Influence learning for cascade diffusion models: focus on partial orders of infections

Sylvain Lamprier, Simon Bourigault, Patrick Gallinari

Influence Learning for Cascade Diffusion Models: Focus on Partial Orders of Infections

Sylvain Lamprier · Simon Bourigault · Patrick Gallinari

Abstract Probabilistic cascade models consider information diffusion as an iterative process in which information transits between users of a network. The problem of diffusion modeling then comes down to learning transmission probability distributions, depending on hidden influence relationships between users, in order to discover the main diffusion channels of the network. Various learning models have been proposed in the literature, but we argue that the diffusion mechanisms defined in most of these models are not well-adapted to deal with noisy diffusion events observed from real social networks, where transmissions of content occur between humans. Classical models usually have some difficulties for extracting the main regularities in such real-world settings. In this paper, we propose a relaxed learning process of the well-known Independent Cascade model that, rather than attempting to explain exact timestamps of users' infections, focus on infection probabilities knowing sets of previously infected users. Furthermore, we propose a regularized learning scheme that allows the model to extract more generalizable transmission probabilities from training social data. Experiments show the effectiveness of our proposals, by considering the learned models for real-world prediction tasks.

Keywords Information Diffusion · Independent Cascade · Machine Learning

Sylvain Lamprier
Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France
CNRS, UMR 7606, LIP6, F-75005, Paris, France
E-mail: sylvain.lamprier@lip6.fr

Simon Bourigault
Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France
CNRS, UMR 7606, LIP6, F-75005, Paris, France
E-mail: simon.bourigault@lip6.fr

Patrick Gallinari
Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France
CNRS, UMR 7606, LIP6, F-75005, Paris, France
E-mail: patrick.gallinari@lip6.fr

1 Introduction

Recently, a huge amount of research has focused on Social Networks and Social Media sites, whose importance continuously grows, as the data they produce become more and more numerous and valuable. They today stand as inescapable sources of social data for several generic problems, such as Community Detection, Collaborative Recommendation or Link Prediction. In this context, the study of temporal content propagation (or information diffusion) corresponds to a very active topic, which may be useful for several concrete tasks. It aims at studying how a given content spreads through the network, via interactions between users, by a so-called word-of-mouth phenomenon. The study of such a phenomenon firstly emerged in epidemiology and social sciences contexts, for predicting and understanding spreads of diseases or marketing innovations. The emergence of social networks opened a very large number of new related research directions. Classical diffusion models, such as the independent cascade model (IC) [4,20] or the linear threshold model (LT) [10], have been applied to social data for capturing the dynamics of observed propagation through networks. In the ground of such general models, many different prediction tasks have recently emerged, such as *Diffusion prediction* - predicting which (or how many) users will be reached by a given content knowing its initial locations in the network [10,14] -, *Buzz detection* - estimating the impact of a content over the network [17]-, or *Leader identification* - identifying most influential users of the network [10,14].

In this paper, we focus on cascade models, that are at the heart of the research literature on information diffusion. In a natural way, these probabilistic models regard the phenomenon of diffusion as an iterative process in which information transits from users to next ones in the network [20,26,5,23]. In such a setting, the problem of diffusion modeling comes down to learn probability distributions depending on hidden influence relationships between users, in order to discover the main communication channels of the network. These iterative models, whose probability learning process consider sequences of infections rather than only dealing with some initial and final sets of infected users, usually leads to discover finer-grained influence relationships, as they enable to distinguish transitive influences in the network.

Various cascade models have been proposed in the literature, each inducing its own learning process to explain some observed diffusion episodes and attempting to extract relevant probability distributions of content transmission between users of the social media. The proposed approaches differ on their way of dealing with observed users' infection timestamps[1]. Some classical models, such as the Independent Cascade Model (IC) [4], iterate over successive time-steps to simulate diffusion episodes. Other models consider asynchronous diffusion processes, in which timestamps of infections are driven by some time delay distributions [19,5,3]. We argue that infection delays are in fact sampled

[1] Throughout this paper, we indifferently talk of infection or contamination to denote the fact that the propagated content has reached a given user of the network.

at near random in diffusion on real-world networks, at least on those where content transmissions occur between human nodes, and then, time regularities are very difficult to extract from such temporally linked data. While this is needed for some applications where dated infection predictions are required, the consideration of diffusion delays may greatly disturb the learning process when the main concern is to extract the transmission relationships of a social media (e.g., for tasks such as best influencers identification, buzz detection, final infections prediction, diffusion-based community detection, etc...).

Therefore, we propose to relax the problem of diffusion by designing a delay-agnostic learning of IC, which does not consider relative timestamps of infection during its training phase. We consider a likelihood defined on partial orders of infections rather than on exact infection time-steps as classically done in [20]. By focusing on infection sequences during its learning phase, our *Delay-Agnostic IC* model is able to better extract the regularities of real-world social data and capture the main diffusion channels of the studied networks.

The paper is organized as follows: Section 2 presents our model and its learning process. Section 3 compares our model to several baselines on real and artificial datasets. Section 4 reviews related works. Section 5 concludes the work.

2 A Delay-Agnostic Diffusion Model

2.1 Background and notations

Traditionally, information diffusion in a network is observed as a set of diffusion episodes $\mathscr{D} = (D_1, D_2, \ldots, D_n)$, where each diffusion episode is a sequence of related events, associated with their timestamps of occurrence. A diffusion episode describes the diffusion of a given content in the network[2]. For instance, it can correspond to a sequence of users' infections by some information at different timestamps: A set of users who "liked" a specific YouTube video, posted a given url, replied to a given message, etc... It describes to whom and when an item spreads through the network, but not *how* diffusion happens: the information of who infected who is unknown in such observed inputs.

Given a social network composed of a set of N users $\mathscr{U} = (u_1, ..., u_N)$, a diffusion episode D is then defined as a set of infected users associated with their timestamp of infection: $D = \{(u, t^D(u)) | u \in \mathscr{U} \wedge t^D(u) < \infty\}$, where $t^D : \mathscr{U} \rightarrow \mathbb{N}$ gives infection timestamps for users infected by the diffusion in concern, or ∞ for non infected ones. Timestamps returned by t^D are relative timestamps given the one of the first infected user (i.e., the source of diffusion, for which t^D then equals 0). In the following, we note U_v^D the set of users having been infected before user v in the diffusion D: $U_v^D = \{u \in \mathscr{U} | t^D(u) < t^D(v)\}$. We also

[2] The extraction of diffusion sequences from the data, which may be not straightforward with non-binary participations to the diffusion or in the case of a polymorphic diffused content, is not of our concern here. We assume diffusion episodes already extracted by a preliminary process.

note U_∞^D the whole set of users that have finally been infected by D and $\bar{U}_\infty^D$ those that have not.

Cascades are richer structures than diffusion episodes, as they explain how a given diffusion happened. A cascade $C = (S_C, U_C, T_C)$ corresponds to a transmission tree starting from sources of diffusion $S_C \subseteq \mathscr{U}$ and reaching a set of infected users $U_C \subseteq \mathscr{U}$ (with $S_C \subseteq U_C$), given a set T_C of timestamped transmission events between users from U_C. Note that, while several transmission events to a same given destination user might succeed during the diffusion process, the cascade structure only contains the first transmission event $(u \rightarrow v)$ that succeeded from any user u to the destination user v (which happens at the infection's timestamp of user v, as reported in diffusion episodes). For a given observed diffusion episode D, the set of possible cascade structures that generated D is thus given by $\mathscr{C}^D = \{C = (U_1^D, U_\infty^D, T_C) | \forall v \in (U_\infty^D \setminus U_1^D) \exists u \in U_v^D, (u \rightarrow v) \in T_C \wedge (\nexists u' \in \mathscr{U} \setminus \{u\}, (u' \rightarrow v) \in T_C)\}$, i.e., each infection is explained by a unique transmission from a previously infected user. Several different cascade structures are possible for a given observed sequence of infections. Cascade models usually perform assumptions on these latent diffusion structures for building their influence graphs.

Cascade models aim at defining an influence oriented graph $G = (\mathscr{U}, \mathscr{I})$, where $\mathscr{I}$ corresponds to the set of influence relationships between users of the network. Depending to the available data and the task, $\mathscr{I}$ can be restricted to relationships from a given known graph of possible influences (the graph of the social network for example), or can be defined as a complete graph allowing influences between all possible pairs of users (see discussion about this point in section 3). In the following, $Preds_u$ and $Succs_u$ respectively correspond to the sets of predecessors and successors of a user u w.r.t. relationships in $\mathscr{I}$ (users that can influence or be influenced by u). $I_{u,v} \in \mathscr{I}$ then corresponds to the directed influence relationship from a user $u \in Preds_v$ to a user $v \in Succs_u$. It is weighted by a function $P_{u,v} : \mathbb{N} \rightarrow [0,1]$ defining the probability of infection $P_{u,v}(t)$ of user v by user u after a time delay t. Note that we focus here on probabilities that do not depend on previous attempts of diffusion: Success or failures of diffusion between users are independent events.

2.2 Delay-Agnostic IC

The goal of the learning process of a cascade model is then to estimate diffusion probability distributions for each relationship among a given set of users U. As pointed out in the introduction, two main kinds of models can be found in the literature to infer these distributions.

On the one hand, time-step based approaches, such as those used for learning diffusion probabilities in *IC* [20], focus on diffusion events belonging to contiguous steps (defining then a probability function that only returns non-null values when the time-delay argument t equals 1). This enables to easily define a likelihood of generating cascades of observed time-steps of infections, since a user can only be infected by a user from the previous step (assuming

that she has been infected by at least one user from the previous step and not by users from preceding ones)[20]. However, assuming that infections can only be observed along contiguous time-steps is a very strong assumption that does not hold in real-world settings: influences between some pairs of users may require more time than between others without being less likely. Moreover, such a model is greatly dependent on the step size that is defined to discretize time and gather infections: with too large steps, a too large amount of users are gathered together which greatly biases the model since diffusion is assumed to only hold between users from two successive steps. With too short steps on the contrary, the process contains several empty steps, which induces a large amount of non-explained infections (infections of users from a step following an empty one cannot be explained by the model) and widely reduces the diffusion expectation (episodes with a possible diffusion along a given relationship are more rare). Even if empty steps were ignored during the learning process (empty steps can indeed be removed to enable more explanations of user's infections during the learning of the model), it still remains that a short step usually reduces the possibilities of latent cascade structures to a unique straight chain of infections. This greatly limits the ability of learning models that well explain the observed infections of users. Figure 1 depicts this dependency with regards to the selected time step size (empty steps are ignored in that figure). With average step sizes, a large variety of latent cascade structures can be considered to explain the infections. With extreme values of time steps however, the variety of possible latent structures is reduced to a single structure (a chain with short steps and a single group with long steps). This greatly reduces the freedom of the learning scheme and then, the effectiveness of the model to represent the main communication channels of the network.

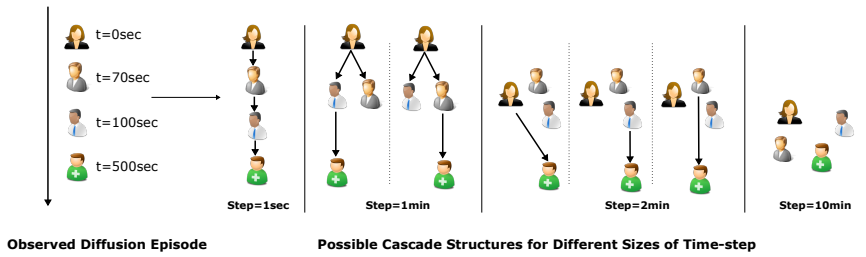

Fig. 1 Possible cascade structures for the *IC* model w.r.t. step size for a given diffusion episode (empty steps are removed for low step sizes).

On the other hand, approaches such as *NetRate* [5] or the Continuous-Time Independent Cascade model (*CTIC*) [18] include time delays in the probability distributions they define. In that way, the *NetRate* model defines decreasing probability functions w.r.t. the time argument t (the greater interval between current time and the timestamp of the infection of a given user, the lower her probabilities to diffuse to other users) [5]. The *CTIC* model learns delay

parameters additionally to diffusion abilities for each pair of users to define an asynchronous diffusion framework [18]. Such models overcome the bias induced by the definition of discrete timesteps. However, regularities over infection timestamps are very difficult to extract from real-world social networks: if regularities may exist regarding the influence of particular users on the activities of others, whose extraction already corresponds to a complex problem, observing tendencies of time delays between sparse activities of pairs of users appears quasi impossible with large social media. In these models, time delays between infections having a great impact on learned influence probabilities, prediction performances in real-world settings may suffer from this great variability of the influence delays.

We argue that including time information in the learning process usually leads to difficulties in extracting diffusion regularities. Moreover, being able to estimate infection timestamps is not essential for many applications, such as buzz prediction, opinion leaders identification or predictions tasks of content diffusion where the focus is given to final infections (who is finally infected, how many users are finally infected, etc...). Based on these two main observations, we propose to relax the problem of diffusion by considering a delay-agnostic model, which exploits infection orders instead of exact infection timestamps. Assuming that the time delay between activities of two related users follows a uniform distribution over the observation window, we consider that the probability of observing the infection of a given user depends on influences from all of its previously infected predecessors. It allows us to learn more about influence tendencies in the network than time explicit models. Note that, although it does not fully use infection timestamps information from the data to gain some generalization ability, our model can still be used to predict probabilities of infections orders and remains relevant for applications where one may be interested in which users are the most likely to be impacted by an advertisement first. Furthermore, time-delays can be learned afterward, in the ground of influence probabilities extracted by our relaxed model.

Our *Delay-Agnostic IC* model (*DAIC*) grounds in the classical *IC* model, but uses a learning process which considers that any previously infected user can explain a newly observed infection. Considering the same example of diffusion episode as in figure 1, figure 2 represents the various possible diffusion cascade structures that could explain the observed successive infections with our model. This highlights the greater freedom of our learning process, which considers each possible structure with equivalent prior probability.

Our model thus focuses on infection probabilities knowing sets of all already infected users. Therefore, our concern is to set time-independent probabilities on relationships of the graph: A diffusion probability value $\theta_{u,v}$ has to be set for each pair of users (u,v) with $I_{u,v} \in \mathscr{I}$. It corresponds to the probability that user u propagates a given content to user v before the end T of the diffusion process[3]: $\theta_{u,v} = \int_{t(u)}^{T} P_{u,v}(t)\ dt$. The influence graph can then be fully

[3] The ending time of diffusion T is arbitrarily set to the infection time-stamp $t^D(u)$ of the latest contaminated user u in the longest diffusion episode D.

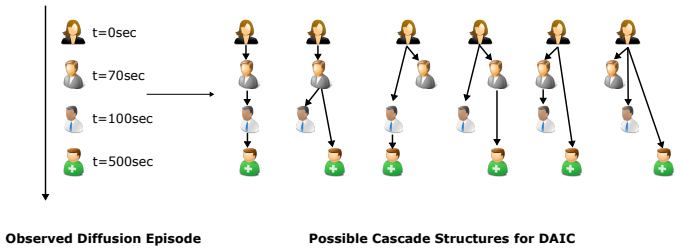

Fig. 2 Possible cascade structures for a given episode with our delay-agnostic model.

described in our model by these pairwise final transmission probabilities $\theta_{u,v}$, whose learning is described below.

2.3 Influence Learning

As we stated that time intervals between successive infections should not be considered during influence relationships learning, we focus on infections (or non-infections) of users knowing previously infected users in training diffusion episodes. In our setting, as in the classical *IC*, a newly infected user has a unique chance to infect each of her non-infected successors. However, we consider here that each of these infection events can happen at anytime in the future (before the end of the observation window T) rather than at the next time-step only. Then, in a similar way as in [20] but without restricting to influences from users whose infection time-stamp falls in a previous contiguous arbitrarily-sized time-step, we consider that the infection of each user in a diffusion episode is due to at least one transmission success from a previously infected user. Given a set of potentially influential users $I \subseteq \mathscr{U}$ (a set of previously infected users), the probability $P(v|I)$ of observing the infection of a user v knowing this set is therefore defined as:

$$P(v|I) = 1 - \prod_{u \in I \cap Preds_v} (1 - \theta_{u,v}) \tag{1}$$

Then, rather than attempting to explain all observed time-stamps of infection, our proposal is to only consider partial orders of infection during influence learning. Considering the pairwise transmission probabilities $\theta_{u,v}$ as the set of parameters θ of the model, we define $P(U_\infty^D|\theta)$ as the probability of observing:

- The infection of each user v infected in the diffusion episode D, knowing the infection configuration of all users at its time-stamp of infection $t^D(v)$;
- The non-infection of each user that does not belong to the set of infected users in the diffusion episode D, knowing all finally infected users U_∞^D in D.

Therefore, $P(U_\infty^D|\theta)$ is defined as:

$$P(U_\infty^D|\theta) = \prod_{v \in U_\infty^D} P(v|U_v^D) \prod_{v \in \bar{U}_\infty^D} (1 - P(v|U_\infty^D)) \tag{2}$$

Also, we consider the following log-likelihood $\mathscr{L}(\theta;\mathscr{D})$ of the parameters θ for all diffusion episodes from the training set $\mathscr{D}$:

$$\begin{aligned}\mathscr{L}(\theta;\mathscr{D}) &= \sum_{D\in\mathscr{D}} \log(P(U_\infty^D|\theta)) \\ &= \sum_{D\in\mathscr{D}} \sum_{v\in U_\infty^D} log(P_v^D) + \sum_{v\in\bar{U}_\infty^D} \sum_{u\in U_\infty^D\cap Preds_v} log(1-\theta_{u,v})\end{aligned} \tag{3}$$

where P_v^D is a shortcut for $P(v|U_v^D)$. This log-likelihood is however very difficult to optimize directly, due to the definition of P_v^D as given by formula 1. Nevertheless, if we knew which attempts of infection succeeded in the observed diffusion process, the optimization problem would become much more easier. Success or failures of influence attempts thus stand as latent factors of the problem. Therefore, following a similar learning methodology as described in [20], we propose to employ an Expectation-Maximization (EM) algorithm considering the following expectation function[4]:

$$\mathscr{Q}(\theta|\hat{\theta}) = \sum_{D\in\mathscr{D}} \Phi^D(\theta|\hat{\theta}) + \sum_{v\in\bar{U}_\infty^D} \sum_{u\in U_\infty^D\cap Preds_v} log(1-\theta_{u,v}) \tag{4}$$

where $\Phi^D(\theta|\hat{\theta})$ corresponds to the expected value, for a given diffusion episode D, of the first term of the log likelihood function, which stands for the log likelihood computed on infected users only. It is computed with respect to the conditional probabilities of success of diffusion between users under the current estimate of the parameters $\hat{\theta}$. Knowing that a user v is infected with an estimated probability $\hat{P}_v^D$ (which is computed via formula 1 with current estimations of transmission probabilities $\hat{\theta}$), the conditional probability $\hat{P}_{u\to v}^D$ that the diffusion from a given previously infected user $u \in Preds_v$ succeeded is given by:

$$\hat{P}_{u\to v}^D = \frac{\hat{\theta}_{u,v}}{1-\prod_{u'\in U_v^D\cap Preds_v}(1-\hat{\theta}_{u',v})} = \frac{\hat{\theta}_{u,v}}{\hat{P}_v^D} \tag{5}$$

Then, we can formulate the expectation $\Phi^D(\theta|\hat{\theta})$ as:

$$\Phi^D(\theta|\hat{\theta}) = \sum_{v\in U_\infty^D} \sum_{u\in U_v^D\cap Preds_v} \frac{\hat{\theta}_{u,v}}{\hat{P}_v^D} \log(\theta_{u,v}) + (1-\frac{\hat{\theta}_{u,v}}{\hat{P}_v^D}) \log(1-\theta_{u,v}) \tag{6}$$

Canceling the derivative of $\mathscr{Q}(\theta|\hat{\theta})$ w.r.t. parameters θ allows us to easily maximize it at each step of the EM algorithm. For each $I_{u,v} \in \mathscr{I}$, we get:

$$\theta_{u,v} = \frac{\sum_{D\in\mathscr{D}_{u,v}^+} \frac{\hat{\theta}_{u,v}}{\hat{P}_v^D}}{|\mathscr{D}_{u,v}^+| + |\mathscr{D}_{u,v}^-|} \tag{7}$$

[4] Note that the second term of formula 3 remains unchanged since this part does not depend on any latent factor and can be considered as it in the optimization process.

This update formula is similar to the one of [20] but with different definitions of positive and negative sets of diffusion episodes for a pair of user (u,v):

$$\mathscr{D}^{+}_{u,v} = \{D \in \mathscr{D} | t^D(u) < t^D(v) \wedge t^D(v) < \infty\} \quad (8)$$

$$\mathscr{D}^{-}_{u,v} = \{D \in \mathscr{D} | t^D(u) < \infty \wedge t^D(v) = \infty\} \quad (9)$$

While $\mathscr{D}^{+}_{u,v}$ corresponds to the set of positive examples of diffusion between user u and user v (diffusion episodes in which an influence can have occurred between user u and user v since they are both infected and the infection of u precedes the one of v), $\mathscr{D}^{-}_{u,v}$ contains diffusion episodes corresponding to examples of no diffusion (or negative examples of diffusion) between these two users (u is infected, v is not). Such sets definition allows our model to be more realistic by assuming influences between all ordered pairs of infected users in a diffusion episode, while avoiding difficulties induced by low time-related regularities in cascade models such as *NetRate* or *CTIC*.

2.4 Improving Robustness with Priors

In our learning model, assumptions are performed on who influenced whom in the observed diffusion episodes. This is done by considering at each step that at least one previous user infected the newly infected one and then, the probability that a diffusion attempt succeeded from user u to user v depends on all diffusion probabilities $\theta_{u',v}$ from users $u' \in Preds_v$ infected before v. This is induced for each diffusion episode D and each pair of users (u,v) by the ratio $(\hat{\theta}_{u,v}/\hat{P}^D_v)$ used in equation 6 (see previous section). While this setting appears rather realistic, it leads to biases resulting from imbalanced representations of users in the training episodes set. Indeed, it is easy to see that, employing the update formula 7, rare examples of diffusion without (or with few) counter-examples[5] in the training set may hide other positive examples on some episodes, even those corresponding to more frequent and therefore more reliable observations. To illustrate this, with $P^{D(i)}_v$ the estimation of the infection probability of v in episode D (computed using formula 1) at the i-th iteration of the learning process, let us consider the following proposition:

Proposition 1 *For every diffusion $D \in \mathscr{D}$ and every user $v \in U^D_\infty$, if it exists at least one user $u \in U^D_v \cap Preds_v$ such that $|\mathscr{D}^{-}_{u,v}| = 0$, then we have:*

$$\lim_{n \to +\infty} P^{D(n)}_v = 1$$

The demonstration of this proposition is given in appendix A. It represents a situation where some infections clearly hide others in the training set $\mathscr{D}$:

[5] In our setting, a counter-example of diffusion from user u to user v is an episode contained in $\mathscr{D}^{-}_{u,v}$ (see formula 9): an episode where u is infected but v is not.

it suffices that at least one relationship $I_{u,v}$ to any user v has no counter-example in the training set for getting the probability of the infection of v converge to 1 for each diffusion episode where u is infected before v. In that case, the infection of user u is enough to fully explain the infection of v. Looking at the update formula (eq. 7), it follows that no other relationship to v can benefit from having its source infected before v in such episodes. Such positive examples of potential influence are lost for the learning of their transmission probability. Going deeper in the analysis of such a problematic case, with $\theta_{u,v}^{(n)}$ the transmission probability from user u to user v at the i-th iteration of the learning process, the following proposition can be stated:

Proposition 2 *For every relationship $I_{u,v} \in \mathscr{I}$ such that $|\mathscr{D}_{u,v}^-| > 0$, if it exists in each $D \in \mathscr{D}_{u,v}^+$ at least one user $u' \in U_v^D \cap Preds_v$ such that $|\mathscr{D}_{u',v}^-| = 0$, then we have:*

$$\lim_{n \to +\infty} \theta_{u,v}^{(n)} = 0$$

The demonstration of this proposition is given in appendix B. It indicates that, if a pair of users (u,v) gets at least one negative example of diffusion (i.e., $\mathscr{D}_{u,v}^-$ is not empty), any other users with no counter-example of diffusion to v can make the transmission likelihood $\theta_{u,v}$ converge to 0. This can be easily deduced from the previous proposition and the update formula (eq. 7), see the appendix.

Then, users participating to a unique diffusion episode may highly perturb the learning process: all infections happening after theirs can be fully explained by transmissions from them if the corresponding relationships exist in $\mathscr{I}$. For instance, imagine a blog where a user v posted a message after u in 99 discussion flows, but missed one discussion in which u participated. Now, consider also that in each one of these 99 positive episodes, another different user, who only appears in this episode, posted a message before v. Then, although owning 99 positive examples over 100, the transmission probability $\theta_{u,v}$ converges to 0, since all the benefits that could have been extracted from these positive examples have been canceled by very rare, and therefore very poorly reliable, participations of users. Figure 3 depicts such a situation with four diffusion episodes starting from the black user. While the grey user is present in 3 over 4 episodes after the black user, the influence probability from the black user to the grey one converges to 0, since all of their positive examples of diffusion can be explained by isolated users.

While the proposition 2 presented above depicts an extreme case (while rather frequent in real datasets), that do not cover every problematic situation related to imbalanced representations of users in the training set, it is representative of over-training problems induced by the fact of considering an infection probability such as the one defined in 1. This problem can be also observed in the learning of classical *IC* as defined in [20]. It is increased here since users' participations to a diffusion episode impact the whole information

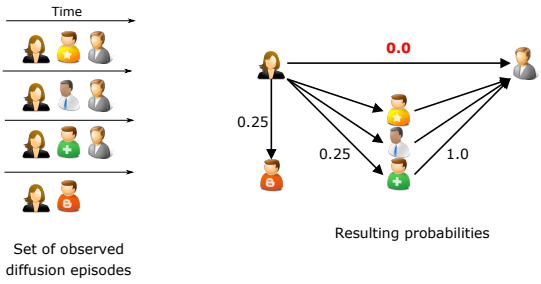

Fig. 3 Influence probabilities learned by our delay-agnostic IC from a set of four diffusion episodes. The influence probability from the black user to the grey one converges to 0, although several positive examples of diffusion have been observed between these two users.

one can extract from this episode rather than only having an impact on the corresponding time-steps as it would be the case with the classical *IC*.

To cope with this identified problem, we propose to consider prior distributions of the transmission probabilities we define, leading then the model to focus on more reliable diffusion channels. Our optimization problem thus becomes a maximum a posteriori estimation, where the estimator is given by:

$$\theta^*(\mathscr{D}) = \arg\max_{\theta} \prod_{D\in\mathscr{D}} P(U_\infty^D|\theta) \prod_{\theta_{u,v}\in\theta} f(\theta_{u,v}) \tag{10}$$

$$= \arg\max_{\theta} \mathscr{L}(\theta;\mathscr{D}) + \sum_{\theta_{u,v}\in\theta} \log(f(\theta_{u,v})) \tag{11}$$

where $f(\theta)$ stands for the prior applied to the transmission probabilities θ of the model. As various prior distribution functions could be considered, an exponential distribution appears a relevant choice since it favors sparse sets of parameters, which well fits with our task of extracting the main communication channels of the network: in proportion w.r.t. the total number of directed edges between users in the network, the set of relationships with high transmission rates is usually very sparse. With an exponential distribution function f, the maximization problem given by formula 11 can be easily simplified to the following formulation:

$$\theta^*(\mathscr{D}) = \arg\max_{\theta} \mathscr{L}(\theta;\mathscr{D}) - \lambda \sum_{\theta_{u,v}\in\theta} \theta_{u,v} \tag{12}$$

where λ corresponds to the parameter of the considered exponential distribution function. Such maximization allows us to cancel the bias mentioned above w.r.t. imbalanced user occurrences in the training set, as it enforces the model to focus on the main diffusion channels by favoring sparse parameter schemes. Following the optimization methodology detailed in the previous section, we get the following second degree polynomial to solve at each maximization step of the EM algorithm for each update of parameter $\theta_{u,v}$ according to current parameters $\hat{\theta}$:

$$\lambda\theta_{u,v}^2 - \beta\theta_{u,v} + \gamma = 0 \tag{13}$$

where $\beta = (|\mathscr{D}^-_{u,v}| + |\mathscr{D}^+_{u,v}| + \lambda)$, $\gamma = \sum_{D \in \mathscr{D}^+_{u,v}} \frac{\hat{\theta}_{u,v}}{\hat{P}^D_v}$ and whose discriminant Δ is equal to: $\beta^2 - 4\lambda\gamma$. Since $|\mathscr{D}^+_{u,v}| \geq \gamma$, we get:

$$\begin{aligned}\Delta &\geq (|\mathscr{D}^-_{u,v}| + |\mathscr{D}^+_{u,v}| + \lambda)^2 - 4\lambda|\mathscr{D}^+_{u,v}| \\ &= (|\mathscr{D}^-_{u,v}| - |\mathscr{D}^+_{u,v}| + \lambda)^2 + 4|\mathscr{D}^-_{u,v}||\mathscr{D}^+_{u,v}| \\ &\geq 0\end{aligned}$$

Then, the polynomial in formula 13 has always at least one solution:

$$\theta_{u,v} = \frac{\beta - \sqrt{\Delta}}{2\lambda} \tag{14}$$

which can be used at each maximization step of the EM algorithm, to find the estimator given by formula (12).

Proposition 3 *Solution given by formula (14) is a consistent probability lying in* $[0,1]$*, which can be used as an update rule at each maximization step of formula (12).*

Following proposition 3, whose demonstration can be found in appendix C, we use the new update formula at each maximization step of the learning process. However, while the use of a prior distribution on parameters to be learned allows us to avoid the convergence of transmission probabilities for rare users to high values, it leads to lowering the diffusion expectation of any information through the network. Therefore, we propose to end the learning process by a classical update (with formula 7), which allows us to benefit from an unbiased basis, resulting from successive updates with priors (with formula 14), while determining influence probabilities that lead to as important spreads of diffusion as observed in the training set of episodes.

3 Experiments

This section aims at evaluating the proposed model *DAIC*, by comparing it with related state of the art approaches.

3.1 Baselines

The following baselines are considered in our experiments:

– **IC**: the classic independent cascade model our works grounds in. Weights are learned as defined in [20].

- **Netrate**: as *IC*, *Netrate* [5] is a cascade model which defines influence probability distributions on the network to model information propagation. It nevertheless considers time-dependent distributions rather than defining static influence probabilities: influence weights used to parametrize probability laws are learned to fit observed infection timestamps. Note that we only report here results obtained with the *exponential* version of the *NetRate* model, as other distributions laws proposed in [5] (i.e., *power* and *rayleigh* laws) lead us to similar results.
- **CTIC**: As defined in [18], *CTIC* is a continuous-time version of the *IC* model. As *NetRate*, it uses exponential distributions to model delays of diffusion between users, but rather than a single parameter for each relationship, delays and influence factors are considered as separated parameters, which leads to more freedom w.r.t. observed diffusion tendencies. Delays and influence parameters are learned conjointly by an EM-like algorithm.

While most of the cascade approaches, such as *IC* or *CTIC*, make the assumption that the graph on which the propagation occurs is known, the social graph defined by an online social network (friends, followers, subscriptions...) is often incomplete, irrelevant [23] or unknown. Nevertheless, most of graph-based models (including all our baselines) remain valid if we consider complete graphs of the set of users. All of our experiments reported in the following are therefore obtained with complete graph structures. During the learning process however, it is possible to drastically reduce computational requirements by only considering relations that own at least one positive example of diffusion in the training set[6].

3.2 Diffusion Prediction Task

As by nature, diffusion probabilities between users are hidden in real-world data, the evaluation of the proposed model cannot be directly done by comparing inferred communication channels (or estimated probabilities in our case) with exact ones, as it is done in several studies with artificial data (see [20] for instance). Therefore, we propose to assess the performances of our proposals on real-world data by considering a related prediction task, in which the diffusion models are used to predict final infections from initial observed ones. This corresponds to the natural task of predicting the spread, over a network, of a diffusion starting from a set of source users. More specifically, the goal is to know which users are likely to be infected at the end of an observation time-window.

Defining final infection probabilities for every user of the network is rather complex with cascade models, as their iterative process requires, for computing infection probabilities at a given step, to consider every possible infection

[6] Relation u,v is considered only if there exists at least one diffusion episode in the training set where u is infected before v. With all approaches studied hereafter, relationships with no positive example would obtain a null weight anyway. They can therefore be ignored during the learning step.

distributions on the previous step, which induces an intractable complexity. Therefore, evaluations are performed on results of monte-carlo simulations of diffusion following the process of the cascade model in concern:

- *IC*: At each time-step (of the same size that was used for learning), each newly infected user attempts to contaminate each not infected one. The success of a contamination depends on the probability set for the relationship between both corresponding users. The process stops when no new contamination has been observed at a given time-step or when or the observation window is exceeded.
- *CTIC*: Simulations for *CTIC* are performed in a similar way as for *IC*, except that new infections do not occur between consecutive time-steps: for each infection success, a continuous time-delay is sampled from an exponential distribution, parametrized during the learning step for the specific relationship between users in concern.
- *NetRate*: *NetRate* discretizes the observation window in different time-steps (100 in our experiments) and, for each of them, samples infections according to the probabilities for users to be infected at this time-step knowing preceding infections and time-dependent distributions defined on the corresponding relationships.
- *DAIC*: The approach proposed in this paper, which is detached from any temporal consideration, performs diffusion simulation same manner as *IC*, but without associating timestamps to infections. What is iteratively built here is simply a set of infected users, with newly infected ones having the possibility of contaminating every other one in the network.

Results obtained from diffusion simulations are evaluated by classical recall (Rec) and precision (Prec) measures, where the recall considers the ratio of users infected in a test episode that have been retrieved as infected in the simulation and the precision renders the ratio of correct infection predictions. Finally, for each simulation, we consider a F1 evaluation measure that proposes a compromise between precision and recall:

$$F1 = \frac{2 \times Prec \times Rec}{Prec + Rec} \tag{15}$$

3.3 Experiments on Synthetic Data

In order to well understand performances of the different approaches, we first performed a preliminary set of experiments on artificial datasets with known properties.

Contrary to experiments on real world data, considering artificial data allows us to assess the ability of the models to extract correct diffusion distributions, since the probabilities of diffusion that have been used to draw the data are known. With such data, comparisons between true $\theta^*_{u,v}$ and inferred $\theta_{u,v}$ probabilities of diffusion for pairs of users (u,v) are then possible by the

mean of a given distance measure. We propose to consider a measure of the mean squared error (MSE) computed over every pair of users of the network:

$$MSE = \frac{1}{|\mathscr{U}| \times (|\mathscr{U}| - 1)} \sum_{(u,v) \in \mathscr{U}^2, u \neq v} (\theta^*_{u,v} - \theta_{u,v})^2 \tag{16}$$

This section first introduces the synthetic data used in these experiments and then presents some experimental results on these data, for both influence probabilities extraction (in term of MSE) and diffusion prediction tasks (in term of $F1$).

Synthetic Datasets Our concern in this section is to understand how behave the different approaches w.r.t. the variability over the delays between successive infections. Starting from a scale-free network of 100 users obtained from the Barabási-Albert algorithm (with each new created node connected to 2 existing ones), influence probabilities are uniformly sampled on these connections between users to obtain an influence graph that can be managed by the *IC model*. Then, we uniformly sampled source users for each diffusion episode to built (1 to 3 source users per diffusion) and performed a diffusion simulation. Note that other settings for data generation have been considered for the construction of the network (including using real-world networks) and the sampling of the diffusion episodes (including using influence probabilities resulting from real-world diffusion observations, obtained by using probability learning schemes proposed by the baseline diffusion models presented above). However, no significant difference have been observed in the results, since what differs between the models is their way of time consideration. We therefore focus on the impact of the variance of time delays on the performances of the approaches.

Following *IC*, each newly infected user attempts to contaminate all its successors in the network according to the probability set on the corresponding relationship. If the contamination attempt succeeds, a delay is chosen to determine the timestamp of the infection in concern. The delay $\delta^D_{u,v}$ is chosen for the relationship u,v and the diffusion episode D in concern:

$$\delta^D_{u,v} = 1 + \gamma_{u,v} + \xi^D_{u,v} \tag{17}$$

where $\gamma_{u,v}$ corresponds to the min delay for any diffusion from u to v and $\xi^D_{u,v}$ stands for an additional delay that can vary for this relationship over the different considered episodes D. These two delays are sampled from exponential distributions:

$$\gamma_{u,v} \sim \frac{1}{\mu} e^{-\frac{x}{\mu}} \qquad \xi^D_{u,v} \sim \frac{1}{\sigma} e^{-\frac{x}{\sigma}} \tag{18}$$

with μ the mean minimal delay for any relationship of the network and σ the mean additional delay over any relationship u,v and any diffusion episode D. While μ allows us to control the variability of the delays over the different

relationships, σ permits to manage the variability of the delays of any diffusion over the various considered episodes and then, enables the evaluation of the approaches for different temporal regularity settings. Note lastly that infections occurring outside of the observation window (i.e., with a timestamp exceeding 1000 in our experiments) are not included in the datasets.

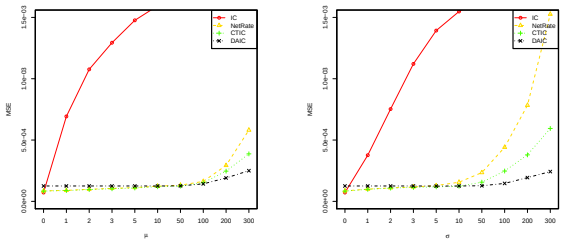

Fig. 4 MSE of the learned diffusion probabilities w.r.t. true distributions, for the experimented models on artificial diffusion data drawn with different delay parameters μ (on the left) and σ (on the right).

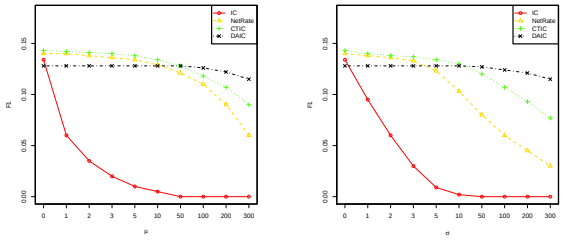

Fig. 5 F1 scores for the experimented models on artificial diffusion data drawn with different delay parameters μ (on the left) and σ (on the right).

Results Figure 4 presents MSE results for models *IC*, *NetRate*, *CTIC* and *DAIC* on artificial datasets built with different settings of infection delay sampling (see formula 17). Curves on the left plot MSE scores w.r.t. the μ parameter that controls the variation of delays between relationships. For these curves, we set $\sigma = 10^{-5}$, which leads, for every relationship, to a very stable delay over the generated diffusion episodes. Curves on the right plot MSE scores w.r.t. σ which controls the variation of delays over relationships and diffusion episodes. For these curves, we set $\mu = 10^{-5}$, which leads to minimal delays for low values of σ. Plotted results are average scores considering 10 datasets for each setting, each containing 1000 episodes for training the models and 1000 other ones for measuring the performances.

When both μ and σ tend to 0 (starting point of both figures), every delay is equal to 1. That is, infections occur between consecutive timestamps, which corresponds to the setting of the classical Independent Cascade Model. For this setting, we can indeed observe that *IC* performs rather well, as its more restrictive process allows it to obtain better results than our proposal (*DAIC*) which has to perform influence assumptions over many more relationships. It nevertheless appears that time dependent cascade models such as *CTIC* or *NetRate* perform better than *IC* in that cases, due to their better generalization abilities (classical *IC* considers very fewer relationships during training than other approaches). Our proposal *DAIC*, which considers that infections can be explained by any previous infection independently from its age, is not well fitted for this setting and therefore infers less accurate probabilities than other approaches, which favors explanations by recent previous infections. As expected, performance of *IC* however collapses when infections can occur between non-successive timestamps (see on every figure, when μ or σ increases), since such long term influences are not considered by its learning process.

From the curves on the left, it may be noticed that *CTIC* behaves better than *NetRate* w.r.t. variations of delays over the different relationships. Its independent consideration of time delays and influence rates allows it to still set good influence levels even for relationships with long delay tendencies (which cannot be done with *NetRate*). As it can be observed from the curves on the right, this also allows it to be more robust w.r.t. variations of delays on each relationship over the different diffusion episodes.

However, from both sets of curves, as values of μ and σ increase, our proposal *DAIC* appears to behave better than these two state of the art approaches. Its effectiveness level is more stable, as delay variations have, by the nature of the model, no effect on it. The increase of all error MSE scores with values of μ or σ greater than 100 may be partly explained by the fact that corresponding diffusion episodes contain less infections, as longer delays induce less infections included in the observation window, and then less positive examples of diffusion are available for learning the models. Nevertheless, another reason is that with such values, delays can cover the whole observation window and then, every observed infection may have been induced by any other previous one, independently from delays between them. Since this matches with the setting of our proposal, we can observe that our delay-agnostic learning of *IC* better resists with great variations of delays than *CTIC* and *NetRate*, which greatly suffer from their time dependent learning process in such cases. Note that, while *CTIC* is able to set quasi-uniform delay distributions when required, its process still tends to converge towards models favoring explanations by more recent infections.

Figure 5 presents F1 results obtained by diffusion simulations performed by *IC*, *NetRate*, *CTIC* and *DAIC* on the same datasets as for MSE curves. Models are also learned and experimented on two distinct sets of 1000 diffusion episodes for each dataset. It is interesting to observe the strong correlation between observations that can be done from these curves compared to those from figure 4, corresponding to MSE errors w.r.t. ground truth diffusion probabil-

ities. This validates that experimental results obtained for a task of diffusion prediction well render the accuracy of the diffusion probabilities extracted by the models: While *CTIC* is quite more robust w.r.t. variations of delays than other existing approaches, our proposal *DAIC* catches up, and then overcomes, the prediction accuracy levels of this state of the art model when values of μ and σ increase.

To summarize, while *CTIC* performs better with regular delays, our delay-agnostic proposal leads to better effectiveness results when delays between infections tend to be drawn from uniform distributions. This corresponds to what we expected to observe on well formatted artificial data. Let's see now what happens on real-world data.

3.4 Experiments on Real-World Data

3.4.1 Real-world datasets

	$\lvert\mathscr{U}\rvert$	$\lvert\mathscr{I}\rvert$	$\lvert\mathscr{D}\rvert$	$\sum_{D\in\mathscr{D}} \frac{\lvert U^D_\infty\rvert}{\lvert\mathscr{D}\rvert}$
Digg	4587	689414	20172	8.26
ICWSM	2270	4773	20027	2.21
Enron	1557	2628	1867	3.30
Twitter	4165	2267310	4815	22.54
Memetracker	30907	1298787	6724	20.21

Table 1 Some statistics about our real datasets.

Five real-world datasets are considered in our experiments:

- Digg: The *Digg* collaborative news portal allows users to post links to *stories* (articles, blog posts, videos...). Other users can then "digg" these stories. Stories appear or not on the front page of *Digg*, on the basis of the amount of "diggs" they have. We use stories as propagated content in diffusion episodes, each "digg" given by a user being considered as a user contamination. We used the Digg stream API to collect the *complete Digg* history (every single story posted, all diggs, and all comments) during a one month time window.
- ICWSM: The International AAAI Conference on Weblogs and Social Media 2009 (*ICWSM*) published a corpus containing 44 millions blog posts collected over a 1-year period [2]. Diffusion episodes are composed of sets of posts which cite a same source blog. A diffusion episode then corresponds to a set of users (authors of the corresponding posts) associated with their infection timestamps (timestamps of the posts).
- Enron: The well-known *Enron* corpus gathers emails from about 150 persons, mostly senior managers of the Enron American corporation. Various

mail addresses are often used for a same person in this corpus. For simplicity, we consider different addresses as different users in the following. The corpus initially contains a total of about 500000 messages. From these, we define diffusion episodes as proposed in [11], by considering sequences of messages that form a conversation about a particular topic. These conversations are extracted by selecting messages that contain at least two common words and whose sender corresponds to a recipient of a previous message in the sequence.

- Twitter: This corpus has been built by collecting messages from the streaming API of the online social network Twitter. First, we collected 5000 users that posted tweets with words "Obama" or "Romney". Then, we followed all their posts during 2 weeks of the US presidential elections (the two weeks before the election day). Diffusion episodes are formed by considering tweets containing the same hashtags. Diffusion episodes with less than 5 users are finally removed to only keep significantly propagated hashtags.
- Memetracker: The *Memetracker* corpus, described in [13], contains diffusion episodes of short phrases (memes) extracted from news websites and blogs collected during the 2008 US presidential campaign.

Table 1 gives some statistics about the datasets. In this table, $|\mathscr{U}|$, $|\mathscr{I}|$ and $|\mathscr{D}|$ respectively correspond to the number of users, the number of relationships and the number of diffusion episodes. The last column corresponds to the average episode size (number of infections). Note that episodes of Twitter and Memetracker corpora contain much more users than those of others.

3.4.2 Results

	Digg	ICWSM	Enron	Twitter	Memetracker
IC	0.036	0.097	0.033	0.013	0.012
NetRate	0.102	0.358	0.105	0.027	0.048
CTIC	0.119	0.482	0.132	0.032	0.061
$DAIC_0$	**0.127**	**0.665**	**0.162**	0.026	**0.073**
$DAIC_5$	**0.128**	**0.665**	**0.164**	0.035	**0.087**
$DAIC_{10}$	**0.127**	**0.665**	**0.164**	**0.044**	**0.082**

Table 2 F1 Results. Scores in bold are significantly greater than *CTIC* ones (99% Student-t test).

Table 2 reports F1 results obtained on real datasets with the different approaches. Each result corresponds to an average score obtained over a set of 1000 diffusion episodes that were not used for learning. We note $DAIC_\lambda$ our approach of delay-agnostic *IC*, with λ the parameter of the exponential prior distribution used in the update formula 14.

In order to learn the parameters of *IC*, a step-size has to be chosen to fit with sequences of infections observed in the datasets. This time-step size is difficult to determine with real-world datasets: if too short, it leads to a lot

of empty infection ties, if too long, most users are gathered in the same time-steps. In both cases, this results in a very low amount of positive examples of diffusion. In our experiments, we set the step-size of *IC* for each dataset as the average delay between two consecutive infections in the training set. This heuristic usually allows one to obtain a reasonable amount of positive examples of diffusion *IC* can ground in. Nevertheless, we observe from table 2 that a classical learning of *IC* presents important difficulties in determining correct infection probabilities in real-world settings, the F1 scores it obtained being greatly lower than those of every other approach for each dataset. It even tends to scores close to 0 for Twitter and Memetracker datasets, which means that for these datasets nearly not any correct infection could be predicted, partly due to the impossibility to find a step-size that fits well for a sufficient amount of training diffusion episodes (no regularities in infection time delays).

Except on the Twitter dataset, our proposal of delay-agnostic learning obtains significantly better results than other approaches. It confirms our claim that real-world time delays of infection should be considered to follow an uniform distribution, an infection at the end of an episode being as likely resulting from an influence by an early infected user as by a recent one. Whereas models such as *CTIC* could be regarded as more realistic, since favoring short delay transmissions, such a setting usually leads to over-fitted distributions, as observed delays in the training set rarely hold for prediction. Moreover, rare users have a strong negative impact on the learned probabilities, as they induce unconstrained infection explanations. While our proposal cannot be used to predict time-stamps of infection (which is, from our point of view, quasi-impossible in general settings with real-world data), it leads to a better identification of the main channels of influence of the network. By only considering partial orders of infections during the learning process rather than attempting to explain full diffusion episodes with exact infection time-stamps, it focuses on who infected whom by emitting diffusion assumptions without favoring any source according to its infection time.

On the Twitter dataset however, it appears that the benefit resulting from this possibility for any infection to be explained by any previously infected user is greatly limited by the unbalanced observations bias mentioned in section 2.4. In this corpus indeed, a lot of diffusion episodes contain very rare users (some of them participating only once in the training set), which induces a loss of generalization ability of the model. Using an exponential prior on transmission probabilities, as proposed in the update formula 14, allows us to cope with this bias and to obtain good results despite great disparities in user's infection frequencies. On datasets with long diffusion episodes, such as the Twitter and Memetracker corpora, considering an exponential prior on the preliminary steps of the learning process (as described in section 2.4) allows one to significantly improve the prediction accuracy. On such datasets with important spreads of diffusion, the observation of infections of rare users is more likely (which induce some noise for the learning process). Our regularization proposal appears to greatly reduce their impact on the prediction accuracy performances. Note at last that the optimal regularization parameter λ to use

in the exponential prior distribution may vary over each dataset: for instance, best performances are obtained on Twitter with $\lambda = 10$, while on Memetracker $\lambda = 5$ performs better. It can nevertheless be easily tuned by a cross-validation process, by selecting the λ value that allows the best generalization ability on a validation set of diffusion episodes.

4 Related work

The recent development of online social networks enabled researchers to suggest methods to explain observations of diffusion across networks. Most of the proposed iterative models ground in the two fundamental models Independent Cascade (IC) [4] and Linear Threshold (LT) [7]. Both are modeling a user-to-user contamination process : while IC models the spread of diffusion as cascades of infections over the network, LT determines infections of users according to thresholds of the influence pressure incoming from the neighborhood of each user. We focus in this paper on IC-like approaches, which appear better fitted to reproduce realistic temporal diffusion dynamics. While parameters of these models (transmission probabilities) initially needed to be set manually, Gruhl et al. defined in [8] a first attempt to automatically learn them. A few years later, [20] proposed the learning methodology we ground in here, which appeared to be an improvement of the one of [8], since it replaces the former "exactly one influencer" assumption by a more realistic "at least one influencer" one.

Thanks to its simplicity and its ability to explain diffusion data, at least artificial ones with regular timestamps, IC has served as a baseline for a large amount of studies in the last decade. It has also been the basis of a lot of approaches, that proposed extensions for improving its effectiveness or for including richer information about the context of the modeled diffusion. [21], [25], [9] or [12] are instances of extensions including user profiles and information content to extract diffusion probabilities. NetInf [6] and then Connie [16] use greedy algorithms to find subsets of links between users that maximize the likelihood of observed diffusions under IC-like diffusion hypothesis. As discussed above, various extensions have also addressed temporal issues, by proposing models that deal with delays between observed infections, such as *CTIC* [18] or *NetRate* [5].

Nevertheless, as widely discussed in this paper, temporal regularities are difficult to observe and attempting to capture them may lead to lower effectiveness for extracting main influence channels of the network. Then, recently various works turned away from such iterative models, making use of classical statistical learning instead of grounding in graph-based approaches. For instance, [22] performs extrapolations grounded in relations between the number of infected users after a short period of time and after a longer one to predict the final volume of infections. [26] infers the volume of diffusion based on infection timestamps of specifically selected subsets of users. [24] proposed a logistic model that considers the density of influenced users at a given dis-

tance of the source after a given time of diffusion. [1] followed a similar idea by projecting the network in a continuous space where information diffusion can be modeled as a heat diffusion process.

Our proposal leads to reconsider the use of cascade models for diffusion predictions on real world networks, since using a temporally relaxed framework while keeping the finer-grained modelization of the cascade models. Note that a close "untemporal" version of IC has also been considered in [15], but in a different context and without experimenting its benefits for influence extraction from real-world social data. We also defined a useful extension to cope with biases related to the usual presence of infrequent users in the training diffusion episodes.

5 Conclusion

In this paper, our contribution is twofold:

- We proposed to use a relaxed learning scheme for the well-known Independent Cascade model, whose parameters are learned by considering partial contamination orders rather than exact observed infection time-stamps. This shows better performances for the prediction of the spread of diffusion on real social networks than greatly more complex time-dependent approaches.
- We introduced a regularization mechanism for *IC* (that can be applied as well with the classical learning scheme as with our delay-agnostic version), that leads to more robust models with great effectiveness improvements on large social networks.

This work enables to reconsider cascade models, and more generally iterative approaches, that lead to finer-grained diffusion explanations and simulations than static models that recently emerged to overcome difficulties of time consideration. Promising effectiveness results obtained with delay-agnostic *IC* let us expect various further developments of the proposed approach. For instance, we are currently working on an embedded version of our delay-agnostic *IC*, which is expected to benefit from geometric constraints related to continuous projection spaces to better capture influence regularities in the networks. Furthermore, as the nature of the propagated information may have a great impact on its spread of diffusion, we are also currently considering mixtures of delay-agnostic IC models that depend on the diffused content.

Appendix

A Proof of Proposition 1

Let us denote $\theta_{u,v}^{(i)}$ the transmission probability from user u to user v at the i-th iteration of the learning process. Let also denote $P_v^{D^{(i)}}$ the estimation of the infection probability of v in

the episode D (computed using formula 1 using current transmission probabilities) at the i-th iteration of the learning process.

First, with $A_{u,v} = \frac{|D^+_{u,v}|}{|D^+_{u,v}|+|D^-_{u,v}|}$, let us consider the following lemma:

Lemma 1

$$\forall i \in \mathbb{N}, \forall I_{u,v} \in \mathscr{I} : \left(\theta^{(i)}_{u,v} \leq A_{u,v}\right)$$

Proof. Lemma 1 can be easily deduced from the update formula applied at each step of the learning process (eq. 7), since we know from (eq. 1) that $\frac{\theta^{(i)}_{u,v}}{P^{D^{(i)}}_v} \leq 1$ for all $I_{u,v} \in \mathscr{I}$ at every iteration $i > 0$ of the process. Note that, without loss of generality, for getting the lemma valid for $i = 0$, we assume that the probabilities θ are all initialized such that for all $I_{u,v} \in \mathscr{I} : \theta^{(0)}_{u,v} \in]0, A_{u,v}[$. □

Let's now consider the following lemma:

Lemma 2

$$\forall I_{u,v} \in \mathscr{I} : (|\mathscr{D}^-_{u,v}| = 0 \implies \forall i \in \mathbb{N} : (\theta^{(i+1)}_{u,v} \geq \theta^{(i)}_{u,v}))$$

Proof. If $|\mathscr{D}^-_{u,v}| = 0$, we get, from formula 7:

$$\frac{\theta^{(i+1)}_{u,v}}{\theta^{(i)}_{u,v}} = \frac{1}{|\mathscr{D}^+_{u,v}|} \sum_{D \in \mathscr{D}^+_{u,v}} \frac{1}{P^{D^{(i)}}_j} \geq \frac{1}{|\mathscr{D}^+_{u,v}|} \sum_{D \in \mathscr{D}^+_{u,v}} 1 = 1$$

where we used the fact that $P^{D^{(i)}}_j$ is included in $]0;1[$.

□

For simplicity, let us now state $I^D_v = (U^D_v \cap Preds_v)$. For every episode $D \in \mathscr{D}$ and every user $v \in U^D_\infty$, we have at any iteration i of the process:

$$\begin{aligned}
P^{D^{(i)}}_v &= 1 - \prod_{u \in I^D_v} (1 - \theta^{(i)}_{u,v}) \\
&= 1 - \prod_{u \in I^D_v, |\mathscr{D}^-_{u,v}| > 0} (1 - \theta^{(i)}_{u,v}) \prod_{u \in I^D_v, |\mathscr{D}^-_{u,v}| = 0} (1 - \theta^{(i)}_{u,v}) \\
&\leq 1 - \prod_{u \in I^D_v, |\mathscr{D}^-_{u,v}| > 0} (1 - A_{u,v}) \prod_{u \in I^D_v, |\mathscr{D}^-_{u,v}| = 0} (1 - \theta^{(i)}_{u,v})
\end{aligned}$$

Let state $B^D_v = \prod_{u \in I^D_v, |\mathscr{D}^-_{u,v}| > 0} (1 - A_{u,v})$. Note that B^D_v is a constant over the whole learning process. Now, let's consider the case of the proposition, where it exists at least one user $u \in I^D_j$ such that $|\mathscr{D}^-_{u,v}| = 0$. In that case, we can rewrite the inequality as :

$$\begin{aligned}
P^{D^{(i)}}_v &\leq 1 - B^D_v (1 - \theta^{(i)}_{u,v}) \prod_{u' \in I^D_v \setminus \{u\}, |\mathscr{D}^-_{u',v}| = 0} (1 - \theta^{(i)}_{u',v}) \\
&\leq 1 - B^D_v (1 - \theta^{(i)}_{u,v}) (1 - \max_{u' \in I^D_v \setminus \{u\}, |\mathscr{D}^-_{u',v}| = 0} \theta^{(i)}_{u',v})^{|\{u' \in I^D_v \setminus \{u\}, |\mathscr{D}^-_{u',v}| = 0\}|} \qquad (19)
\end{aligned}$$

Now, let us consider the sequence V defined as:

$$V_n = (1 - \max_{u' \in I^D_v \setminus \{u\}, |\mathscr{D}^-_{u',v}| = 0} \theta^{(n)}_{u',v})^{|\{u' \in I^D_v \setminus \{u\}, |\mathscr{D}^-_{u',v}| = 0\}|}$$

From lemma 2, we know that V is decreasing, since any component of the max function does not own any counter-example in the training set. Moreover, This sequence is lower-bounded by 0. Then, V converges towards its fixed point, which we denote as l. From this, two possibilities: either l equals 0 or is strictly greater than 0.

If $l = 0$, then we know that:

$$\lim_{n\to\infty} \max_{u' \in I_v^D \setminus \{u\}, |\mathscr{D}^-_{u',v}|=0} \theta^{(n)}_{u',v} = 1$$

Now, the formula 1 leads to know that, at every iteration i, $\forall u' \in I_v^D : P_v^{D^{(i)}} \geq \theta^{(i)}_{u',v}$. Therefore, at every iteration i, we have: $P_v^{D^{(i)}} \geq \max_{u' \in I_v^D \setminus \{u\}, |\mathscr{D}^-_{u',v}|=0} \theta^{(i)}_{u',v}$. Since we know that $P_v^{D^{(i)}}$ is also upper-bounded by 1 at every iteration i, we can state that, in that case, $\lim_{n\to\infty} P_v^{D^{(n)}} = 1$.

Else, we have at every iteration i:

$$(1 - \max_{u' \in I_v^D \setminus \{u\}, |\mathscr{D}^-_{u',v}|=0} \theta^{(i)}_{u',v})^{|\{u' \in I_v^D \setminus \{u\}, |\mathscr{D}^-_{u',v}|=0\}|} \geq l$$

Plugging this in inequality 19, we get for every i:

$$P_v^{D^{(i)}} \leq 1 - lB_j^D(1 - \theta^{(i)}_{u,v}) \leq 1 - \lambda + \lambda\,\theta^{(i)}_{u,v}$$

with $\lambda = lB_v^D$. Then, we can rewrite the update formula 7 as:

$$\theta^{(i+1)}_{u,v} = \frac{\sum_{D' \in \mathscr{D}^+_{u,v} \setminus D} \frac{\theta^{(i)}_{u,v}}{P_v^{D'(i)}} + \frac{\theta^{(i)}_{u,v}}{P_v^{D^{(n)}}}}{|\mathscr{D}^+_{u,v}|} \geq \frac{(|\mathscr{D}^+_{u,v}| - 1)\theta^{(i)}_{u,v} + \frac{\theta^{(i)}_{u,v}}{1-\lambda+\lambda\,\theta_{u,v}}}{|\mathscr{D}^+_{u,v}|} \tag{20}$$

Let us consider now the sequence W such that:

$$\begin{cases} W_0 = \theta^{(0)}_{u,v} \\ W_{n+1} = \frac{(|\mathscr{D}^+_{u,v}|-1)W_n + \frac{W_n}{1-\lambda+\lambda W_n}}{|\mathscr{D}^+_{u,v}|} \end{cases}$$

Then, since W takes its values in $]0;1[$, and that λ is also in $]0;1[$, we can state that:

$$\frac{W_{n+1}}{W_n} = \frac{|\mathscr{D}^+_{u,v}| - 1 + \frac{1}{1-\lambda+\lambda W_n}}{|\mathscr{D}^+_{u,v}|} > 1$$

The sequence is thus strictly increasing. Since it is upper bounded by its fixed point 1, we know that it converges to 1. Now, since we know that, from inequality 20, $\forall n : \theta^{(n)}_{u,v} \geq W_n$, we can get that $\lim_{n\to\infty} \theta^{(n)}_{u,v} = 1$. This concludes the proof since therefore: $\lim_{n\to\infty} P_v^{D^{(n)}} = 1$.

B Proof of Proposition 2

If, for a given relationship $I_{u,v} \in \mathscr{I}$ such that $|\mathscr{D}^-_{u,v}| > 0$, it exists in each $D \in \mathscr{D}^+_{u,v}$ at least one user $u' \in U_v^D \cap Preds_v$ such that $|\mathscr{D}^-_{u',v}| = 0$, we can deduce from proposition 1 that:

$$\forall D \in \mathscr{D}^+_{u,v} : \lim_{n\to+\infty} P_v^{D^{(n)}} = 1$$

In that case, we can state that, after a given iteration m, it exists a value $x \in]A_{u,v};1[$ such that $\forall D \in \mathscr{D}^+_{u,v} : P_v^{D^{(n)}} > x$. Then, we know that: $\forall n > m, \theta^{(n+1)}_{u,v} < \theta^{(n)}_{u,v} \frac{A_{u,v}}{x} = \gamma\theta^{(n)}_{u,v}$, with $\gamma = \frac{A_{u,v}}{x}$. Note that $\gamma \in]0;1[$ since $x > A_{u,v}$. Let us consider now the following sequence V:

$$\begin{cases} V_0 = \theta^{(0)}_{u,v} \\ V_{n+1} = \gamma V_n \end{cases}$$

This sequence converges to its unique fixed point 0 since $\gamma \in]0;1[$. Since we know that: $\forall n > m, \theta^{(n)}_{u,v} \leq V_n$ and that $\theta^{(n)}_{u,v}$ is lower bounded by 0, then we get: $\lim_{n\to+\infty} \theta^{(n)}_{u,v} = 0$.

C Proof of Proposition 3

Proving that the solution given by (14), denoted hereafter $\theta^*_{u,v}$, is non-negative is straightforward. Inequality $\theta^*_{u,v} \geq 0$ can indeed be transformed into $\beta \geq \sqrt{\Delta}$ whose both sides are non-negative terms and which can thus be verified by considering its square: as $\Delta - \beta^2 = -4\lambda\gamma \leq 0$, $\beta^2 \geq \Delta$ is always true.

Proving that $\theta^*_{u,v} \leq 1$ requires showing that $\beta - \sqrt{\Delta} \leq 2\lambda$, which is equivalent to $\beta - 2\lambda \leq \sqrt{\Delta}$. If $\lambda \geq (|\mathscr{D}^-_{u,v}| + |\mathscr{D}^+_{u,v}|)$, the verification of the latter is direct since in that case $\beta - 2\lambda \leq 0$ (and we know that $\sqrt{\Delta} \geq 0$). In the opposite case, both sides of the inequality are non-negative. It is then possible to consider the square of the inequality: $(\beta - 2\lambda)^2 \leq \Delta$ is equivalent to $|\mathscr{D}^-_{u,v}| + |\mathscr{D}^+_{u,v}| - \gamma \geq 0$, that is always true since we know that $|\mathscr{D}^+_{u,v}| \geq \gamma$. Then, $\theta^*_{u,v}$ always lies in $[0,1]$.

Proving that the solution given by (14) can be used as an update rule at each maximization step for solving the estimator of formula (12) implies to show that it maximizes, for any pair (u,v), the quantity $Q = \mathscr{Q}(\theta|\hat{\theta}) - \lambda \sum_{\theta_{u,v} \in \theta} \theta_{u,v}$. Since we already know that $\theta^*_{u,v}$ corresponds to one of the two possible solutions of the cancellation of the derivative of Q from equation (13), it suffices to show that it corresponds to a maximum. This can be easily verified by considering the second derivative of Q w.r.t. $\theta_{u,v}$, which equals:

$$\frac{\partial Q}{\partial \theta^2_{u,v}} = -\sum_{D \in \mathscr{D}^+_{u,v}} \left(\frac{\hat{\theta}^D_{u \to v}}{\theta^2_{u,v}} + \frac{(1 - \hat{\theta}^D_{u \to v})}{(1 - \theta_{u,v})^2} \right) - \sum_{D \in \mathscr{D}^-_{u,v}} \frac{1}{(1 - \theta_{u,v})^2}$$

where $\hat{\theta}^D_{u \to v}$ is a shortcut for $\frac{\hat{\theta}_{u,v}}{\hat{P}^D_v}$. From this formulation, it is easy to see that the second derivative of Q w.r.t. $\theta_{u,v}$ is always negative on $]0;1[$, which concludes the proof: taking $\theta^*_{u,v}$ as an update of $\theta_{u,v}$ allows us to maximize Q at each step of the EM algorithm.

Acknowledgments

This work has been partially supported by the REQUEST project (projet Investissement d'avenir, 2014-2017) and the project ARESOS from the CNRS program MASTODONS.

References

1. Bourigault, S., Lagnier, C., Lamprier, S., Denoyer, L., Gallinari, P.: Learning social network embeddings for predicting information diffusion. In: B. Carterette, F. Diaz, C. Castillo, D. Metzler (eds.) WSDM, pp. 393–402. ACM (2014)
2. Burton, K., Java, A., Soboroff, I.: The icwsm 2009 spinn3r dataset. In: Proceedings of the Third Annual Conference on Weblogs and Social Media (2009)
3. Du, N., Song, L., Gomez-Rodriguez, M., Zha, H.: Scalable influence estimation in continuous-time diffusion networks. In: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (eds.) Advances in Neural Information Processing Systems 26, pp. 3147–3155. Curran Associates, Inc. (2013)
4. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing letters **12**(3), 211–223 (2001)
5. Gomez-Rodriguez, M., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), ICML '11, pp. 561–568. ACM (2011)
6. Gomez Rodriguez, M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10. ACM, New York, NY, USA (2010). DOI 10.1145/1835804.1835933. URL http://doi.acm.org/10.1145/1835804.1835933

7. Granovetter, M.: Threshold Models of Collective Behavior. American Journal of Sociology **83**(6), 1420–1143 (1978)
8. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. In: Proceedings of the 13th International Conference on World Wide Web, WWW '04, pp. 491–501. ACM, New York, NY, USA (2004)
9. Guille, A., Hacid, H.: A predictive model for the temporal dynamics of information diffusion in online social networks. In: Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion. ACM (2012)
10. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03, pp. 137–146. ACM (2003)
11. Klimt, B., Yang, Y.: Introducing the Enron corpus. In: First Conference on Email and Anti-Spam (CEAS) (2004)
12. Lagnier, C., Denoyer, L., Gaussier, E., Gallinari, P.: Predicting information diffusion in social networks using content and user's profiles. In: European Conference on Information Retrieval, ECIR '13 (2013)
13. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09, pp. 497–506. ACM, New York, NY, USA (2009). DOI 10.1145/1557019.1557077
14. Ma, H., Yang, H., Lyu, M.R., King, I.: Mining social networks using heat diffusion processes for marketing candidates selection. In: Proceedings of the 17th ACM conference on Information and knowledge management, CIKM '08, pp. 233–242. ACM, New York, NY, USA (2008). DOI 10.1145/1458082.1458115. URL http://doi.acm.org/10.1145/1458082.1458115
15. Mathioudakis, M., Bonchi, F., Castillo, C., Gionis, A., Ukkonen, A.: Sparsification of influence networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, pp. 529–537. ACM, New York, NY, USA (2011). DOI 10.1145/2020408.2020492. URL http://doi.acm.org/10.1145/2020408.2020492
16. Myers, S.A., Leskovec, J.: On the convexity of latent social network inference. CoRR **abs/1010.5504** (2010)
17. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: Proceedings of the 20th international conference on World wide web, pp. 695–704. ACM (2011)
18. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning, ACML '09, pp. 322–337. Springer-Verlag, Berlin, Heidelberg (2009)
19. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Generative models of information diffusion with asynchronous timedelay. Journal of Machine Learning Research - Proceedings Track **13**, 193–208 (2010)
20. Saito, K., Nakano, R., Kimura, M.: Prediction of information diffusion probabilities for independent cascade model. In: Proceedings of the 12th international conference on Knowledge-Based Intelligent Information and Engineering Systems, Part III, KES '08, pp. 67–75. Springer-Verlag (2008)
21. Saito, K., Ohara, K., Yamagishi, Y., Kimura, M., Motoda, H.: Learning diffusion probability based on node attributes in social networks. In: M. Kryszkiewicz, H. Rybinski, A. Skowron, Z.W. Ras (eds.) ISMIS, *Lecture Notes in Computer Science*, vol. 6804, pp. 153–162. Springer (2011)
22. Szabo, G., Huberman, B.A.: Predicting the popularity of online content. Communications of the ACM **53**(8), 80–88 (2010)
23. Ver Steeg, G., Galstyan, A.: Information-theoretic measures of influence based on content dynamics. In: Proceedings of the sixth ACM international conference on Web search and data mining, WSDM '13, pp. 3–12. ACM, New York, NY, USA (2013). DOI 10.1145/2433396.2433400

24. Wang, F., Wang, H., Xu, K.: Diffusive logistic model towards predicting information diffusion in online social networks. In: Proceedings of the 2012 32nd International Conference on Distributed Computing Systems Workshops, ICDCSW '12, pp. 133–139. IEEE Computer Society (2012)
25. Wang, L., Ermon, S., Hopcroft, J.E.: Feature-enhanced probabilistic models for diffusion network inference. In: Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II, ECML PKDD'12, pp. 499–514. Springer-Verlag (2012)
26. Yang, J., Leskovec, J.: Modeling information diffusion in implicit networks. In: Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM '10, pp. 599–608. IEEE Computer Society, Washington, DC, USA (2010). DOI 10.1109/ICDM.2010.22