Survival Factorization on Diffusion Networks

Nicola Barbieri[1], Giuseppe Manco[2], and Ettore Ritacco[2]

[1] Tumblr, 35 E 21st St, 10010, New York, USA,
nicola@tumblr.com
[2] ICAR - CNR, via Pietro Bucci 7/11C, 87036 Arcavacata di Rende (CS), ITALY,
giuseppe.manco@icar.cnr.it, ettore.ritacco@icar.cnr.it

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 a 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.

Keywords: Social Network Analysis, Survival Analysis, Information Diffusion, Influence Propagation, Adoption Prediction.

1 Introduction

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?

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. Moreover, in many scenarios of interest for this paper, we are able to only observe cascades, having no information about the network structure (users' interconnections).

In this setting, 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 topical pairwise transmission rates $\lambda^k_{u,v}$, explaining the influence exerted by user v on u according to the topic k (see e.g. [18]). This approach has three main drawbacks: a large number of parameters (i.e. it's prone to overfitting); the parameter inference does not scale well; poor estimates when 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, $\varphi_{w,k}$.

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

- We review previous studies on information diffusion (Sec. 2) and briefly introduce a survival framework for modeling information diffusion (Sec. 3).
- Next, we introduce a factorization model (Sec. 3.1) that expresses topical pairwise transmission rates in terms of user authority and susceptibility, by coupling the topical content of a cascade and the observed activation times.
- We devise a highly scalable expectation maximization algorithm (Sec. 4) for the model parameter learning.
- 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.

2 Related work

Starting from seminal studies [13, 21, 9], the research on information diffusion 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 [8, 7] studies a different perspective, where the social network is not given in 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. For example, in NetRate [7], if node u succeeds in activating v, then the contagion of the latter happens after an incubation period sampled from a chosen distribution. According to this propagation model, the likelihood of a propagation cascade can be formulated by applying standard survival analysis [14]. Recent extensions of the survival diffusion process exploit Poisson [12] or Hawkes processes [5, 22].

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

	Time	*Req. Network*	*Inference*	*Side Info*	*Clustering*
NetRate [7]	contin.	no	$O(N^2)$	no	no
MONET [17]	contin.	yes	$O(N^2)$	nodes	no
MMRate [18]	contin.	no	$O(N^2)$	no	cascades
CSDK [4]	contin.	no	$O(NM)$	cascades	no
LIS [19]	discrete	no	$O(N^2)$	no	no
AIR [1]	discrete	yes	$O(N)$	cascades	nodes
CCN [2]	discrete	yes	$O(N)$	no	nodes
CWN [3]	both	no	$O(N)$	no	nodes
Our method	contin.	no	$O(N)$	cascades	cascades

A different research line extends the diffusion process by considering enhancements based on features [17], or topics which characterize cascades [18, 6, 3, 11, 10]. These models assume that the diffusion speed depends on node connections, features characterizing users and cascades, and node topical affinity [18, 6, 10].

Recent works have also focused on alternative ways of representing interactions between nodes, using latent-dimensional embedding techniques. In [4] 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 approaches described so far do not explicitly consider the diffusion process as a result of the interaction between influence and susceptibility. In [1, 3], the probability of activation is modeled as the effect of the influence of neighbor nodes within the cascades and/or the network. Further, the approaches [2, 19] 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.

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

A cascade represents the propagation of a piece of information (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 when the 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 reposted 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 it 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 [13], we assume that node 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 [7], 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 [7]:

$$\begin{aligned}\Pr(\mathbf{t}^c|\Theta) = &\prod_{u \in \mathcal{V}^-(c)} \prod_{v \in \mathcal{V}^+(c)} S(T^c - t_v(c); \lambda_{v,u}) \cdot \\ &\prod_{u \in \mathcal{V}^+(c)} \prod_{v \prec_c u} S(\Delta^c_{u,v}; \lambda_{v,u}) \cdot \sum_{v' \prec_c u} h(\Delta^c_{u,v'}; \lambda_{v',u}) \,,\end{aligned} \tag{3.1}$$

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 $\mathcal{W}$:

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

3.1 Factorization Model

We start from the idea that the temporal dynamics, governing the activations of each node within observed cascades, depends 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 discover 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 likelihood 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:

$$\lambda_{v,u,k} = A_{v,k} \cdot S_{u,k}\,. \tag{3.2}$$

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* [16] 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.

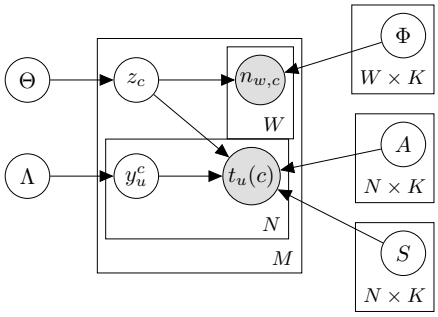


Fig. 1: Graphical model of Survival Factorization.

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:

$$f(t_u(c)|t_v(c), \lambda_{v,u,k}) = \mathcal{W}eib(\Delta_{u,v}^c; \lambda_{v,u,k}, \rho)\,. \tag{3.3}$$

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:

$$h(t; \lambda, \rho) = \rho\lambda t^{\rho-1}\,, \tag{3.4}$$

$$S(t; \lambda, \rho) = e^{-\lambda t^\rho}\,. \tag{3.5}$$

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 $y_{u,v}^c$, with the constraint $\sum_{v\in\mathcal{V}} y_{u,v}^c = 1$. Let $\mathbf{Y}$ denote a $M \times N \times N$ binary matrix, where $y_{u,v}^c = 1$ 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 $y_{u,v}^c = 1$ is governed by a multinomial distribution Λ [3].
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:

$$\begin{aligned}\Pr(\mathbf{t}^c|\mathbf{Z}, \mathbf{Y}, \mathbf{A}_k, \mathbf{S}_k) = \prod_k \Bigg(\prod_{u\in\mathcal{V}^-(c)} \prod_{v\in\mathcal{V}^+(c)} S(T^c - t_v(c); \lambda_{v,u,k}, \rho) \\ \cdot \prod_{u\in\mathcal{V}^+(c)} \prod_{v\prec_c u} h(\Delta_{u,v}^c; \lambda_{v,u,k}, \rho)^{y_{u,v}^c} \cdot S(\Delta_{u,v}^c; \lambda_{v,u,k}, \rho) \Bigg)^{z_{c,k}}\,.\end{aligned} \tag{3.6}$$

[3] In the next we shall assume that this distribution is uniform, i.e., each v has equal chances of activating u.

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. 3.1, the above model exhibits two main differences. First, cascade are characterized by a topic which also governs the propagation speed. Second, we explicitly model influencers by introducing the $\mathbf{Y}$ matrix. In fact, eq. 3.6 is a refined extension of eq. 3.1, 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* [16], 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:

$$\Pr(\mathbf{w}^c|\mathbf{\Phi}_k) = \prod_{w} \frac{(|\mathbf{w}^c| \cdot \varphi_{w,k})^{n_{w,c}} \exp\{-|\mathbf{w}^c| \cdot \varphi_{w,k}\}}{n_{w,c}!} \,. \tag{3.7}$$

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

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

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

$$\Pr(\mathcal{C}, \Xi) = \sum_{\mathbf{Z},\mathbf{Y}} \Pr(\mathcal{C}|\mathbf{Z}, \mathbf{Y}, \Xi) \Pr(\mathbf{Z}, \mathbf{Y}, \Xi) \,. \tag{4.1}$$

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. That is, it is easy to devise an iterative alternating strategy consisting of the following two steps:

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

M step: exploit the posterior to solve

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

Both steps are tractable and the estimation produces closed formulas. The details of the derivations can be found in the appendix submitted as supplemental material.

In particular, for the E step the estimation of $\Pr(\mathbf{Z},\mathbf{Y}|\mathcal{C},\boldsymbol{\Xi}^{(n)})$ 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

$$\eta^k_{c,u,v} = \frac{h(\Delta^c_{u,v};\lambda_{v,u,k},\rho)}{\sum_{v' \prec_c u} h(\Delta^c_{u,v'};\lambda_{v',u,k},\rho)}\,, \tag{4.2}$$

$$\gamma_{c,k} = \frac{\Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k)\Pr(\mathbf{w}^c|\Phi_k)\pi_k}{\sum_k \Pr(\mathbf{t}^c|\mathbf{A}_k,\mathbf{S}_k)\Pr(\mathbf{w}^c|\Phi_k)\pi_k}\,. \tag{4.3}$$

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 3.7, 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 3.6 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 $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ can be obtained directly:

$$\pi_k = \frac{1}{M}\sum_c \gamma_{c,k} \tag{4.4}$$

$$\varphi_{w,k} = \frac{\sum_c \gamma_{c,k} n_{w,c}}{\sum_c \gamma_{c,k}|\mathbf{w}^c|} \tag{4.5}$$

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

$$S_{u,k} = \frac{\sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k}}{\sum_{c=1}^{M}\sum_{v\prec_c u} \gamma_{c,k}\cdot(\Delta^c_{u,v})^\rho \cdot A_{v,k}} \tag{4.6}$$

$$A_{v,k} = \frac{\sum_{c:v\in\mathcal{V}^+(c)} \sum_{\substack{u\in\mathcal{V}^+(c)\\ v\prec_c u}} \eta^k_{c;u,v}\cdot\gamma_{c,k}}{\sum_{c:v\in\mathcal{V}^+(c)}\sum_u \gamma_{c,k}\cdot(\Delta^c_{u,v})^\rho \cdot S_{u,k}} \tag{4.7}$$

We deliberately choose not to optimize the ρ parameter, and to investigate the case $\rho = 1$.

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 (see appendix[4] for details). Algorithm 1 describes the overall procedure for estimating the parameters.

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.

Algorithm 1 Optimized Survival Factorization EM

Require: $\mathcal{C}$, the number of latent features K
Ensure: matrices $\mathbf{A}$, $\mathbf{S}$ and $\boldsymbol{\Phi}$
Randomly initialization for $\mathbf{A}$, $\mathbf{S}$, $\boldsymbol{\Phi}$;
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. 4.10;
  end for
  for all topic k **do**
    Update π_k according to Eq. 4.4;
    for all users u **do**
      Compute $S_{u,k}$ according to Eq. 4.8;
    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. 4.9;
    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. 4.5;
    end for
  end for
  $n \leftarrow n + 1$
end while

$$S_{u,k} = \frac{\sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k}}{\sum_{c:u\in\mathcal{V}^+(c)} \gamma_{c,k}\left(t_u(c)A_{c,u,k} - \tilde{A}_{c,u,k}\right) + \sum_{c:u\in\mathcal{V}^-(c)} \gamma_{c,k}\left(T^c A_{c,k} - \tilde{A}_{c,k}\right)} \tag{4.8}$$

$$A_{v,k} = \frac{A_{v,k}^{(n-1)} \sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} R_{c,v,k}}{\sum_{c:v\in\mathcal{V}^+(c)} \gamma_{c,k} \left\{ \begin{array}{c} \tilde{S}_{c,k} - \tilde{S}_{c,v,k} \\ +T^c(S_k - S_{c,k}) \\ -t_v(c)(S_k - S_{c,v,k}) \end{array} \right\}} \tag{4.9}$$

$$\begin{aligned} \log \Pr(\mathbf{t}^c|\mathbf{A}_k, \mathbf{S}_k) =& L_{c,k} - (S_k - S_{c,k})(T^c A_{c,k} - \tilde{A}_{c,k}) \\ &+ \sum_{u\in\mathcal{V}^+(c)} \left\{ \log A_{c,u,k} - S_{u,k} \ (t_u(c) A_{c,u,k} - \tilde{A}_{c,u,k}) \right\} \end{aligned} \tag{4.10}$$

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

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

Proof. See appendix[4]. □

5 Evaluation

The following experimental evaluation is aimed at exploring the following aspects: (1) Investigate 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 we used in the experiments can be found at `http://github.com/gmanco/SurvivalFactorization`.

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.

We base the generation on the assumption (studied, e.g., in [20]) that vertices are connected and the diffusion of information happens through the links of the underlying network. Thus, to generate synthesized data, we, firstly, build networks with a known community structure by varying connectivity structure of the network. To this aim, we borrow the synthetic networks studied in [3].

Given a network $G = (V, E)$, we next generate synthetic propagation cascades by simulating a propagation process which spreads over E. The process gener-

ates $|\mathcal{I}|$ 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 remaning 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.

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.

	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.

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 $\mathcal{C}_{test}$[4]. 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. 4.3. 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 a chronological split proportion of 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

[4] 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.

$u \in c_3$; 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.

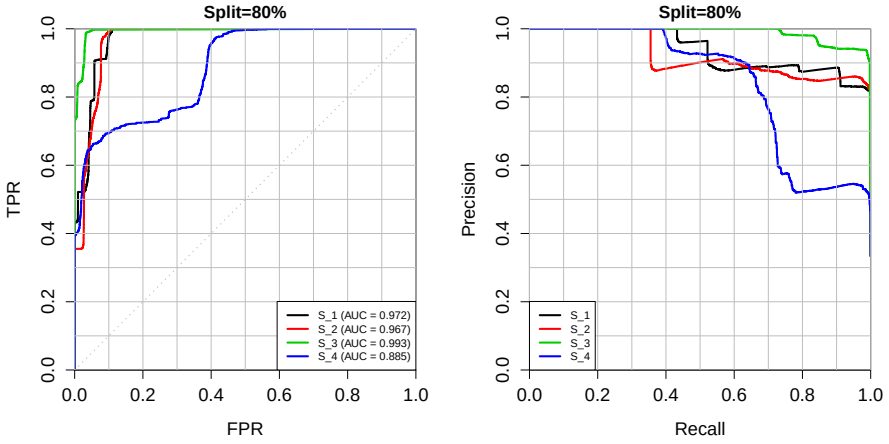


Fig. 3: AUC and Precision/Recall on predicting the activation time over synthetic data.

The results of the experiments, reported in Fig. 3, show that the proposed method is effective in predicting activation behaviour 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.

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. Secondly, we analyze and discuss topical and diffusion patterns inferred on the Memetracker dataset.

Twitter The following analysis is based on a sample of real-world propagation cascades crawled from the public timeline of *Twitter* and studied in [2]. 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). 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 4.

	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 4: Summary of the Twitter data used for evaluation.

Predicting activation times. We apply the testing protocol detailed in Sec. 5.1 on the Twitter datasets for predicting users retweet times, by considering two training/test chronological split (80%) and measuring prediction accuracy by ROC analysis. Results, reported in Fig. 4, 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.

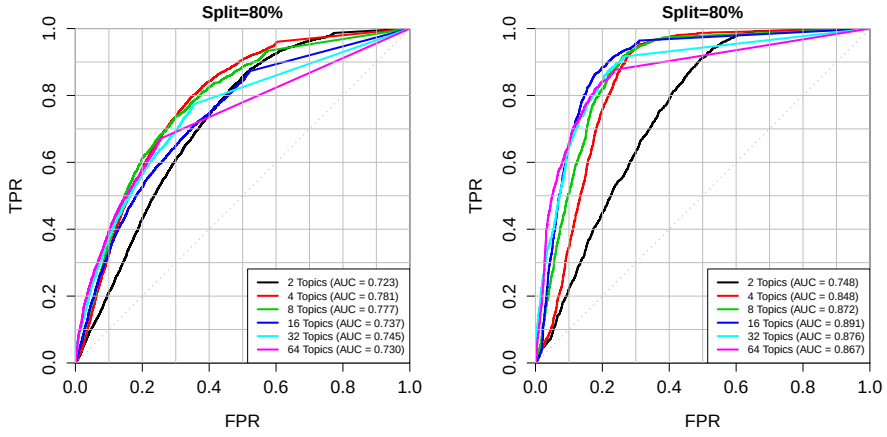


Fig. 4: Accuracy on predicting user's retweet time on *Twitter-Large* (on the left) and *Twitter-Small* (on the right).

Memetracker The evaluation on the Memetracker dataset [15] 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 cluster. We consider each 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 cluster. The raw dataset was cleaned from cascades with less than 10 activations and less than 10 words as content, and from vertices that belong to less than 10 cascades. The final dataset contains 7k vertices and 28k cascades, the word dictionary contains 3.5k tokens, with of 16 words for cascade on average.

For the sake of presentation, we run the survival factorization learning algorithm setting $K = 8$. Table 5 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.

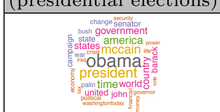
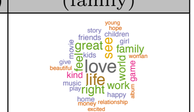
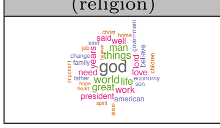
Topic 1 (economy)	Topic 2 (France/Germany)
	
Topic 3 (presidential elections)	**Topic 4 (family)**
	
Topic 5 (international crisis)	**Topic 6 (news in spanish)**
	
Topic 7 (religion)	**Topic 8 (sport)**
	

Table 5: Most relevant terms for each topic.

Topic 1 (economy)	Topic 2 (France/Germany)
	
Topic 3 (presidential elections)	**Topic 4 (family)**
	
Topic 5 (international crisis)	**Topic 6 (news in spanish)**
	
Topic 7 (religion)	**Topic 8 (sport)**
	

Table 6: Most influential hosts for each topic.

Next, we analyze each cascade and compute:

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

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

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 7: Characterization of the cascade trees for each topic.

Finally, Table 6 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 5.

6 Conclusions

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.

There are different points worth further investigation. For example, we showed that the instantiation based on the Exponential distribution admit an efficient implementation. In future work we will study if this property holds on other models, e.g. Rayleigh. Also, the robustness of the model can be improved by relying on a full bayesian framework.

References

1. N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. In *ICDM*, pages 81–90, 2012.
2. N. Barbieri, F. Bonchi, and G. Manco. Cascade-based community detection. In *WSDM*, pages 33–42, 2013.
3. N. Barbieri, F. Bonchi, and G. Manco. Influence-based network-oblivious community detection. In *ICDM*, pages 955–960, 2013.
4. S. Bourigault et al. Learning social network embeddings for predicting information diffusion. In *WSDM*, pages 393–402, 2014.
5. 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.
6. N. Du, L. Song, H Woo, and H. Zha. Uncover topic-sensitive information diffusion networks. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 229–237, 2013.
7. M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *ICML*, pages 561–568, 2011.
8. Manuel Gomez-Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *KDD*, 2010.
9. A. Goyal, F. Bonchi, and L. Lakshmanan. Learning influence probabilities in social networks. In *WSDM*, 2010.
10. 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.
11. 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.
12. 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.
13. David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.
14. E. T Lee and J. Wang. *Statistical methods for survival data analysis*. Wiley-Interscience, 2003.
15. Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, pages 497–506, 2009.
16. Qiaozhu Mei, Hui Fang, and ChengXiang Zhai. A study of poisson query generation model for information retrieval. In *SIGIR*, pages 319–326, 2007.
17. Liaoruo Wang, Stefano Ermon, and John E Hopcroft. Feature-enhanced probabilistic models for diffusion network inference. In *ECMLPKDD*, pages 499–514. 2012.
18. 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.
19. Y. Wang, H. Shen, S. Liu, and X. Cheng. Learning user-specific latent influence and susceptibility from information cascades. In *AAAI*, 2015.
20. L. Weng, F. Menczer, and Y. Ahn. Virality prediction and community structure in social networks. *Scientific reports*, 3, 2013.
21. R. Xiang, J. Neville, and M. Rogati. Modeling relationship strength in online social networks. In *WWW*, pages 981–990, 2010.
22. S. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. In *ICML*, pages 1–9, 2013.