4
Communities	9	7	11	6
Activations	215,608	275,633	171,501	313,972
Median activations/cascade	86	139	73	127
Median activations/user	220	276	173	314
Min activations/user	192	250	145	231

TABLE 3: Statistics for the synthesized cascades.

In the following experiments, we set $p = 0.9$, $|\mathcal{I}| = 2,048$ and run the generation of cascades on 4 networks, with different degrees of overlapping. The main properties of the synthesized data are summarized in Table 3.

Predicting activation times. The first experiment is meant to evaluate the accuracy in estimating the activation times. Given a training and test sets $\mathcal{C}_{train}$ and $\mathcal{C}_{test}$ of cascades, we train the model on $\mathcal{C}_{train}$ and measure the accuracy of the predictions on $\mathcal{C}_{test}$.[1] We chronologically split each cascade $c \in \mathcal{C}_{test}$ into c_1 and c_2 (for each $u \in c_1$ and $v \in c_2$, $t_u(c) < t_v(c)$) and pick a random subset c_3 of vertices that did not participate to corresponding cascade. We use c_1 to predict the most likely topic k by exploiting Eq. 12. Then, for each user in $c_2 \cup c_3$ we compute $\delta_u = \min_{v \in c_1} (A_{v,k} S_{u,k})^{-1}$.

We set a 90:10 training/test proportion and vary the chronological split proportion from 50% to 80%. Given a target delay horizon H, the prediction on u is considered as: true positive (TP) if $\delta_u < H$ and $u \in c_2$; true negative (TN) if $\delta_u > H$ and $u \in c_3$; false positive (FP) if $\delta_u < H$ and $u \in c_3$; and false negative (FN) if $\delta_u > H$ and $u \in c_2$. By varying H, we can plot ROC and F curves.

The results of the experiments, reported in Fig. 4, show that the proposed method is effective in predicting activation behavior even when the propagation happens on networks with an overlapping community structure. The best performances are achieved on the network *S3*, despite the fact that some communities are strongly interconnected. A possible explanation is the higher number of communities in the dataset, which also makes cascades shorter and the co-occurrence of nodes less likely in cascades where they are not susceptible/authoritative.

Network reconstruction. The purpose of this experiment is to evaluate whether connectivity patterns between users can be inferred by the susceptibility and influence matrices. We express the likelihood of the existence of the link $u \rightarrow v$ as $\lambda_{u,v} = \sum_k S_{u,k} \cdot A_{v,k}$. When reconstructing the network, we consider all pairs of users which exhibit a connection as positive examples. For the negative examples, we focus on two-hops non-existing links [2].

We compare the proposed algorithm with the standard NetRate benchmark [14] described in section 2. NetRate learns a higher number of parameters ($O(N^2)$, compared to $O(NK)$ for our model). Thus, it is in principle more prone to overfitting. However, since the estimation in NetRate is accomplished through global optimization, it does not

[1] The two sets are obtained by randomly splitting the original dataset by ensuring that there is no overlap among the cascades of the two sets, but there is no vertex in the test that has not been observed in the training.

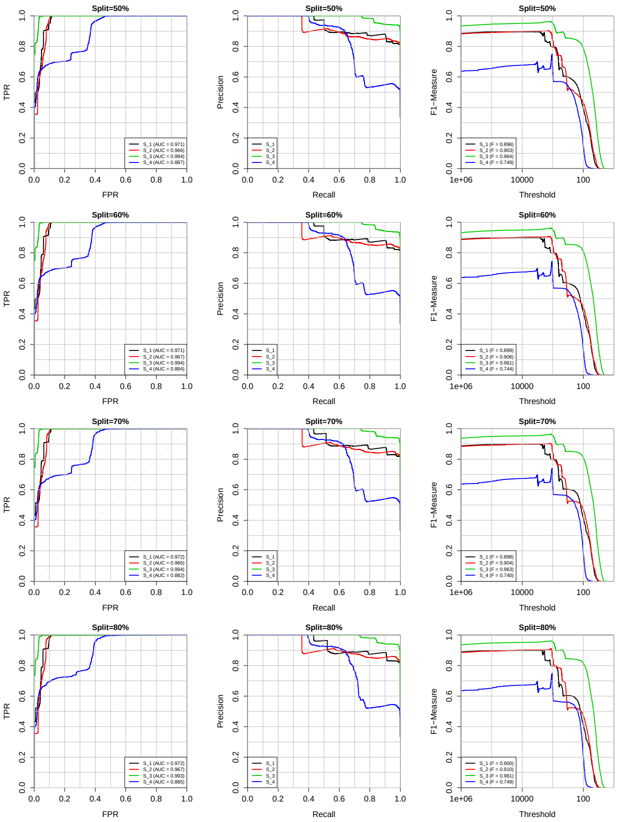


Fig. 4: Evaluation on the Synthetic Datasets. The rows report splits from 50% to 80%. The columns report respectively AUC, Precision/Recall, and F-measure/threshold on activation delay.

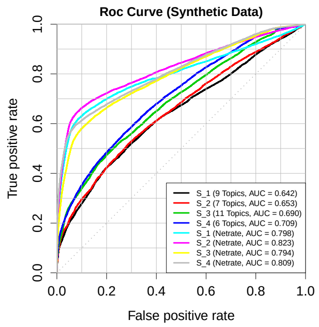


Fig. 5: Network reconstruction on Synthetic Datasets.

explicitly computes, for each activation, the most likely influencer. As a consequence, it does not suffer from a specific bias occurring with our model, where the most likely influencer dominates the likelihood of links. This is visible in fig. 5, where NetRate exhibits a higher accuracy in predicting links. Notwithstanding, our model is still capable of achieving sufficient accuracy in inferring the connectivity.

	S1	S2	S3	S4
F-measure	0.74	0.70	0.80	0.74
ARI	0.61	0.56	0.64	0.58
NMI	0.55	0.43	0.64	0.38
χ_A	0.77	0.87	0.68	0.88
χ_S	0.64	0.49	0.24	0.54
χ_Φ	0.00045	0.0008	0.00097	0.001

TABLE 4: Evaluation of cascade clustering and matrix reconstruction on synthetic data.

Cascade clustering. In this set of experiments, we evaluate the capability of the algorithm at detecting the correct topic for each cascade, based on the hypothesized survival factorization model. Since the synthetic data contains ground truth, it is possible to evaluate the true cascade assignments with the assignments that can be obtained from the final $\gamma_{c,k}$. We measure the quality of the discovered assignments w.r.t. the known ground truth communities using the Adjusted Rand Index [20] (ARI), as well as the F-Measure and the Normalized Mutual Information [1] (NMI). The results are reported in Table 4. The algorithm achieves a relatively good accuracy at separating cascades by topics. Some topics tend to be mixed, probably as a result to the fact that nodes are heavily connected. For example, for *S1* we observe 6 topics with regards to the 9 expected: the original communities (3,5), (4,8) and (7,9) are mixed together.

Inference accuracy. To evaluate the effectiveness of the learning algorithm, we compare the original $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Phi}$ matrices with those inferred. However, this setting presents two issues, namely possible difference in scale and the fact that the column components of these matrices can be shuffled. We approach this problem by introducing a measure which addresses both issues: given a matrix M and its estimation $\tilde{M}$, we define $\chi_M = \sum_k \min_h \|M_{\cdot,k} - \alpha_{k,h}\tilde{M}_{\cdot,h}\|^2/\|M\|^2$, where each column k in the inferred matrix $\tilde{M}$ is associated with the column h in the source matrix M with minimal distance. [2] In this formulation the term $\alpha_{k,h}$ represents the scaling factor relative to columns k and h, and it normalizes the differences to comparable results. The optimal value for $\alpha_{k,h}$ can be obtained by solving $\arg\min_\alpha \|M_{\cdot,k} - \alpha\tilde{M}_{\cdot,h}\|^2$ and thus yielding the following closed formula for χ:

$$\chi_M = \frac{1}{\|M\|^2}\sum_k \min_h \left[\sum_u M_{u,k}^2 - \frac{\left(\sum_u M_{u,k}\tilde{M}_{u,h}\right)^2}{\sum_u \tilde{M}_{u,h}^2}\right]$$

The values of χ for $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Phi}$ are reported in Table 4. The closest the values are to zero, the higher is the correspondence between the matrices. These results show that the estimation is faithful to the original values, with some differences due to the occasional merging of topics.

Scalability. In the last set of experiments, we measure the performance of algorithm. To this purpose, we generate synthetic datasets following the same protocol described above, but where the number of users and the number of cascades differ. In particular, we adopt the parameter $\mu = 0.01$ and range the number of users from $1,000$ to

[2] In principle two or more columns in the source matrix can be associated with the same column h in the inferred matrix. Notice, however, that by construction the source matrices we consider are practically block-diagonal (with a block including a single column and multiple rows). Thus, this situation can only happen if the target column represents two or more topics: That is, when the target matrix exhibits higher values on two or more topics, and there are no other "pure" columns on those topics. We consider this a legit situation that would nevertheless result in a larger distance between the target and the source columns.

N	M	Min/Med/Max N_c
1,000	907	5/64/195
5,000	3,279	5/148/335
10,000	5,696	5/128/341
50,000	57,438	5/132/1782
100,000	113,011	5/120/1949

TABLE 5: Synthetic data for increasing values of N and M.

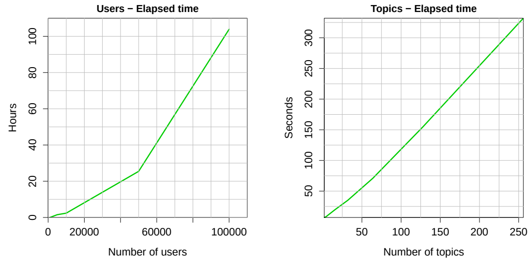


Fig. 6: Scalability analysis on synthetic.

$100,000$, and the corresponding cascade sizes accordingly. Table 5 reports the statistics for each generated datasets. Figure 6, reports the running times for these datasets with $k = 16$, as well as those for the first dataset, with varying number of topics.

5.2 Real data

In this section, we assess the performances of the proposed method on real data, from a quantitative and qualitative perspective. First, we evaluate the accuracy of the model at predicting when a user will retweet a post, and at inferring the underlying network of diffusion (network reconstruction) on information cascades extracted from Twitter. Secondly, we analyze and discuss topical and diffusion patterns inferred on the Memetracker dataset.

5.2.1 Twitter

The following analysis is based on a sample of real-world propagation cascades crawled from the public timeline of *Twitter* and studied in [3]. The propagation of information on Twitter happens by retweet and in this dataset tracks the propagation of URLs over the Twitter network during a period of one month (July 2012). Each activation/adoption corresponds to the instance when a user tweets a certain URL. Note that this dataset does not provide side-information (e.g. hashtags associated to each tweet, or the actual URL being shared). The relevant features of the dataset, called *Twitter-Large*, can be observed in Fig. 7. We also select a subset of the dataset by considering users who participated in at least 15 cascades and retweet cascades that involved at least 5 users. We refer to this dataset as *Twitter-Small*. A summary of the properties of both datasets is shown in Table 6.

	Twitter-Large	Twitter-Small
Nodes	28,585	6,030
Edges	1,636,451	259,568
Activations	516,412	187,941
Cascades	8,541	3,983
Max Delay	2,380,651	2,141,136
Avg Delay	36,775	50,117
Median activations/cascade	18	17
Median activations/user	15	26
Min activations/cascade	1	7
Min activations/user	11	15

TABLE 6: Summary of the Twitter data used for evaluation.

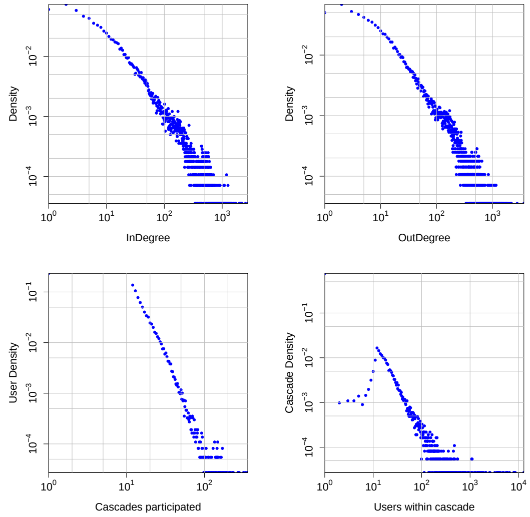


Fig. 7: Distributions within the *Twitter-Large* dataset.

Predicting activation times. We apply the testing protocol detailed in Sec. 5.1 on the Twitter datasets for predicting users retweet times, by varying the training/test chronological splits (from 50% to 80%). Results, reported in Fig. 8 and Fig. 9, show that the model achieves high accuracy in predicting which are the users more likely to become active on each cascade within the prediction window. The prediction accuracy is higher on *Twitter-Small*. This result is compatible with the intuition that the inference works better when the focus is on users who actively participate into cascades. Finally, like in the case of synthesized data, the accuracy is not affected by the size of the cascade used for inferring the optimal topic.

Network reconstruction. We also report the results on network reconstruction for real data. The results of this test are shown in Fig. 10. The best AUC value on *Twitter-small* is 0.7, obtained by by employing 4 topics, and it is 31% improvement over the accuracy achieved by NetRate. While NetRate took roughly 27 days to execute, our method converges in about 6 minutes. Also, the effects of overfitting for NetRate can be clearly seen, as the latter does not replicate the positive results obtained on synthesized data. A large number of users results in a too sparse influence matrix and consequently a poor estimation of the connectivity $u \to v$, which instead can be more accurately reconstructed by resorting to factorization.

On *Twitter-Large* the best AUC achieved is 0.595, by employing 2 topics. The learning algorithm converges in about 89 minutes. Due to time constraints it was not possible to run NetRate over this dataset. The better AUC of *Twitter-Small* over *Twitter-Large* can be explained by the more active participation of users within cascades (see statistics in Table 6 and fig. 7). These represent users which, albeit

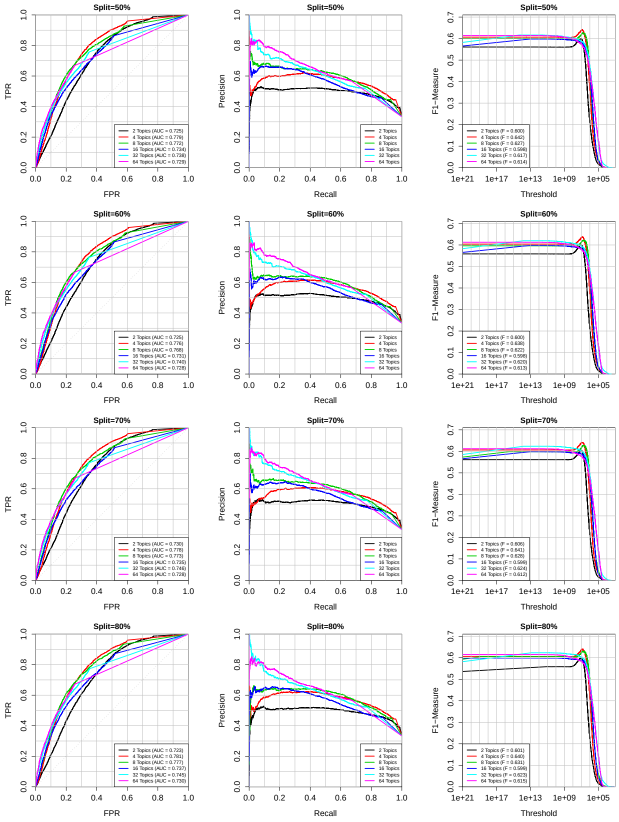


Fig. 8: Activation time prediction on *Twitter-Large*.

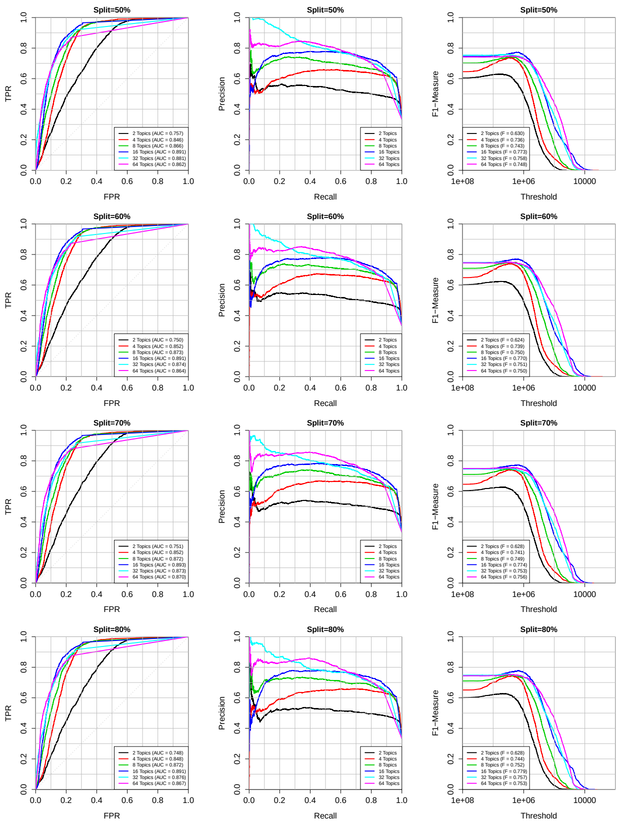


Fig. 9: Activation time prediction on *Twitter-Small*.

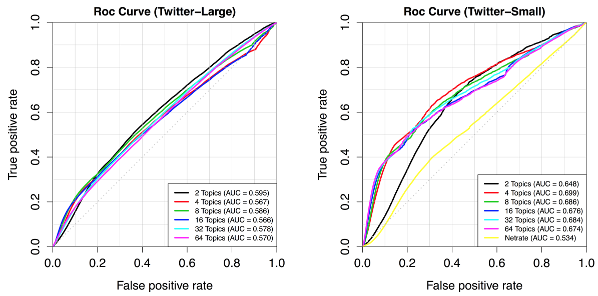


Fig. 10: Evaluation on network reconstruction on *Twitter-Large* (left) and *Twitter-Small* (right).

exhibiting following relationships do not retweet any information. For these nodes the transmission rates are low, and as a consequence the prediction is biased.

Analysis. That said, the learning process is still capable of detecting a clear structure within the propagation process. In fact, a closer look at the matrices produced by the algorithm shows an evident block-diagonal structure. Within figure 11, we cluster the nodes according to the topic k corresponding to the highest value, and then plot the matrices by permuting the rows accordingly. Interestingly, each block exhibits a small number of authoritative users compared to the susceptible ones.

Finally, fig. 12 report the learning times also for the real datasets. We can see that on *Twitter-Large*, learning 64 topics requires about 10 hours.

5.2.2 Memetracker

The evaluation on the Memetracker dataset [23] is aimed at assessing the alignment between the topical and social influence structure. This dataset tracks phrases and quotes over online-news providers and blogs; textual variants of the same phrase are clustered together and the dataset specifies each timestamp at which a particular blog mentioned a phrase belonging to a phrase cluster. We consider each phrase cluster as a separate cascade, the root-phrase as the content being diffused and the hostname extracted from the url of the blog as vertex identifier. In this case, an activation within a information cascade represent the first timestamp at which a given blog mentioned a phrase belonging to the considered phrase cluster. The raw dataset was preprocessed to filter out cascades with less than 10 activations and with less than 10 words as content, and vertices that belong to less than 10 cascades. This step resulted in a dataset that contains 7k vertices and 28k cascades, while the word dictionary contains 3.5k tokens, with an average of 16 words for each cascade.

For the sake of presentation, we run the survival factorization learning algorithm setting $K = 8$. Table 7 reports the most relevant words for each topic, i.e. the words w which exhibit the highest value of $\varphi_{w,k}$ for each k, and our interpretation of the topic is reported in the headings of the table.

Next, we analyze each cascade and compute

- The most-likely topic as $\tilde{k}_c = \arg\max_k \gamma_{c,k}$;

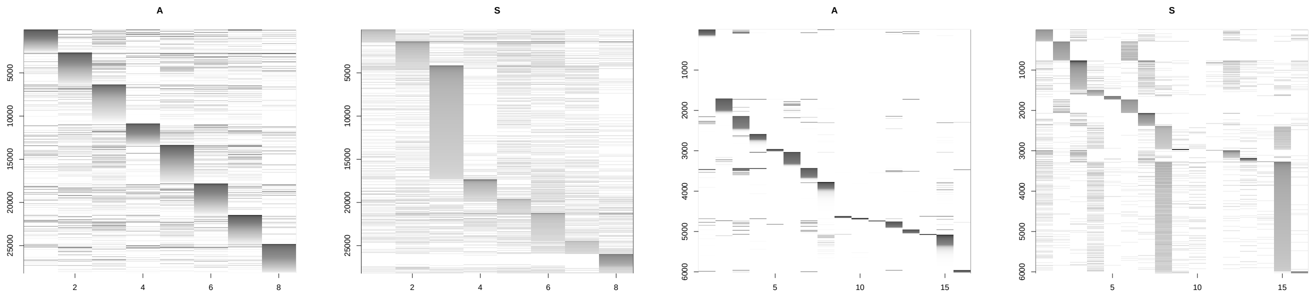


Fig. 11: **A** and **S** matrices on *Twitter-Large* (8 topics) and *Twitter-small* (16 topics).

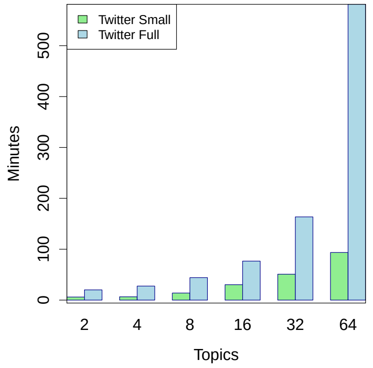


Fig. 12: Execution times for the learning algorithm on Twitter.

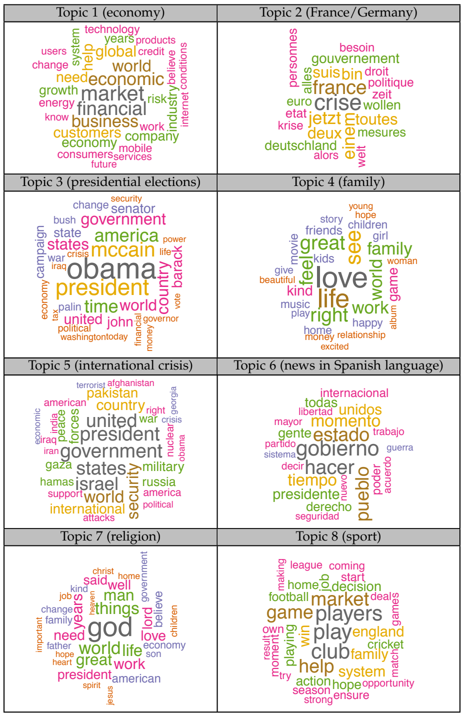

Topic 1 (economy)
technology system years products users global credit help change world need economic growth market risk energy financial know business work customers company economy mobile consumers services future industry believe internet conditions
Topic 2 (France/Germany)
personnes besoin gouvernement alles suis bin droit france politique euro crise wollen etat jetzt zeit krise deux toutes mesures deutschland alors eine welt
Topic 3 (presidential elections)
security change senator bush government state america power campaign states mccain war crisis iraq obama barack country president economy tax palin time world vote united john governor political washingtontoday financial money
Topic 4 (family)
young hope story children friends movie great girl see family give feel kids woman beautiful love kind world game music life play right work happy home money relationship album excited
Topic 5 (international crisis)
terrorist afghanistan pakistan american country right georgia economic india peace forces united war crisis iraq president nuclear iran government obama gaza states military hamas israel russia support world america security international political attacks
Topic 6 (news in Spanish language)
internacional todas libertad unidos mayor momento gente trabajo partido estado sistema gobierno guerra decir hacer tiempo nuevo poder presidente acuerdo derecho pueblo seguridad
Topic 7 (religion)
christ home government said well kind job man things family change years god lord believe need love children important father world life economy hope son heart great work president american spirit jesus
Topic 8 (sport)
league coming making start home club decision football market deals game players games own play result england win playing club moment cricket family help system try action hope opportunity season ensure strong match

TABLE 7: Most relevant terms for each topic on Memetracker.

- The most-likely cascade tree for each cascade $\tilde{T}_c$ by computing the parent of each active node (excluding the root) as $par(u)_c = \arg\max_v \eta^{\tilde{k}_c}_{c,u,v}$;

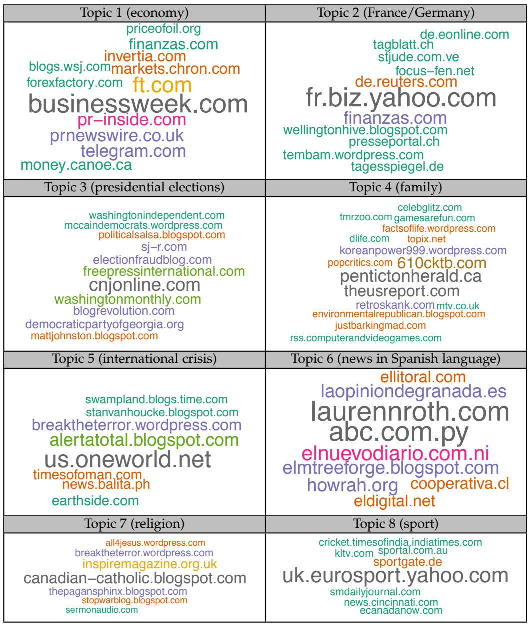

Topic 1 (economy)
priceofoil.org
finanzas.com
invertia.com
blogs.wsj.com
markets.chron.com
forexfactory.com
ft.com
businessweek.com
pr–inside.com
prnewswire.co.uk
telegram.com
money.canoe.ca
Topic 2 (France/Germany)
de.eonline.com
tagblatt.ch
stjude.com.ve
focus–fen.net
de.reuters.com
fr.biz.yahoo.com
finanzas.com
wellingtonhive.blogspot.com
presseportal.ch
tembam.wordpress.com
tagesspiegel.de
Topic 3 (presidential elections)
washingtonindependent.com
mccaindemocrats.wordpress.com
politicalsalsa.blogspot.com
sj–r.com
electionfraudblog.com
freepressinternational.com
cnjonline.com
washingtonmonthly.com
blogrevolution.com
democraticpartyofgeorgia.org
mattjohnston.blogspot.com
Topic 4 (family)
celebglitz.com
tmrzoo.com
gamesarefun.com
factsoflife.wordpress.com
dlife.com
topix.net
koreanpower999.wordpress.com
popcritics.com
610cktb.com
pentictonherald.ca
theusreport.com
retroskank.com
mtv.co.uk
environmentalrepublican.blogspot.com
justbarkingmad.com
rss.computerandvideogames.com
Topic 5 (international crisis)
swampland.blogs.time.com
stanvanhoucke.blogspot.com
breaktheterror.wordpress.com
alertatotal.blogspot.com
us.oneworld.net
timesofoman.com
news.balita.ph
earthside.com
Topic 6 (news in Spanish language)
ellitoral.com
laopiniondegranada.es
laurennroth.com
abc.com.py
elnuevodiario.com.ni
elmtreeforge.blogspot.com
howrah.org
cooperativa.cl
eldigital.net
Topic 7 (religion)
all4jesus.wordpress.com
breaktheterror.wordpress.com
inspiremagazine.org.uk
canadian–catholic.blogspot.com
thepagansphinx.blogspot.com
stopwarblog.blogspot.com
sermonaudio.com
Topic 8 (sport)
cricket.timesofindia.indiatimes.com
kltv.com
sportal.com.au
sportgate.de
uk.eurosport.yahoo.com
smdailyjournal.com
news.cincinnati.com
ecanadanow.com

TABLE 8: Most influential hosts for each topic on Memetracker.

- For each cascade c the delay $\Delta^c_{u,v}$ for each pair u, v such that $par(u)_c = v$, and compute the average delay over cascades in each topic;
- The Wiener index for each cascade tree, and use this information to compute the average Wiener index for a topic k as $\bar{w}_k = avg_{c:\ \tilde{k}_c=k}\ W(\tilde{T}_c)$;
- The depth of each cascade tree, which is averaged across cascades in the same topic to compute the average cascade topical depth.

The outcome of this analysis is summarized in Table 9. The topic labeled as "sports" exhibits the shortest average transmission delay, followed by "international crisis" and "news in Spanish language". In general, cascade trees are shallow, which suggests that the propagation of information is due to few influencers. The highest average Wiener index is observed on the topic "religion".

Finally, Table 8 shows the top influencers for each topic, computed by counting the number of children of each node in each cascade and aggregating this info at the topic level. The top influential blogs are well aligned with the topical structure shown in Table 7.

Topic	Average delay	Avg Wiener index	Avg depth
1	20.6h	1.73	1.77
2	22.8h	1.69	1.74
3	43.5h	1.82	1.93
4	21h	1.76	1.83
5	12.7h	1.80	1.92
6	12h	1.85	2.13
7	23.8h	1.89	2.20
8	7.8h	1.83	2.08

TABLE 9: Characterization of the cascade trees for each topic.

6 Conclusion

In this work we proposed a model for information diffusion where adoptions can be explained in terms of susceptibility and authoritativeness. The latter concepts can be expressed as latent factors over a low-dimensional space representing topical interests. We showed the adequacy of the resulting probabilistic model both from a mathematical and an experimental point of view.

The experiments show that the number of topics actually affects the fitness of the model. However, we did not specifically cover the problem of detecting the optimal K. As a matter of fact, this is an orthogonal issue which is extensively covered in the literature and viable solutions exist. For example, within the context of the EM algorithm that we adopt here, a suitable solution consists in choosing a reasonably high value for K and then allow an annihilation procedure by choosing an appropriate prior for Θ, as described in [12]. Besides the capability of automatically detecting the optimal number of topics, this would have the further advantage of robustness to random initialization: By starting with an arbitrarily large number of components, it would avoid the pitfalls of local maxima, since the whole parameter space would be likely covered. Even the Gibbs sampling scheme for Bayesian formulation can be easily extended to a non-parametric setting by exploiting Dirichlet Process Mixtures [13].

It is worth comparing the Survival Factorization approach proposed here with the CWN approach [6] discussed in section 2. CWN tackles the problem of detecting social communities when the social graph is not available, but only cascades are available. Within CWN, each user has a likelihood of belonging to a community, which can explain its activations according to an influential user within the community. Under this perspective, both CWN and Survival Factorization approach the problem of social contagion with similar tools but from different perspectives. The former focuses on the user by modeling its susceptibility implicitly as community membership. As such, it is not interested in detecting whether each activation in a cascade refers to a same topic. By contrast, the latter latter focuses specifically on this aspect and measures each activation as the result of both susceptibility and influence. It would be interesting to see how these two approaches can be combined to model both community membership and cascade topic.

Finally, within the model, we chose not to correlate the content with the time delay in the propagation process. As a matter of fact, the content can actually influence the diffusion: Think e.g. of an hashtag relative to a earthquake. Clearly, combining susceptibility, authoritativeness and strength of the content can help better characterize the diffusion process and this should be taken into account in the inference procedure.

References

[1] L.N.F. Ana and A.K. Jain. Robust data clustering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR'03)*, 2003.

[2] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. In *WSDM*, pages 635–644, 2011.

[3] N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In *WSDM*, pages 33–42, 2013.

[4] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. *Knowl. Inf. Syst.*, 37(3):555–584, 2013.

[5] N. Barbieri, F. Bonchi, and G. Manco. Who to follow and why: Link prediction with explanations. In *KDD*, pages 1266–1275, 2014.

[6] N. Barbieri, F. Bonchi, and G. Manco. Efficient methods for influence-based network-oblivious community detection. *ACM Trans. Intell. Syst. Technol.*, 8(2):32:1–32:31, 2016.

[7] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 2006.

[8] S. Bourigault et al. Learning social network embeddings for predicting information diffusion. In *WSDM*, pages 393–402, 2014.

[9] N. Du, M. Farajtabar, A. Ahmed, A. J. Smola, and L. Song. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *KDD*, pages 219–228, 2015.

[10] N. Du, L. Song, H Woo, and H. Zha. Uncover topic-sensitive information diffusion networks. In *AISTATS*, pages 229–237, 2013.

[11] N. Du, L. Song, Ming Y., and Alex J. S. Learning networks of heterogeneous influence. In *NIPS*, pages 2780–2788. 2012.

[12] M. A. T. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(3):381–396, 2002.

[13] S. J. Gershman and D. M. Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

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

[15] Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *KDD*, 2010.

[16] A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *WSDM*, 2010.

[17] X. He, T. Rekatsinas, J. R. Foulds, L. Getoor, and Y. Liu. Hawkestopic: A joint model for network inference and topic modeling from text-based cascades. In *ICML*, pages 871–880, 2015.

[18] Q. Hu, S. Xie, S. Lin, W. Fan, and P.S. Yu. Frameworks to encode user preferences for inferring topic-sensitive information networks. In *SDM*, pages 442–450, 2015.

[19] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. In *KDD*, pages 266–274, 2013.

[20] Anil K. Jain and Richard C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[21] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[22] E. T Lee and J. Wang. *Statistical methods for survival data analysis*. Wiley-Interscience, 2003.

[23] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.

[24] S. Lin, Q. Hu, J Zhang, and P. Yu. Discovering audience groups and group-specific influencers. In *ECMLPKDD*, pages 559–575, 2015.

[25] Qiaozhu Mei, Hui Fang, and ChengXiang Zhai. A study of poisson query generation model for information retrieval. In *SIGIR*, pages 319–326, 2007.

[26] T. Minka. Estimating a gamma distribution. Available at http://research.microsoft.com/en-us/um/people/minka/papers/minka-gamma.pdf, 2002.

[27] T. Minka. Estimating a dirichlet distribution. Available at http://www.msr-waypoint.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf, 2012.

[28] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD*, pages 807–816, 2009.

[29] Liaoruo Wang, Stefano Ermon, and John E Hopcroft. Feature-enhanced probabilistic models for diffusion network inference. In *ECMLPKDD*, pages 499–514, 2012.

[30] S. Wang, X. Hu, P.S. Yu, and Z. Li. MMRate: Inferring multi-aspect diffusion networks with multi-pattern cascades. In *KDD*, pages 1246–1255, 2014.

[31] Y. Wang, H. Shen, S. Liu, and X. Cheng. Learning user-specific latent influence and susceptibility from information cascades. In *AAAI*, 2015.

[32] L. Weng, F. Menczer, and Y. Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3, 2013.

[33] R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW*, pages 981–990, 2010.

[34] S. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. In *ICML*, pages 1–9, 2013.

Acknowledgments

This work was partially supported by the 2014 YAHOO Faculty Research and Engagement Program.

Giuseppe Manco Giuseppe Manco graduated summa cum laude in computer science and received the PhD degree in computer science from the University of Pisa. He is currently a senior researcher at the Institute of High Performance Computing and Networks (ICAR-CNR) of the National Research Council of Italy and a contract professor at University of Calabria, Italy. His current research interests include knowledge discovery and data mining, Web databases, and semi-structured data, Recommender Systems and Social Network Analysis. He has been the coordinator of several national and international research projects. He has been serving in the program committee of several international/national conferences, including: IEEE ICDM, ECMLPKDD, SIAM SDM, PAKDD, and he was program co-chair of ECML-PKDD 2016. He is serving as an associate editor for the Journal of Intelligent Information Systems, Knowledge and Information Systems and Machine Learning Journal.

Ettore Ritacco Ettore Ritacco is a researcher at Institute for High Performance Computing and Networking (ICAR) of the Italian National Research Council (CNR). He received his M.S. and Ph.D. degree in computer science in 2006 and 2011, respectively, at the University of Calabria (UNICAL). He is expert on data science, data analytics and enabling technologies for data analytics. Currently, his research activities are focusing on User Profiling and Behavioral Modeling, Social Network Analysis, Recommendation, Information Propagation and Diffusion, Profiling for Cyber Security, Latent Factor and Deep Learning models, Time Series analysis, Outlier Detection and Hardware Component Failure Prevention. He is interested in new frontiers of Computer Science and Technology aimed at analyzing Complex Big Data.

Nicola Barbieri Nicola Barbieri, Ph.D, has worked as a Staff Data Scientist on the Market-Places Personalization team at Schibsted Media Group and as a Sr. Data Scientist at Tumblr (New York City), working on Search & Discovery. Prior to that he held a Research Scientist position at Yahoo Labs (Barcelona/London) working on Ad Quality and Retrieval. He has more than 5 years experience in developing large scale search and recommendation techniques, with important contributions to sponsored search relevance at Yahoo and to blog recommendation at Tumblr. His research focuses on the development of novel data mining and machine learning techniques for information diffusion, social influence analysis, viral marketing and community detection. He is co-author of the book ”Probabilistic Approaches to Recommendations” published by Chapman & Hall/CRC Press.

Appendix A
Derivation of the E and M steps

A.1 Recap: The EM algorithm

The log-likelihood of the observed cascades can be written as

$$\begin{aligned}\log \Pr(\mathcal{C},\boldsymbol{\Xi}) &= \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z},\mathbf{Y}) \log \Pr(\mathcal{C},\Xi)\\ &= \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z},\mathbf{Y}) \left\{ \log \frac{\Pr(\mathcal{C},\mathbf{Z},\mathbf{Y},\Xi)}{\Pr(\mathbf{Z},\mathbf{Y}|\mathcal{C},\Xi)} \right\}\\ &= \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z},\mathbf{Y}) \left\{ \log \frac{\Pr(\mathcal{C},\mathbf{Z},\mathbf{Y},\Xi)}{q(\mathbf{Z},\mathbf{Y})} \right\}\\ &\quad - \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z},\mathbf{Y}) \left\{ \log \frac{\Pr(\mathbf{Z},\mathbf{Y}|\mathcal{C},\Xi)}{q(\mathbf{Z},\mathbf{Y})} \right\}\\ &= \mathcal{L}(q;\mathcal{C},\boldsymbol{\Xi}) + \mathbf{KL}(q\| \Pr(\mathbf{Z},\mathbf{Y}|\mathcal{C},\Xi))\end{aligned}$$

where q is an arbitrary instrumental distribution over the latent variables $\mathbf{Z},\mathbf{Y}$ and

$$\begin{aligned}\mathcal{L}(q;\mathcal{C},\boldsymbol{\Xi}) &= \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z},\mathbf{Y}) \log \Pr(\mathcal{C},\mathbf{Z},\mathbf{Y},\Xi)\\ &\quad - \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z},\mathbf{Y}) \log q(\mathbf{Z},\mathbf{Y})\\ &= \mathbb{E}_q[\log \Pr(\mathcal{C},\mathbf{Z},\mathbf{Y},\Xi)] + \mathbb{H}[q]\end{aligned}$$

In practice, since the Kullback-Leibler divergence $\mathbf{KL}(q\|p) \geq 0$ we have that $\log \Pr(\mathcal{C}|\boldsymbol{\Xi}) \geq \mathcal{L}(q;\mathcal{C},\boldsymbol{\Xi})$ The bound is tight for the exact posterior, due to the fact that $\mathbf{KL}(p\|p) = 0$. That is,

$$\arg\max_{q(\mathbf{Z},\mathbf{Y})} \mathcal{L}(q;\mathcal{C},\boldsymbol{\Xi}) = \Pr(\mathbf{Z},\mathbf{Y}|\mathcal{C},\boldsymbol{\Xi})$$

and for that specific value we have $\log \Pr(\mathcal{C}|\boldsymbol{\Xi}) = \mathcal{L}(q;\mathcal{C},\boldsymbol{\Xi})$. This suggests a simple iterative procedure for estimating the optimal parameters in the usual EM setting, by computing the variational approximations given by the E and M steps:

E step: $q^{(n)}(\mathbf{Z},\mathbf{Y}) = \Pr(\mathbf{Z},\mathbf{Y}|\mathcal{C},\boldsymbol{\Xi}^{(n-1)})$
M step: $\boldsymbol{\Xi}^{(n)} = \arg\max_{\boldsymbol{\Xi}} \sum_{\mathbf{Z},\mathbf{Y}} q^{(n)}(\mathbf{Z},\mathbf{Y}) \log \Pr(\mathcal{C},\mathbf{Z},\mathbf{Y},\Xi)$

It is easy to see that iterating between the above steps produces a progressive improvement of the estimation of the optimal parameters:

$$\begin{aligned}&\log \Pr(\mathcal{C},\boldsymbol{\Xi}^{(n)})\\ &\quad = \mathcal{L}(q^{(n+1)};\mathcal{C},\boldsymbol{\Xi}^{(n)}) && \text{(by definition)}\\ &\quad \leq \mathcal{L}(q^{(n+1)};\mathcal{C},\boldsymbol{\Xi}^{(n+1)}) && \text{(M step)}\\ &\quad \leq \mathcal{L}(q^{(n+2)};\mathcal{C},\boldsymbol{\Xi}^{(n+1)}) && \text{(E step)}\\ &\quad = \log \Pr(\mathcal{C},\boldsymbol{\Xi}^{(n+1)}) && \text{(by definition)}\end{aligned}$$

In the following we omit the $^{(n)}$ superscript when the components to be updated are clear from the context.

A.2 Expected Likelihood

We can rewrite the above expressions by adopting an alternative notation. In the following, With an abuse of notation, for a generic binary variable x the event $x = 1$ within a probability function is simply denoted as x, e.g. $p(x)$ denotes $p(x = 1)$. Recall that $\Pr(\mathbf{t}^c|\mathbf{Z},\mathbf{Y},\mathbf{A},\mathbf{S})$ is defined as in eq. 8 and $\Pr(\mathbf{w}^c|\boldsymbol{\Phi}_k)$ as in eq. 9. We also assume multinomial distributions over $\mathbf{Z}$ and $\mathbf{Y}$. However, we assume that the in the initial setting, every user is equally a viable influencer. That is to say, the prior relative to $\mathbf{Y}$ is uniform. Concerning the prior $\Theta = \{\pi_1,\ldots,\pi_K\}$ on $\mathbf{Z}$, we have

$$\Pr(\mathbf{Z}|\Theta) = \prod_{c=1}^{M}\prod_{k} \Pr(z_{c,k}|\Theta) = \prod_{c=1}^{M}\prod_{k} \pi_k^{z_{c,k}} \tag{25}$$

Finally, we hypothesize no priors on $\mathbf{A},\mathbf{S}$ and Φ for the moment. We shall analyze in details the effects of these priors in a further section in the following.

Define

$$\mathcal{Q}(\Xi;q,\mathcal{C}) = \sum_{\mathbf{Z},\mathbf{Y}} q(\mathbf{Z},\mathbf{Y}) \log \Pr(\mathcal{C},\mathbf{Z},\mathbf{Y},\Xi)$$

We can observe that

$$\begin{aligned}\mathcal{Q}(\Xi;q,\mathcal{C}) &= \sum_{c=1}^{M}\sum_{k} q(z_{c,k}) \log \pi_k\\ &+ \sum_{c=1}^{M}\sum_{u\in\mathcal{V}^+(c)}\sum_{v\prec_c u}\sum_{k} q(z_{c,k},y^c_{u,v}) \log h(\Delta^c_{u,v};\lambda_{v,u,k},\rho)\\ &+ \sum_{c=1}^{M}\sum_{u\in\mathcal{V}^+(c)}\sum_{v\prec_c u}\sum_{k} q(z_{c,k}) \log S(\Delta^c_{u,v}\lambda_{v,u,k},\rho)\\ &+ \sum_{c=1}^{M}\sum_{u\in\mathcal{V}^-(c)}\sum_{v\in\mathcal{V}^+(c)}\sum_{k} q(z_{c,k}) \log S(T^c - t_v(c);\lambda_{v,u,k},\rho)\\ &+ \sum_{c=1}^{M}\sum_{w}\sum_{k} q(z_{c,k}) \left\{ n_{w,c}\log\varphi_{w,k} - |\mathbf{w}^c|\cdot\varphi_{w,k} \right\}\end{aligned} \tag{26}$$

By exploiting definitions 5 and 6, we have:

$$\begin{aligned}\mathcal{Q}(\Xi;q,\mathcal{C}) &= \sum_{c=1}^{M}\sum_{k} q(z_{c,k}) \log \pi_k + (\log\rho)\sum_{c=1}^{M} N_c\\ &+(\rho-1)\sum_{c=1}^{M}\sum_{u\in\mathcal{V}^+(c)}\sum_{v\prec_c u}\sum_{k} q(z_{c,k},y^c_{u,v})\cdot\log\Delta^c_{u,v}\\ &+\sum_{c=1}^{M}\sum_{u\in\mathcal{V}^+(c)}\sum_{v\prec_c u}\sum_{k} q(z_{c,k},y^c_{u,v})\cdot\log\lambda_{v,u,k}\\ &-\sum_{c=1}^{M}\sum_{u\in\mathcal{V}^+(c)}\sum_{v\prec_c u}\sum_{k} q(z_{c,k})\cdot\lambda_{v,u,k}\cdot(\Delta^c_{u,v})^\rho\\ &-\sum_{c=1}^{M}\sum_{u\in\mathcal{V}^-(c)}\sum_{v\in\mathcal{V}^+(c)}\sum_{k} q(z_{c,k})\lambda_{v,u,k}\,(T^c - t_v(c))^\rho\\ &+\sum_{c=1}^{M}\sum_{w}\sum_{k} q(z_{c,k}) \left\{ n_{w,c}\log\varphi_{w,k} - |\mathbf{w}^c|\cdot\varphi_{w,k} \right\}\end{aligned} \tag{27}$$

A.3 E step

As shown to eq. 26, $q(\mathbf{X}, \mathbf{Y})$ (and consequently $\Pr(\mathbf{Z}, \mathbf{Y}|\mathcal{C}, \Xi)$) can be decomposed into the specific activations. In particular, for each cascade c and pair of users u, v with $v \prec_c u$, the E step consists in estimating the following:

$$\begin{aligned} q(z_{c,k}) &= \Pr(z_{c,k}|\mathbf{t}^c, \mathbf{w}^c, \Xi^{(n-1)}) \\ q(z_{c,k}, y^c_{u,v}) &= \Pr(z_{c,k}, y^c_{u,v}|\mathbf{t}^c, \mathbf{w}^c, \Xi^{(n-1)}) \\ &= \Pr(z_{c,k}|\mathbf{t}^c, \mathbf{w}^c, \Xi^{(n-1)}) \cdot \Pr(y^c_{u,v}|z_{c,k}, \mathbf{t}^c, \Xi^{(n-1)}) \end{aligned}$$

Hence, we need to estimate the quantities $\Pr(z_{c,k}|\mathbf{t}^c, \mathbf{w}^c, \Xi)$ and $\Pr(y^c_{u,v}|z_{c,k}, \mathbf{t}^c, \Xi)$. For the former, we have:

$$\Pr(z_{c,k}|\mathbf{t}^c, \mathbf{w}^c, \Xi) = \frac{\Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k) \cdot \Pr(\mathbf{w}^c|\Phi_k) \cdot \Pr(z_{c,k}|\Theta)}{\sum_{\tilde{k}} \Pr\left(\mathbf{t}^c|\mathbf{A}_{\tilde{k}}, \mathbf{S}_{\tilde{k}}\right) \cdot \Pr\left(\mathbf{w}^c|\Phi_{\tilde{k}}\right) \cdot \Pr\left(z_{c,\tilde{k}}|\Theta\right)}$$

In the above formula we have $\Pr(z_{c,k}|\Theta) = \pi_k$, $\Pr\left(\mathbf{w}^c|\Phi_k\right)$ defined as in eq. 9 and $\Pr\left(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k\right)$ obtained by marginalizing $\Pr(\mathbf{t}^c|\mathbf{Y}, \mathbf{A}_k, \mathbf{S}_k)$ in eq. 8 with respect to $\mathbf{Y}$:

$$\begin{aligned} \Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k) &= \sum_{\mathbf{Y}} \Pr(\mathbf{t}^c|\mathbf{Y}, \mathbf{A}_k, \mathbf{S}_k) \\ &= \prod_{u \in \mathcal{V}^-(c)} \prod_{v \in \mathcal{V}^+(c)} S(T^c - t_v(c); \lambda_{v,u,k}, \rho) \\ &\cdot \prod_{u \in \mathcal{V}^+(c)} \prod_{v \prec_c u} S(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho) \sum_{v \prec_c u} h(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho) \end{aligned} \tag{28}$$

Concerning $\Pr(y^c_{u,v}|z_{c,k}, \mathbf{t}^c, \Xi)$, we can observe the following:

$$\begin{aligned} \Pr(y^c_{u,v}|z_{c,k}, \mathbf{t}^c, \Xi) &= \frac{\Pr(\mathbf{t}^c|y^c_{u,v}, \mathbf{A}_k, \mathbf{S}_k)}{\sum_{\tilde{v}} \Pr(\mathbf{t}^c|y^c_{u,\tilde{v}}, \mathbf{A}_k, \mathbf{S}_k)} \\ &= \frac{h(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho)}{\sum_{v' \prec_c u} h(\Delta^c_{u,v'}; \lambda_{v',u,k}, \rho)} \\ &\cdot \frac{\prod_{v' \prec_c u} S(\Delta^c_{u,v'} \lambda_{v',u,k}, \rho)}{\prod_{v' \prec_c u} S(\Delta^c_{u,v'} \lambda_{v',u,k}, \rho)} \\ &\cdot \frac{\prod_{\substack{w \in \mathcal{V}^+(c) \\ w \neq u}} \sum_{v \prec_c w} h(\Delta^c_{w,v}; \lambda_{v,w,k}, \rho)}{\prod_{\substack{w \in \mathcal{V}^+(c) \\ w \neq u}} \sum_{v \prec_c w} h(\Delta^c_{w,v}; \lambda_{v,w,k}, \rho)} \\ &\cdot \frac{\prod_{v' \prec_c w} S(\Delta^c_{w,v'} \lambda_{v',w,k}, \rho)}{\prod_{v' \prec_c w} S(\Delta^c_{w,v'} \lambda_{v',w,k}, \rho)} \\ &= \frac{h(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho)}{\sum_{v' \prec_c u} h(\Delta^c_{u,v'}; \lambda_{v',u,k}, \rho)} \end{aligned}$$

A.4 M Step

Let us denote $q(z_{c,k})$ by $\gamma_{c,k}$ and $\Pr(y^c_{u,v}|z_{c,k}, \mathbf{t}^c, \Xi)$ by $\eta^k_{c,u,v}$. By exploiting the definition $\lambda_{u,v,k} = A_{v,k} \cdot S_{u,k}$ we can hence revrite the likelihood as

$$\begin{aligned} \mathcal{Q}(\Xi; \boldsymbol{\gamma}, \boldsymbol{\eta}, \mathcal{C}) &= \sum_{c=1}^{M} \sum_{k} \gamma_{c,k} \log \pi_k + (\log \rho) \sum_{c=1}^{M} N_c \\ &+ (\rho - 1) \sum_{c=1}^{M} \sum_{u \in \mathcal{V}^+(c)} \sum_{k} \gamma_{c,k} \sum_{v \prec_c u} \eta^k_{c,u,v} \cdot \log \Delta^c_{u,v} \\ &+ \sum_{c=1}^{M} \sum_{u \in \mathcal{V}^+(c)} \sum_{k} \gamma_{c,k} \sum_{v \prec_c u} \eta^k_{c,u,v} \cdot (\log A_{v,k} + \log S_{u,k}) \\ &- \sum_{c=1}^{M} \sum_{u \in \mathcal{V}^+(c)} \sum_{v \prec_c u} \sum_{k} \gamma_{c,k} \cdot A_{v,k} \cdot S_{u,k} \cdot (\Delta^c_{u,v})^\rho \\ &- \sum_{c=1}^{M} \sum_{k} \gamma_{c,k} \sum_{u \in \mathcal{V}^-(c)} \sum_{v \in \mathcal{V}^+(c)} A_{v,k} \cdot S_{u,k} \, (T^c - t_v(c))^\rho \\ &+ \sum_{c=1}^{M} \sum_{w} \sum_{k} \gamma_{c,k} \left\{ n_{w,c} \log \varphi_{w,k} - |\mathbf{w}^c| \cdot \varphi_{w,k} \right\} \end{aligned} \tag{29}$$

Optimizing $\mathcal{Q}(\Xi; \boldsymbol{\gamma}, \boldsymbol{\eta}, \mathcal{C})$ for π_k under the constraint $\sum_k \pi_k = 1$ yields the following:

$$\pi_k = \frac{1}{M} \sum_{c=1}^{M} \gamma_{c,k} \tag{30}$$

Concerning $\mathbf{S}$ and A, the we need to resort to block coordinate ascent optimization. In particular, given $\mathbf{A}$, we can optimize with respect to $\mathbf{S}$. First of all, we can observe that, for a given u and k, it holds that $\sum_{c: u \in \mathcal{V}^+(c)} \sum_{v \prec_c u} \eta_{c;u,v} \cdot \gamma_{c,k} = \sum_{c: u \in \mathcal{V}^+(c)} \gamma_{c,k}$ by definition of $\eta^k_{c,u,v}$. Consequently, we have:

$$S_{u,k} = \frac{\sum_{c: u \in \mathcal{V}^+(c)} \gamma_{c,k}}{\bar{S}_{u,k}} \tag{31}$$

where:

$$\begin{aligned} \bar{S}_{u,k} = & \sum_{c: u \in \mathcal{V}^+(c)} \sum_{v \prec_c u} \gamma_{c,k} \cdot A_{v,k} \cdot (\Delta^c_{u,v})^\rho \\ & + \sum_{c: u \in \mathcal{V}^-(c)} \sum_{v \in \mathcal{V}^+(c)} \gamma_{c,k} \cdot A_{v,k} \, (T^c - t_v(c))^\rho \end{aligned} \tag{32}$$

Also, given $\mathbf{S}$, we can optimize for $\mathbf{A}$ to obtain

$$A_{v,k} = \frac{\sum_{c: v \in \mathcal{V}^+(c)} \gamma_{c,k} \sum_{\substack{u \in \mathcal{V}^+(c) \\ v \prec_c u}} \eta^k_{c,u,v}}{\sum_{c: v \in \mathcal{V}^+(c)} \gamma_{c,k} \bar{A}_{v,k,c}} \tag{33}$$

where:

$$\bar{A}_{v,k,c} = \sum_{u: v \prec_c u} S_{u,k} \cdot (\Delta^c_{u,v})^\rho + \sum_{u \in \mathcal{V}^-(c)} S_{u,k} \, (T^c - t_v(c))^\rho \tag{34}$$

Finally, optimizing with respect to $\mathbf{\Phi}$ yields:

$$\varphi_{w,k} = \frac{\sum_{c=1}^{M} \gamma_{c,k} n_{w,c}}{\sum_{c=1}^{M} \gamma_{c,k} |\mathbf{w}^c|} \tag{35}$$

A.5 Priors

The model can be further refined by assuming some priors on the parameters $\mathbf{A}, \mathbf{S}$ and Φ. Concerning Θ, we assume that each user is equally likely to be influenced and/or

Term	*Definition*	*Term*	*Definition*
$A_{c,u,k}$	$\sum_{v \prec_c u} A_{v,k}$	$S_{c,u,k}$	$\sum_{v \preceq_c u} S_{v,k}$
$\tilde{A}_{c,u,k}$	$\sum_{v \prec_c u} t_v(c) A_{v,k}$	$\tilde{S}_{c,u,k}$	$\sum_{v \preceq_c u} t_v(c) S_{v,k}$
$\hat{A}_{c,u,k}$	$\sum_{v \prec_c u} t_v(c)^2 A_{v,k}$	$\hat{S}_{c,u,k}$	$\sum_{v \preceq_c u} t_v(c)^2 S_{v,k}$
$A_{c,k}$	$\sum_{v \in \mathcal{V}+(c)} A_{v,k}$	$S_{c,k}$	$\sum_{v \in \mathcal{V}+(c)} S_{v,k}$
$\tilde{A}_{c,k}$	$\sum_{v \in \mathcal{V}+(c)} t_v(c) A_{v,k}$	$\tilde{S}_{c,k}$	$\sum_{v \in \mathcal{V}+(c)} t_v(c) S_{v,k}$
$\hat{A}_{c,k}$	$\sum_{v \in \mathcal{V}+(c)} t_v(c)^2 A_{v,k}$	$\hat{S}_{c,k}$	$\sum_{v \in \mathcal{V}+(c)} t_v(c)^2 S_{v,k}$
$R_{c,u,k}$	$\sum_{\substack{v \in \mathcal{V}+(c) \\ u \prec_c v}} (A_{c,v,k})^{-1}$	S_k	$\sum_v S_{v,k}$
$\hat{R}_{c,v,k}$	$\sum_{\substack{u \in \mathcal{V}+(c) \\ v \prec_c u}} \left(t_u(c) A_{c,u,k} - \tilde{A}_{c.u,k}\right)^{-1}$	$L_{c,k}$	$\sum_{v \in \mathcal{V}+(c)} \log S_{v,k}$
$\tilde{R}_{c,v,k}$	$\sum_{\substack{u \in \mathcal{V}+(c) \\ v \prec_c u}} t_u(c) \left(t_u(c) A_{c,u,k} - \tilde{A}_{c.u,k}\right)^{-1}$		

TABLE 10: Counters on the cascades.

influence. This is modeled by assuming an exponential prior with rate N (or, equivalently, a Gamma prior with parameters 1 and N):

$$\begin{aligned} \Pr(A_{v,k}) &= \exp\left\{-N \cdot A_{v,k} + \log N\right\} \\ \Pr(S_{u,k}) &= \exp\left\{-N \cdot A_{u,k} + \log N\right\} \end{aligned} \tag{36}$$

Concerning Φ, we want that the expected frequency of each term to be 0, unless different evidence. This means that $\sum_k \varphi_{w,k}$ should be less than one. We ensure this condition by adopting a gamma prior with shape 2 and rate $2K^2 + 2$:

$$\Pr(\varphi_{w,k}) = (2K^2 + 2) \cdot \varphi_{w,k} \cdot e^{-(2K^2+2)\varphi_{w,k}} \tag{37}$$

We deliberately choose not to optimize the ρ parameter, and to investigate two particular cases, namely $\rho = 1$ and $\rho = 2$. In both cases, the above equations can be further simplified and are amenable an efficient implementation as described in the following section.

A.6 Scaling up the estimation

When $\rho = 1$ or $\rho = 2$ the Weibull distribution simplifies into to an exponential distribution or the Rayleigh distribution, respectively. In such a case, we can introduce the counters described in table 10 and rewrite the updated equations by relying on such counters. We analyze the two cases separately.

A.6.1 Exponential distribution

When $\rho = 1$, the term $h(\Delta^c_{u,v}; \lambda_{v,u,k}, \rho)$ simplifies to $A_{v,k} \cdot S_{u,k}$. As as consequence, the term $\eta^k_{c,u,v}$ can be rewritten as $\eta^k_{c,u,v} = A_{v,k}/A_{c,u,k}$. Then, we can can state the following.

Theorem 2. Equations 31 and 32 can be rewritten as

$$\boxed{\begin{aligned} S_{u,k} &= \frac{\sum_{c:u \in \mathcal{V}+(c)} \gamma_{c,k}}{\bar{S}_{u,k}} \\ \bar{S}_{u,k} &= \sum_{c:u \in \mathcal{V}+(c)} \gamma_{c,k} \left(t_u(c) A_{c,u,k} - \tilde{A}_{c,u,k}\right) \\ &\quad + \sum_{c:u \in \mathcal{V}-(c)} \gamma_{c,k} \left(T^c A_{c,k} - \tilde{A}_{c,k}\right) \end{aligned}} \tag{38}$$

by exploiting the counters defined in 10.

PROOF. We can observe that:

$$\begin{aligned} &\sum_{c:u \in \mathcal{V}+(c)} \sum_{v \prec_c u} \gamma_{c,k} \cdot \Delta^c_{u,v} \cdot A_{v,k} = \\ &\sum_{c:u \in \mathcal{V}+(c)} \gamma_{c,k} \sum_{v \prec_c u} (t_u(c) - t_v(c)) A_{v,k} = \\ &\sum_{c:u \in \mathcal{V}+(c)} \gamma_{c,k} \left(t_u(c) A_{c,u,k} - \tilde{A}_{c,u,k}\right) \end{aligned}$$

and

$$\begin{aligned} &\sum_{c:u \in \mathcal{V}-(c)} \sum_{v \in \mathcal{V}-(c)} \gamma_{c,k} \cdot \Delta^c_{u,v} \cdot A_{v,k} = \\ &\sum_{c:u \in \mathcal{V}-(c)} \gamma_{c,k} \sum_{v \in \mathcal{V}-(c)} (T^c - t_v(c)) A_{v,k} = \\ &\sum_{c:u \in \mathcal{V}-(c)} \gamma_{c,k} \left(T^c A_{c,k} - \tilde{A}_{c,k}\right) \end{aligned}$$

By plugging the above results in 31 we obtain the claim. □

Theorem 3. Equation 33 can be rewritten as

$$\boxed{\begin{aligned} A_{v,k} &= \frac{A'_{v,k} \sum_{c:v \in \mathcal{V}^+(c)} \gamma_{c,k} R_{c,v,k}}{\sum_{c:v \in \mathcal{V}^+(c)} \gamma_{c,k} \bar{A}_{v,k}} \\ \bar{A}_{v,k} &= \tilde{S}_{c,k} - \tilde{S}_{c,v,k} + T^c (S_k - S_{c,k}) \\ &\quad - t_v(c)(S_k - S_{c,v,k}) \end{aligned}} \tag{39}$$

by exploiting the counters defined in 10 and assuming that $A'_{v,k}$ is computed in the preceding step.

PROOF. First of all, Notice that, in this new formulation, $\eta^k_{u,v}$ does not need to be explicitly computed. In fact, it is decomposed within the update equation of $A_{v,k}$, by resorting to a previous value of the same variable (denoted as $A^{(n-1)}_{u,k}$ in the formula) and the $R_{c,u,k}$ counter:

$$\begin{aligned} &\sum_{c:v \in \mathcal{V}^+(c)} \sum_{\substack{u \in \mathcal{V}^+(c) \\ v \prec_c u}} \eta_{c;u,v} \cdot \gamma_{c,k} \\ &\qquad = \sum_{c:v \in \mathcal{V}^+(c)} \gamma_{c,k} \sum_{\substack{u \in \mathcal{V}^+(c) \\ v \prec_c u}} A'_{v,k} \left(A'_{c,u,k}\right)^{-1} \\ &\qquad = A'_{v,k} \sum_{c:v \in \mathcal{V}^+(c)} \gamma_{c,k} R_{c,v,k} \end{aligned}$$

Further,

$$
\begin{aligned}
&\sum_{c:v\in\mathcal{V}^+(c)}\sum_{u:v\prec_c u}\gamma_{c,k}\cdot\Delta^c_{u,v}\cdot S_{u,k}\\
&\quad+\sum_{c:v\in\mathcal{V}^+(c)}\sum_{u\in\mathcal{V}^-(c)}\gamma_{c,k}\cdot(T^c-t_v(c))\cdot S_{u,k}\\
&=\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}\sum_{u:v\prec_c u}(t_u(c)-t_v(c))S_{u,k}\\
&\quad+\sum_{c:v\in\mathcal{V}^+(c)}\sum_{u\in\mathcal{V}^-(c)}\gamma_{c,k}\cdot(T^c-t_v(c))\cdot S_{u,k}\\
&=\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}\sum_{u:v\prec_c u}t_u(c)S_{u,k}-\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}t_v(c)\sum_{u:v\prec_c u}S_{u,k}\\
&\quad+\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}\sum_{u\in\mathcal{V}^-(c)}T^cS_{u,k}-\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}t_v(c)\sum_{u\in\mathcal{V}^-(c)}S_{u,k}\\
&=\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}(\tilde{S}_{c,k}-\tilde{S}_{c,v,k})-\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}t_v(c)(S_{c,k}-S_{c,v,k})\\
&\quad+\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}T^c(S_k-S_{c,k})-\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}t_v(c)(S_k-S_{c,k})\\
&=\sum_{c:v\in\mathcal{V}^+(c)}\gamma_{c,k}\left\{\tilde{S}_{c,k}-\tilde{S}_{c,v,k}+T^c(S_k-S_{c,k})-t_v(c)(S_k-S_{c,v,k})\right\}
\end{aligned}
$$

□

By plugging the above results in 33 we obtain the claim. □

Theorem 4. The logarithm of $\Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k)$ as defined in eq. 28 can be expressed as

$$
\boxed{
\begin{aligned}
\log\Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k)=\;&L_{c,k}\\
&-(S_k-S_{c,k})(T^cA_{c,k}-\tilde{A}_{c,k})\\
&+\sum_{u\in\mathcal{V}^+(c)}\{\log A_{c,u,k}\\
&\qquad-S_{u,k}\,(t_u(c)A_{c,u,k}-\tilde{A}_{c,u,k})\}
\end{aligned}
}
\tag{40}
$$

by exploiting the counters defined in 10.

PROOF. We can observe the following:

$$
\begin{aligned}
\log\Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k)=&\sum_{u\in\mathcal{V}^+(c)}\log S_{u,k}+\sum_{u\in\mathcal{V}^+(c)}\log\left(\sum_{v\prec_c u}A_{v,k}\right)\\
&-\sum_{u\in\mathcal{V}^+(c)}S_{u,k}\sum_{v\prec_c u}\Delta^c_{u,v}\cdot A_{v,k}\\
&-\sum_{u\in\mathcal{V}^-(c)}S_{u,k}\sum_{v\in\mathcal{V}^+(c)}(T^c-t_v(c))\cdot A_{v,k}\\
=&L_{c,k}+\sum_{u\in\mathcal{V}^+(c)}\log A_{c,u,k}\\
&-\sum_{u\in\mathcal{V}^+(c)}S_{u,k}\sum_{v\prec_c u}(t_u(c)-t_v(c))\cdot A_{v,k}\\
&-\sum_{u\in\mathcal{V}^-(c)}S_{u,k}\sum_{v\in\mathcal{V}^+(c)}(T^c-t_v(c))\cdot A_{v,k}\\
=&L_{c,k}+\sum_{u\in\mathcal{V}^+(c)}\log A_{c,u,k}\\
&\;-\sum_{u\in\mathcal{V}^+(c)}S_{u,k}\,(t_u(c)A_{c,u,k}-\tilde{A}_{c,u,k})\\
&-(S_k-S_{c,k})(T^cA_{c,k}-\tilde{A}_{c,k})
\end{aligned}
$$

A.6.2 Rayleigh Distribution

Similar results can be obtained for the case $\rho = 2$. We start by noticing that, in this case, the EM algorithm relies on the following key components:

$$
\begin{aligned}
S_{u,k}&=\frac{\sum_{c:u\in\mathcal{V}^+(c)}\gamma_{c,k}}{\bar{S}_{u,k}}\\
\bar{S}_{u,k}&=\sum_{c:u\in\mathcal{V}^+(c)}\sum_{v\prec_c u}\gamma_{c,k}\cdot(\Delta^c_{u,v})^2\cdot A_{v,k}\\
&\quad+\sum_{c:u\in\mathcal{V}^-(c)}\sum_{v\prec_c u}\gamma_{c,k}\cdot(T^c-t_v(c))^2\cdot A_{v,k}
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
A_{v,k}&=\frac{\sum_{c:u\in\mathcal{V}^+(c)}\sum_{\substack{u\in\mathcal{V}^+(c)\\ v\prec_c u}}\eta^k_{c;u,v}\cdot\gamma_{c,k}}{\bar{A}_{v,k}}\\
\bar{A}_{v,k}&=\sum_{c:v\in\mathcal{V}^+(c)}\sum_{u:v\prec_c u}\gamma_{c,k}\cdot(\Delta^c_{u,v})^2\cdot S_{u,k}\\
&\quad+\sum_{c:v\in\mathcal{V}^+(c)}\sum_{u\in V^-(c)}\gamma_{c,k}\cdot(T_c-t_v(c))^2\cdot S_{u,k}
\end{aligned}
\tag{42}
$$

$$
\begin{aligned}
\Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k)=&\prod_{u\in\mathcal{V}^-(c)}\prod_{v\in\mathcal{V}^+(c)}e^{-A_{v,k}\cdot S_{u,k}\cdot(T^c-t_v(c)^2)}\\
&\cdot\prod_{u\in\mathcal{V}^+(c)}\prod_{v\prec_c u}e^{-A_{v,k}\cdot S_{u,k}\cdot(\Delta^c_{u,v})^2}\sum_{v\prec_c u}2A_{v,k}\cdot S_{u,k}\cdot\Delta^c_{u,v}
\end{aligned}
\tag{43}
$$

The following results hold.

Theorem 5. Equation 31 can be rewritten as

$$\boxed{\begin{aligned} S_{u,k} &= \frac{\sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k}}{\bar{S}_{u,k}} \\ \bar{S}_{u,k} &= \sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k} \Big(t_u(c)^2 A_{c,u,k} + \hat{A}_{c,u,k} \\ &\qquad -2t_u(c)\tilde{A}_{c,u,k} \Big) \\ &\quad + \sum_{c:u\in\mathcal{V}^-(c)} \gamma_{c,k} \Big(T_c^2 A_{c,k} + \hat{A}_{c,k} - 2T_c \tilde{A}_{c,k} \Big) \end{aligned}} \tag{44}$$

by exploiting the counters defined in 10.

PROOF. We can observe that

$$\begin{aligned} &\sum_{c:u\in\mathcal{V}^+(c)} \sum_{v\prec_c u} \gamma_{c,k} \cdot [t_u(c) - t_v(c)]^2 \cdot A_{v,k} \\ &= \sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k} \sum_{v\prec_c u} [t_u(c)^2 + t_v(c)^2 - 2t_u(c)t_v(c)] \cdot A_{v,k} \\ &= \sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k} \cdot t_u(c)^2 \sum_{v\prec_c u} A_{v,k} \\ &\quad + \sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k} \sum_{v\prec_c u} t_v(c)^2 A_{v,k} \\ &\quad - 2 \sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k} \cdot t_u(c) \sum_{v\prec_c u} t_v(c) A_{v,k} \\ &= \sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k} \Big(t_u(c)^2 \cdot A_{c,u,k} + \hat{A}_{c,u,k} - 2t_u(c) \cdot \tilde{A}_{c,u,k} \Big). \end{aligned}$$

Analogously

$$\begin{aligned} &\sum_{c:u\in\mathcal{V}^-(c)} \sum_{v\in V^+(c)} \gamma_{c,k} \cdot (T^c - t_v(c))^2 \cdot A_{v,k} \\ &\qquad = \sum_{c:u\in\mathcal{V}^-(c)} \gamma_{c,k} \Big((T^c)^2 \cdot A_{c,k} + \hat{A}_{c,k} - 2T^c \tilde{A}_{c,k} \Big) \end{aligned}$$

By replacing the formulas in the denominator of eq. 41 we obtain the claim. □

Theorem 6. Equation 33 can be rewritten as

$$\boxed{\begin{aligned} A_{v,k} &= \frac{A'_{v,k} \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \Big(\tilde{R}_{c,v,k} - \hat{R}_{c,v,k} \Big)}{\sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \bar{A}_{v,k,c}} \\ \bar{A}_{v,k,c} &= \Big(\hat{S}_{c,k} - \hat{S}_{c,v,k} \Big) + t_v(c)^2 \left(S_{c,k} - S_{c,v,k} \right) \\ &\quad - 2t_v(c) \Big(\tilde{S}_{c,k} - \tilde{S}_{c,v,k} \Big) \\ &\quad + \left(T_c^2 + t_v(c)^2 - 2T_c t_v(c) \right) \left(S_k - S_{c,k} \right) \end{aligned}} \tag{45}$$

where $A'_{v,k}$ is the value computed in the preceding step, and by exploiting the counters defined in 10.

PROOF. For the denominator, we have:

$$\begin{aligned} &\sum_{c:v\in\mathcal{V}^+(c)} \sum_{u:v\prec u} \gamma_{c,k} \cdot (t_u(c) - t_v(c))^2 \cdot S_{u,k} \\ &= \sum_{c:v\in\mathcal{V}^+(c)} \sum_{u:v\prec u} \gamma_{c,k} \cdot t_u(c)^2 S_{u,k} \\ &\quad + \sum_{c:v\in\mathcal{V}^+(c)} \sum_{u:v\prec u} \gamma_{c,k} \cdot t_v(c)^2 S_{u,k} \\ &\quad - 2 \sum_{c:v\in\mathcal{V}^+(c)} \sum_{u:v\prec u} \gamma_{c,k} \cdot t_u(c) t_v(c) \cdot S_{u,k} \\ &= \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \sum_{u:v\prec u} t_u(c)^2 S_{u,k} \\ &\quad + \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \cdot t_v(c)^2 \sum_{u:v\prec u} S_{u,k} \\ &\quad - 2 \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \cdot t_v(c) \sum_{u:v\prec u} t_u(c) S_{u,k} \\ &= \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \Big\{ \Big(\hat{S}_{c,k} - \hat{S}_{c,v,k} \Big) + t_v(c)^2 \left(S_{c,k} - S_{c,v,k} \right) \\ &\quad -2t_v(c) \Big(\tilde{S}_{c,k} - \tilde{S}_{c,v,k} \Big) \Big\} \end{aligned}$$

and

$$
\begin{aligned}
&\sum_{c:v\in\mathcal{V}^+(c)} \sum_{u\in V^-(c)} \gamma_{c,k}\cdot(T_c - t_v(c))^2 \cdot S_{u,k}\\
&= \sum_{c:v\in\mathcal{V}^+(c)} \sum_{u\in V^-(c)} \gamma_{c,k}\cdot T_c^2 S_{u,k}\\
&+ \sum_{c:v\in\mathcal{V}^+(c)} \sum_{u\in V^-(c)} \gamma_{c,k}\cdot t_v(c)^2 S_{u,k}\\
&-2 \sum_{c:v\in\mathcal{V}^+(c)} \sum_{u\in V^-(c)} \gamma_{c,k}\cdot t_v(c) T_c \cdot S_{u,k}\\
&= \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} T_c^2 \sum_{u\in V^-(c)} S_{u,k}\\
&+ \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} t_v(c)^2 \sum_{u\in V^-(c)} S_{u,k}\\
&-2 \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} T_c t_v(c) \sum_{u\in V^-(c)} S_{u,k}\\
&= \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \left\{ \left(T_c^2 + t_v(c)^2 - 2T_c t_v(c)\right)\left(S_k - S_{c,k}\right)\right\}
\end{aligned}
$$

Also, for the numerator, we have:

$$
\begin{aligned}
&\sum_{c:v\in\mathcal{V}^+(c)} \sum_{\substack{u\in\mathcal{V}^+(c)\\ v\prec_c u}} \eta^k_{c;u,v}\cdot\gamma_{c,k} =\\
&\sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \sum_{\substack{u\in\mathcal{V}^+(c)\\ v\prec_c u}} \frac{A_{v,k}\Delta^c_{u,v}}{\sum_{v\prec u} A_{v,k}\Delta^c_{u,v}} =\\
&\sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} A_{v,k} \sum_{\substack{u\in\mathcal{V}^+(c)\\ v\prec_c u}} \frac{t_u(c) - t_v(c)}{t_u(c)A_{c,u,k} - \tilde{A}_{c.u,k}} =\\
&\sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} A_{v,k} \left\{ \sum_{\substack{u\in\mathcal{V}^+(c)\\ v\prec_c u}} t_u(c)\left(t_u(c)A_{c,u,k} - \tilde{A}_{c.u,k}\right)^{-1}\right.\\
&\qquad \left. -t_v(c) \sum_{\substack{u\in\mathcal{V}^+(c)\\ v\prec_c u}} \left(t_u(c)A_{c,u,k} - \tilde{A}_{c.u,k}\right)^{-1}\right\} =\\
&\sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} A_{v,k}\left(\tilde{R}_{c,v,k} - \hat{R}_{c,v,k}\right)
\end{aligned}
$$

By replacing the respective components in eq. 42 we obtain the claim. □

We also notice that $\eta^k_{c,u,v}$ can be simplified into:

$$\eta^k_{c,u,v} = \frac{A_{v,k}\Delta^c_{u,v}}{\sum_{v\prec u} A_{v,k}\Delta^c_{u,v}} \tag{46}$$

$$\Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k) = \exp\{llk(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k)\} \tag{47}$$

Theorem 7. The logarithm of $\Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k)$ as defined in eq. 28 can be expressed as

$$
\boxed{
\begin{aligned}
&\log\Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k) = N_c\log 2 + L_{c,k}\\
&\quad + \sum_{u\in\mathcal{V}^+(c)} \log\left(t_u(c)A_{c,u,k} - \tilde{A}_{c,u,k}\right)\\
&\quad - \sum_{u\in\mathcal{V}^+(c)} S_{u,k}\left\{t_u(c)^2 A_{c,u,k} + \hat{A}_{c,u,k}\right.\\
&\qquad\qquad\qquad \left. -2t_u(c)\tilde{A}_{c,u,k}\right\}\\
&\quad - S_k\left(T_c^2 A_{c,k} + \hat{A}_{c,k} - 2T_c\tilde{A}_{c,k}\right)
\end{aligned}
}
\tag{48}
$$

by exploiting the counters defined in 10.

PROOF. By analyzing eq. 43, we observe that

$$
\begin{aligned}
&\log\Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k) = N_c\log 2\\
&\quad + \sum_{u\in\mathcal{V}^+(c)} \log\left(\sum_{v\prec_c u}\Delta^c_{u,v} A_{v,k} S_{u,k}\right)\\
&\quad - \sum_{u\in\mathcal{V}^+(c)} \sum_{v\prec_c u} (\Delta^c_{u,v})^2 \cdot A_{v,k} S_{u,k}\\
&\quad - \sum_{u\in\mathcal{V}^-(c)} \sum_{v\in\mathcal{V}^+(c)} (T^c - t_v(c))^2 \cdot A_{v,k} S_{u,k}
\end{aligned}
$$

Within the equation, the second row expands as follows:

$$
\begin{aligned}
&\sum_{u\in\mathcal{V}^+(c)} \log\left(\sum_{v\prec_c u}(t_u(c) - t_v(c))A_{v,k}S_{u,k}\right) =\\
&\sum_{u\in\mathcal{V}^+(c)} \log S_{u,k} + \log\left(t_u(c)\sum_{v\prec_c u} A_{v,k} - \right.\\
&\qquad\qquad \left. \sum_{u\in\mathcal{V}^+(c)} \sum_{v\prec_c u} t_v(c) A_{v,k}\right) =\\
&L_{c,k} + \sum_{u\in\mathcal{V}^+(c)} \log\left(t_u(c)A_{c,u,k} - \tilde{A}_{c,u,k}\right)
\end{aligned}
$$

Concerning the third row, we observe:

$$
\begin{aligned}
&\sum_{u\in\mathcal{V}^+(c)} \sum_{v\prec_c u} (t_u(c) - t_v(c))^2 \cdot A_{v,k}S_{u,k} =\\
&\sum_{u\in\mathcal{V}^+(c)} \sum_{v\prec_c u} (t_u(c)^2 A_{v,k}S_{u,k} + t_v(c)^2 A_{v,k}S_{u,k}\\
&\qquad\qquad - 2t_u(c)t_v(c)A_{v,k}S_{u,k}) =\\
&\sum_{u\in\mathcal{V}^+(c)} S_{u,k}\left\{t_u(c)^2 A_{c,u,k} + \hat{A}_{c,u,k} - 2t_u(c)\tilde{A}_{c,u,k}\right\}
\end{aligned}
$$

Finally, for the fourth row we have:

$$\sum_{u\in\mathcal{V}^-(c)} \sum_{v\in\mathcal{V}^+(c)} (T_c - t_v(c))^2 \cdot A_{v,k}S_{u,k} =$$

$$\sum_{u\in\mathcal{V}^-(c)} S_{u,k} \sum_{v\in\mathcal{V}^+(c)} (T_c^2 A_{v,k} + t_v(c)^2 A_{v,k} - 2T_c t_v(c) A_{v,k}) =$$

$$S_k \left(T_c^2 A_{c,k} + \hat{A}_{c,k} - 2T_c \tilde{A}_{c,k}\right)$$

Putting all together proves the claim. □

A.6.3 Complexity result

Algorithm 2 Optimized Survival Factorization EM

```
Require: $\mathcal{C}$, the number of latent features $K$
Ensure: matrices $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Phi}$
 1: Randomly initialization for $\mathbf{A}$, $\mathbf{S}$, $\mathbf{\Phi}$;
 2: Compute all counters of table 10;
 3: $n \leftarrow 0$
 4: while Increment in Likelihood is negligible do
 5:    for all cascades $c$ and topic $k$ do
 6:       Compute $\gamma_{c,k}$ exploiting $\log \Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k)$ as defined in Eq. 19
          (exponential) or Eq. 48 (Rayleigh);
 7:    end for
 8:    for all topic $k$ do
 9:       Update $\pi_k$ according to Eq. 13;
10:       for all users $u$ do
11:          Compute $S_{u,k}$ according to Eq. 17 (exponential) or Eq. 44
             (Rayleigh);
12:       end for
13:       Update all counters relative to $\mathbf{S}$ as defined in table 10;
14:       for all users $u$ do
15:          Compute $A_{u,k}$ according to Eq. 18 (exponential) or Eq. 45
             (Rayleigh);
16:       end for
17:       Update counters relative to $\mathbf{A}$ as defined in table 10;
18:       for all words $w$ do
19:          Compute $\phi_{w,k}$ according to Eq. 14;
20:       end for
21:    end for
22:    $n \leftarrow n + 1$
23: end while
```

The general scheme of the algorithm is shown in figure 2. For this algorithm, we can finally state the main complexity result.

Theorem 8. *Algorithm 2 has complexity* $O(\sum_c N_c \log N_c + nK(N + W + \sum_c N_c))$ *time (where* n *is the total number of iterations) and* $O(KN)$ *space.*

PROOF. The crucial steps of the algorithm are in line 1 and 10-17. First of all, let us consider the counters of table 10. Each of them can be computed through at least two scans over the set $\mathcal{C}$ of cascades, by considering that when cascades are sorted counters can be computed incrementally for adjacent nodes. This ensures that steps 2,13 and 17 can be computed in $K\sum_c N_c$. Also, notice that a further scan over $\mathcal{C}$ allows to incrementally compute, for each u within a cascade, the contributions to equations 19, 18 and 17: for the latter, the second term in the denominator can be obtained by considering the difference between the sum over all cascades and the sum over the cascades that contain u. Thus, updating $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{\Phi}$ only requires a loop over all the elements of the matrices, provided that all the counters and the terms of the equations are conveniently accumulated as the cascades are iterated. □

The major difference between the two instantiations is the underlying definition of the $\eta^k_{c,u,v}$ parameter. Within the exponential distribution, this is expressed as $\eta^k_{c,u,v} = \frac{A_{v,k}}{\sum_{v \prec u} A_{v,k}}$. Since $\eta^k_{c,u,v}$ represents the probability that node v influences the activation of u within c, the above formulation resolves in a “winner takes all” situation. That is, the element with the higher degree of authoritativeness is deemed as the influencer, no matter the time delay. Thus for example, assume that user v_1 activates at time $t_{v_1} = 1$, user v_2 at time $t_{v_2} = 100$ and user u at time $t_u = 101$. Also, assume that $A_{v_1,k} = 2$, $A_{v_2,k} = 1$ and $S_{u,k} = 1$. By virtue of the definition of η, the influencer would be v_1, but the delay $t_u - t_{v_1}$ would be extremely unlikely according to the given parameters.

This bias does not hold for the Rayleigh distribution, for which the $\eta^k_{c,u,v}$ includes the delay, $\eta^k_{c,u,v} = \frac{A_{v,k}\Delta^c_{u,v}}{\sum_{v \prec u} A_{v,k}\Delta^c_{u,v}}$, thus recovering the time dependency in the detection of the influencer.