2012 IEEE 12th International Conference on Data Mining

Topic-aware Social Influence Propagation Models

Nicola Barbieri
Yahoo! Research Barcelona, Spain
barbieri@yahoo-inc.com

Francesco Bonchi
Yahoo! Research Barcelona, Spain
bonchi@yahoo-inc.com

Giuseppe Manco
ICAR-CNR, Rende, Italy
manco@icar.cnr.it

***Abstract*—We study social influence from a topic modeling perspective. We introduce novel topic-aware influence-driven propagation models that experimentally result to be more accurate in describing real-world cascades than the standard propagation models studied in the literature. In particular, we first propose simple topic-aware extensions of the well-known Independent Cascade and Linear Threshold models. Next, we propose a different approach explicitly modeling authoritativeness, influence and relevance under a topic-aware perspective. We devise methods to learn the parameters of the models from a dataset of past propagations. Our experimentation confirms the high accuracy of the proposed models and learning schemes.**

I. Introduction

Social influence and the phenomenon of influence-driven propagations in social networks have received tremendous attention in the last years. One of the key computational problems in this area is the identification of a set of influential users, which are more likely to produce large influence-driven cascades: these are the users that should be "targeted" by a viral marketing campaign. This problem has received a good deal of attention by the data mining research community in the last decade [1], but quite surprisingly, the *characteristics of the item being the subject of the viral marketing campaign has been left out of the picture.*

Kempe *et al.* [2] formalize the *influence maximization* problem for a generic item: for a given budget k, find k "seed" nodes in the network, such that by activating them we can maximize the expected number of nodes that eventually get activated, according to a chosen *propagation model*, that governs how influence diffuses or propagates through the network. Kempe *et al.* [2] mainly focus on two propagation models – the *Independent Cascade* (IC) and the *Linear Threshold* (LT) models. Following this seminal work, a substantial research effort has been dedicated to develop algorithms for influence maximization under these two propagation models (see Sec. II). However, these propagation models suffer various limitations when it comes to model real-world cascades: e.g., the discrete treatment of time and the very large number of parameters. The latter is a serious issue both for efficiency and scalability, but more importantly, for the risk of overfitting.

In this paper we start from the observations that (i) users have different interests, (ii) items have different characteristics and (iii) similar items are likely to interest the same users. Thus we take a topic-modeling perspective to jointly learn items characteristics, users' interests and social influence, resulting in new propagation models that experimentally are proven to be more accurate in describing real-world cascades.

More in details our contributions are as follows:

- We extend the classic IC and LT models to be topic-aware. The propagation models we obtain are dubbed *Topic-aware Independent Cascade* (TIC) model and *Topic-aware Linear Threshold* (TLT) model. We show that the *expected spread* remains submodular for both models, thus the simple greedy algorithm provides a $(1 - 1/e - \phi)$-approximation of the optimal solution.
- We devise an *expectation maximization* (EM) approach for estimating the parameters of the TIC model.
- Starting from a discussion on the limits of the TIC and TLT models, we introduce a new influence propagation model, dubbed AIR (*Authoritativeness-Interest-Relevance*). Instead of considering user-to-user influence, the proposed model focuses on user authoritativeness and interests in a topic, leading to a drastic reduction of the number of parameters of the model.
- We devise a *generalized expectation maximization* (GEM) approach to learn the parameters that maximize the likelihood for the AIR model.
- Our experiments on real-world social networks show that topic-aware influence propagation models outperform the traditional "topic-blind" IC model in predicting adoption of a specific item, thus in modeling real-world cascades.
- The benefits of keeping in consideration the characteristics of the item being propagated, are confirmed by our experiments on influence maximization: topic-aware methods exhibit a consistent gain over state-of-the-art approach that just considers a generic item, ignoring its characteristics.

Although topic-wise social influence has been studied before, to the best of our knowledge we are the first to study it within the context of *viral marketing* and the *influence maximization* problem, as discussed in the next section.

In Sec. III we introduce the TIC and TLT models, while Sec. IV is devoted to the AIR model. Sec. V reports our experimental analysis, while Sec. VI discusses future work.

II. Background and Related Work

In this section, we provide the needed background for the paper while discussing the most relevant related work.

Influence maximization. Suppose we are given a social network, that is a directed graph whose nodes are users and arcs represent social relations among the users. Suppose we are also given the estimates of reciprocal influence between

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.122

individuals connected in the network, that is a weight (or probability) $p_{v,u}$ associated top each arc (v, u).

As said in the previous section, a basic computational problem is that of selecting the set of initial users that are more likely to influence the largest number of users in the social network. The first algorithmic treatment of the problem was provided by Domingos and Richardson [3], [4], who modeled the diffusion process in terms of Markov random fields, and proposed heuristic solutions to the problem.

Later, Kempe *et al.* [2] studied influence maximization as a discrete optimization problem focusing on two fundamental propagation models, named *Independent Cascade Model* (IC) and *Linear Threshold Model* (LT). In both these models, at a given timestamp, each node is either active (an adopter of the innovation, or a customer which already purchased the product) or inactive, and each node's tendency to become active increases monotonically as more of its neighbors become active. An active node never becomes inactive again.

In the IC model, when a node v first becomes active, say at time t, it is considered contagious. It has one chance of influencing each inactive neighbor u with probability $p_{v,u}$, independently of the history thus far. If the tentative succeeds, u becomes active at time $t + 1$.

In the LT model, each node u is influenced by each neighbor v according to a weight $p_{v,u}$, such that the sum of incoming weights to u is no more than 1. Each node u chooses a threshold θ_u uniformly at random from $[0, 1]$. At any timestamp t, if the total weight from the active neighbors of an inactive node u is at least θ_u, then u becomes active at timestamp $t + 1$. In both the models, the process repeats until no new node becomes active.

Given a propagation model m (e.g., IC or LT) and a seed set $S \subseteq V$, the expected number of active nodes at the end of the process is denoted by $\sigma_m(S)$. The *influence maximization problem* requires to find the set $S \subseteq V$, $|S| = k$, such that $\sigma_m(S)$ is maximum.

Under both the IC and LT propagation models, the problem is $\mathbf{NP}$-hard [2]. Kempe *et al.*, however, show that the function $\sigma_m(S)$ is *monotone* (i.e., $\sigma_m(S) \leq \sigma_m(T)$ whenever $S \subseteq T$) and *submodular* (i.e., $\sigma_m(S \cup \{w\}) - \sigma_m(S) \geq \sigma_m(T \cup \{w\}) - \sigma_m(T)$ whenever $S \subseteq T$). When equipped with such properties, the simple greedy algorithm that at each iteration greedily extends the set of seeds with the node providing the largest marginal gain, produces a solution with provable approximation guarantee $(1 - 1/e)$ [5]. Though simple, the greedy algorithm is computationally prohibitive, since the step of selecting the node providing the largest marginal gain is $\#\mathbf{P}$-hard under both the IC and the LT model. In their paper, Kempe *et al.* run Monte Carlo simulations for sufficiently many times to obtain an accurate estimate of the expected spread. In particular, they show that for any $\phi > 0$, there is a $\delta > 0$ such that by using $(1 + \delta)$-approximate values of the expected spread, we obtain a $(1 - 1/e - \phi)$-approximation for the influence maximization problem. However, running many propagation simulations is extremely costly on very large real-world social networks. Therefore, following [2], considerable effort has been devoted to develop methods for improving the efficiency of influence maximization [6], [7], [8], [12].

The approaches discussed above, assume a weighted social graph as input and do not address *how* the link influence weights (or probabilities) can be obtained; [10], [11], [12], [13] instead focus on the latter problem and propose specific solutions. Saito *et al.* [10] for example, study how to learn the probabilities for the IC model from a set of past propagations. They formalize this as a likelihood maximization problem and then apply the Expectation Maximization (EM) algorithm to solve it. We will extend this contribution to deal with topic-wise influence in Section III-A.

Goyal *et al.* [12] also study the problem of learning influence probabilities but under a different model, i.e., an instance of the General Threshold Model. They extend this model by introducing temporal decay, as well as factors such as the influenceability of a specific user, and influence-proneness of a certain action. They also show that their methods can be used to predict *whether* a user will perform an action and *when*.

Topic modeling. The key idea at the basis of topic modeling, is to introduce an hidden variable Z for each co-occurrence user-item. The hidden variable can range among K states. Each topic (i.e., state of the latent variable) represents an abstract interest/pattern and intuitively models the underlying cause for each data observation. Among the probabilistic approaches for topic modeling, besides mixture models that are widely investigated in the literature [14], *Probabilistic Latent Semantic Analysis* (pLSA) [15] is considered the progenitor of a wide range of recent approaches, which include e.g. the popular *Latent Dirichlet Allocation* (LDA) [16].

Topic-aware influence analysis. Regardless the fact that users authoritativeness, expertise, trust and influence are evidently topic-dependent, the research on social influence has surprisingly largely overlooked this aspect. To the best of our knowledge only few papers have looked at social influence from the topics perspective [11], [19], [17], [18].

Tang *et al.* [11] study the problem of learning user-to-user topic-wise influence strength. The input to their problem is the social network and a prior topic distribution for each node, which is given as input and inferred separately. As a consequence, they do not consider the simultaneous learning of topics and topic-wise influence. Further, their main focus is expert finding, and hence they do not propose any propagation model, nor study influence maximization.

A probabilistic model for the joint inference of the topic distribution and topic-wise influence strength has been proposed by Liu *et al.* [17]. Here the input is an heterogenous social network with nodes that are users and documents. The goal is to learn users' interest (topic distribution) and user-to-user influence. Gibbs-Sampling algorithm is used to estimate both topic distribution and influence weights.

Lin *et al.* [18] study the joint modeling of influence and topics, by adopting textual models. According to the generative semantic of the proposed approach, each document is generated by a mixture model on topics. The topic sam-

pling process takes into account document-to-documents non negative weights which models influence (in this case topic-inheritance), while novel aspects of the document are modeled by the evolution component.

Weng *et al.* [19] analyze topic-wise influence in `Twitter` by means of a two-step process. First, topics of interest for each user are extracted by means of LDA and topic-specific relationship networks are constructed. Then, in order to measure the influence of each user, they propose TwitterRank, an extension of the PageRank algorithm taking into account both the topic similarity and the social link structure.

What mentioned before for [11] holds for [17][18][19] too: *none of these papers define an influence propagation model* nor study the influence maximization problem.

III. Simple Topic-aware Propagation Models

As a first step towards topic-aware modeling of social influence, we extend the classic IC and LT models to their topic-aware versions.

Topic-aware Independent Cascade Model (TIC). In the topic-aware version of the IC model the user-to-user influence probabilities depend on the topic. Therefore, for each arc $(v,u) \in E$ and each topic $z \in [1,K]$ we are given a probability $p^z_{v,u}$, representing the strength of the influence exerted by user v on user u on topic z. Moreover for each item i that propagates in the network, we have a distribution over the topics, that is for each topic $z \in [1,K]$ we are given $\gamma_i^z = P(Z = z|i)$, with $\sum_{z=1}^{K} \gamma_i^z = 1$.

In this model a propagation happens like in the IC model: when a node v first becomes active on item i, has one chance of influencing each inactive neighbor u, independently of the history thus far. The tentative succeeds with a probability that is the weighted average of the link probability w.r.t. the topic distribution of the item i:

$$p^i_{v,u} = \sum_{z=1}^{K} \gamma_i^z p^z_{v,u}. \tag{1}$$

Topic-aware Linear Threshold Model (TLT). For each arc $(v,u) \in E$ and each topic $z \in [1,K]$ we are given a weight $p^z_{v,u}$, such that the sum of incoming weights in each node and for each topic is no more than 1. Each node u chooses a threshold θ_u uniformly at random from $[0,1]$. At time t, a node u which is not yet active on item i, is submitted to an influence weight

$$W_i^t(u) = \sum_{z=1}^{K} \sum_{v \in \mathcal{F}_i(u,t)} \gamma_i^z p^z_{v,u}. \tag{2}$$

where $\mathcal{F}_i(u,t)$ denotes the set of users that have a link to u and that at time t have already adopted the item i. If $W_i^t(u) \geq \theta_u$, then u will activate on item i at time $t+1$.

Observation 1: For both the TIC and TLT models the submodularity of the expected spread $\sigma_m(S)$ is directly inherited from the IC and LT models, respectively. In fact, in both cases only the model parameters are topic-aware, while the overall mechanism of propagation does not change. In particular, given an item i just let $p_{v,u} := \sum_{z=1}^{K} \gamma_i^z p^z_{v,u}$ (Eq. 1) to reduce from TIC to IC and from TLT to LT.
Hence it holds the following.

TABLE I: Some of the notation used

$z \in [1,K]$	a topic
$p^z_{v,u}$	strength of influence of v on u, on topic z
γ_i^z	topic distribution for item i
ϑ_u^z	topic distribution for user u
$t_i(v)$	the time at which v adopts item i
$D_i(t)$	$\{v \in V \mid t_i(v) = t\}$
$\underline{t_i}$ and $\overline{t_i}$	min and max t s.t. $D_i(t) \neq \emptyset$
$C_i(t)$	$\bigcup_{t' \leq t} D_i(t')$
$\mathcal{F}_i(u,t)$	$\{v \in V \mid (v,u) \in E \wedge v \in C_i(t)\}$
p_v^z	authoritativeness of user v in topic z
φ_i^z	relevance of item i in topic z

Proposition 1: The expected spread $\sigma_m(S)$ remains monotone and submodular for $m =$TIC or $m =$TLT.

Proof: The proof follows directly from the proofs for IC and LT in [2] and Observation 1. ■

A direct corollary is that the greedy algorithm provides an $(1 - 1/e - \phi)$-approximation for the influence maximization problem also under the TIC and TLT propagation models.

Next we define an Expectation Maximization (EM) method for learning the parameters of the TIC model.

A. Learning topic-aware influence

The problem of learning the parameters of the TIC models takes in input the social graph $G = (V,E)$, a log of past propagations $\mathbb{D}$, and an integer K. The propagation log is a relation `(User,Item,Time)` where a tuple $(u,i,t) \in \mathbb{D}$ indicates that user u adopted item i at time t. We assume that no user adopts the same item more than once. Moreover we assume that the projection of $\mathbb{D}$ on `User` is contained in the set of nodes V of the social graph G. We let $\mathcal{I}$ denote the universe of items, i.e., the projection of $\mathbb{D}$ on the second column. We also use D_i to denote the propagation trace of i, that is the selection of the tuples of $\mathbb{D}$ where `Item` $= i$, while $D_i(t)$ will denote the set of users that adopted i at time t, and $C_i(t) = \bigcup_{t' \leq t} D_i(t')$. Finally we use $\underline{t_i}$ and $\overline{t_i}$ to denote the first and last timestamp of adoption of item i.

The output of the learning problem is the set of all parameters of the TIC propagation model, which we denote Θ: these are γ_i^z and $p^z_{v,u}$ for all $i \in \mathcal{I}$, $(v,u) \in E$, and $z \in [1,K]$.

Assuming that each propagation trace is independent from the others, the likelihood of the data given the model parameters Θ, can be expressed as:

$$\mathcal{L}(\Theta; \mathbb{D}) = \sum_{i \in \mathcal{I}} \log \mathcal{L}(\Theta; D_i). \tag{3}$$

Saito *et al.* [10] assume that the input propagations have the same shape as they were generated by the IC model itself. This means that the propagation trace of an item i must be a sequence of sets of users $D_i(0), \ldots, D_i(n)$, corresponding to the discrete time steps of the IC propagation. Moreover for each node $u \in D_i(t)$ there exists a neighbor v of u such that $v \in D_i(t-1)$.

Following [20] we adopt a delay threshold Δ to define influencers. Specifically, suppose that u adopted i at time $t_i(u)$, and let $t_i(u) = \infty$ if u does not adopt i, then we define $\mathcal{F}^+_{i,u}$ as the set of u' neighbors that potentially influenced u in the selection of i:

$$\mathcal{F}^+_{i,u} = \{v|(v,u) \in E, 0 \leq t_i(u) - t_i(v) \leq \Delta\}.$$

The set $\mathcal{F}^-_{i,u}$ of u's neighbors who definitely failed in influencing u over i is defined similarly:

$$\mathcal{F}^-_{i,u} = \{v|(v,u) \in E, t_i(u) - t_i(v) > \Delta\}.$$

The main difference between the IC model and TIC, is that while in the former the probability that user v will succeed influencing u is the same for every item i, in the latter $p^i_{v,u}$ is a mixture over the user-to-user influence probabilities, where the mixture weights γ^i_z and the influence probabilities $p^z_{v,u}$ are the parameters to be learned. However, directly unpacking $p_{v,u}$ in order to expose γ^z_i and $p^z_{v,u}$ would lead us to a likelihood formulation which is not tractable in a closed form. We can tackle this problem by resorting to the "complete data" approach [14], which allows us to provide an effective closed form estimation of the parameters γ^z_i and $p^z_{v,u}$. The likelihood of a propagation trace D_i within the z-th component of the model can be defined as $P(D_i|z;\Theta) = \prod_u P^{i,z}_{u,+} P^{i,z}_{u,-}$, where

$$P^{i,z}_{u,+} = 1 - \prod_{v \in \mathcal{F}^+_{i,u}} (1 - p^z_{v,u}) \quad \text{and}$$

$$P^{i,z}_{u,-} = \begin{cases} \prod_{v \in \mathcal{F}^-_{i,u}} (1 - p^z_{v,u}) & \text{if } \mathcal{F}^-_{i,u} \neq \emptyset, \\ 1 & \textit{otherwise}. \end{cases}$$

In the rest of the paper, following the standard EM notation, $\hat{\Theta}$ will represent the current estimate of the set of parameters Θ. Assuming that each active neighbors v succeeds to activate u w.r.t. the generic item i with probability

$$R^i_z(u,v;\Theta) = \frac{p^z_{v,u}}{P^{i,z}_{u,+}} \tag{4}$$

the *Complete-Data Expectation Likelihood* [14] is given by:

$$\mathcal{Q}(\Theta;\hat{\Theta}) = \sum_i \sum_{z=1}^{K} Q_i(z;\hat{\Theta}) \left\{ \log \pi_z + \sum_u \right.$$

$$\left\{ \sum_{v \in \mathcal{F}^+_{i,u}} \left\{ R^i_z(u,v;\hat{\Theta}) \log p^z_{v,u} + \left(1 - R^i_z(u,v;\hat{\Theta})\right) \log(1 - p^z_{v,u}) \right\} \right.$$

$$\left. \left. + \sum_{v \in \mathcal{F}^-_{i,u}} \log(1 - p^z_{v,u}) \right\} \right\} \tag{5}$$

where π_z is the prior probability that a generic item is assigned to topic z. The mixture parameters γ^z_i, which define the TIC model in Eq. 1, are given by the values of $Q_i(z;\hat{\Theta})$ at the end of the learning procedure.

Let $S^+_{v,u} = \{i|v \in \mathcal{F}^+_{i,u}\}$, and similarly $S^-_{v,u} = \{i|v \in \mathcal{F}^-_{i,u}\}$. Moreover let

$$\kappa^+_{v,u,z} = \sum_{i \in S^+_{v,u}} Q_i(z;\hat{\Theta}), \text{ and } \kappa^-_{v,u,z} = \sum_{i \in S^-_{v,u}} Q_i(z;\hat{\Theta}).$$

Algorithm 1: EM inference of parameters for TIC

Input : Social graph $G = (V, E)$, data $\mathbb{D}$, and $K \in \mathbb{N}^+$.
Output: The set of all parameters of TIC, Θ, that is:
$\forall (v,u) \in E, \forall i \in \mathcal{I}, \forall z \in [1,K] : p^z_{v,u},\ \pi_z$ and γ^z_i.

$init(\pi_z, p^z_{v,u})$;
repeat
  E-step:
  forall the $i \in \mathcal{I}$ **do**
    forall the $z = \{1, \cdots, K\}$ **do**
      $Q_i(z;\hat{\Theta}) \leftarrow \frac{P(D_i|z;\hat{\Theta})\pi_z}{\sum_{\tilde{z}} P(D_i|\tilde{z};\hat{\Theta})\pi_{\tilde{z}}}$;
      forall the $(u,v) \in E$ **do**
        $R^i_z(u,v;\hat{\Theta}) \leftarrow \frac{p^z_{v,u}}{P^{i,z}_{u,+}}$;
      end
    end
  end
  M-step:
  forall the $z = \{1, \cdots, K\}$ **do**
    $\pi_z \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Q_i(z;\hat{\Theta})$;
    forall the $(u,v) \in E : S^+_{v,u} \neq \emptyset$ **do**
      $p^z_{v,u} \leftarrow \frac{1}{\kappa^+_{v,u,z} + \kappa^-_{v,u,z}} \sum_{i \in S^+_{v,u}} Q_i(z;\hat{\Theta}) R^i_z(u,v;\hat{\Theta})$
    end
  end
until convergence;

The Expectation-Maximization method for learning the parameters of the TIC model is given in Algorithm 1: it starts with a random initialization of parameters π_z (ensuring that $\sum_z \pi_z = 1$) and $p^z_{v,u}$ for all pair $\langle v,u \rangle$ such that $S^+_{v,u} \neq \emptyset$. Then it alternates the E-step and the M-step, measuring at each iteration the gain of log-likelihood (Eq. 5) w.r.t. the previous iteration. When the gain is below a given threshold, the algorithm has converged.

B. Dealing with new items in TIC

TIC model assumes that for each item we are given distribution over the topics and we have shown how to estimate this distribution by log-likelihood maximization. However, an interesting case is to apply the model to a new item never seen before, e.g., when we want to push a new product in the market. In this case we cannot directly apply the parameter estimation procedure described above, since no propagation trace of the new item is available yet. We have to rely on background knowledge about the item. For instance, the marketing expert might directly define the distribution over the topics for the given new item. Alternatively, item features (e.g., genre, price, etc.) might be available, or a small set of initial adopters might have provided tags.

In the most general setting, let us assume that multiple descriptions, in the form of sets of tags from a vocabulary $\mathcal{T}$, exist for item i. Let $\mathbf{w}_i$ denote the bag of tags obtained by joining all the descriptions of i and let w_n denote the n-th tag in $\mathbf{w}_i$. Then, we can extend the expected-likelihood formulation in order to take into account tag-assignments for items and maximize their likelihood. Let $\beta_{w_n,k}$ denote the probability of observing the n-th tag in the k-mixture:

$\beta_{w_n,k} = P(w_n|z_k)$. Assuming that influence probabilities and tags assignments are conditionally independent given the topic, the probability that the trace of item i will be generated by the z-th component is:

$$P(D_i|z;\Theta) = \prod_u P^{i,z}_{u,+} P^{i,z}_{u,-} \prod_{n=1}^{\mathbf{w}_i} \beta^{N(w_n,i)}_{w_n,k}$$

where $N(w_n, i)$ is the number of times that the tag w_n has been assigned to the item i. Then, the Complete-Data expectation likelihood becomes:

$$\mathcal{Q}(\Theta;\hat{\Theta})' = \mathcal{Q}(\Theta;\hat{\Theta}) + \sum_i \sum_{z=1}^{K} Q(z;i,\hat{\Theta}) \sum_{n=1}^{|\mathbf{w}_i|} N(w_n,i) \log \beta_{w_n,k}$$

and in the M-step we need to update the β distribution as:

$$\beta_{w_n,k} = \frac{1 + \sum_{i\in\mathcal{I}} Q_i(z;\hat{\Theta}) N(w_n,i)}{|\mathcal{T}| + \sum_{n'=1}^{|\mathcal{T}|} \sum_{i\in\mathcal{I}} Q_i(z;\hat{\Theta}) N(w_{n'},i)}.$$

Since no diffusion trace has been observed yet, it follows that both $\mathcal{F}^+_{i,u}$ and $\mathcal{F}^-_{i,u}$ are empty. In this case, the $P^{i,z}_{u,+}$ and $P^{i,z}_{u,-}$ components equal 1 and the overall probability reduces to the probability of observing the tags within topic z. In addition, γ^z_i can be computed as

$$\gamma^z_i = \frac{\prod_{n=1}^{|\mathbf{w}|} \beta_{w_n,z} \cdot \pi_z}{\sum_{z'} \prod_{n=1}^{|\mathbf{w}|} \beta_{w_n,z'} \cdot \pi_{z'}}. \tag{6}$$

C. Discussion

The traditional IC and LT models suffer various limitations when it comes to apply them in practice. One first limitation is the treatment of time and the consequent need for some discretization, as we have already highlighted in Section III-A. Another important limitation is the number of parameters. In fact both LT and IC have influence weights (or probabilities) for each pair of connected users. However, having $|E|$ parameters is unsuitable for real-world social networks where the number of edges is usually extremely large (for instance, Facebook nowadays exhibits $|E| > 130$ billion). The very large number of parameters, on the one hand makes the learning phase computationally prohibitive (the EM-based method needs to update the influence probability associated to each edge in each iteration), and on the other hand it makes the model prone to overfitting.

These limitations are not solved in the topic-aware TIC and TLT models that we have introduced in this section. Indeed, in TIC and TLT we have $K(|E| + |\mathcal{I}|)$ parameters.

The huge number of parameters can jeopardize the applicability of topic-modeling techniques. In the next section we introduce the AIR (*Authoritativeness-Interest-Relevance*) propagation model, which assumes that social influence depends on a user authority in the context of a given topic and the interest of the user social neighborhood for that topic. This assumption greatly reduces the number of parameters.

IV. The "AIR" Propagation Model

The AIR model has the following parameters:

- **Authoritativeness of a user in a topic:** For each user $v \in V$ and for each topic $z \in [1, K]$, we are given a weight $p^z_v \in \mathbb{R}$ which measures the strength of v's influence on the topic z. A positive value represent *authoritativeness*, i.e., given a topic, the activation of v with respect to an item will influence v's neighbors to select the item as well; on the other hand, negative values model *distrust*, i.e., the activation of v will discourage the activation of her neighbors.
- **Interest of a user for a topic:** each user u is defined by a distribution $\vec{\vartheta}_u$ over topics: i.e., $\vartheta^z_u = P(Z = z|u)$ denotes the interest of the user u in the topic z and $\sum_{z=1}^{K} \vartheta^z_u = 1$.
- **Relevance of an item for a topic:** each topic z is defined by a set of weights $\vec{\varphi}_z \in \mathbb{R}^{|\mathcal{I}|}$, with $\varphi^z_i \in \mathbb{R}$ being the relevance (or selection) weight for the item i in the topic z. Each topic can be hence characterized by the set of the most relevant items. For example, in the topic "Politics" the weight associated with the selection of the "NYT" is expected to be greater than the one corresponding to "Sport Illustrated".

The working principle of AIR is a *general threshold model* [2]. At the beginning of the process each user u chooses a threshold θ_u uniformly at random from $[0, 1]$. At time t, the decision of u to activate for a given item i depends on the influence exerted by her neighbors who have already activated on i (their authoritativeness) and on topic-wise u's interests and i's relevance. In details, at time t user u actives on i iff

$$P(i|u,t) = \sum_z P(z|u) P(i|u,z,t) \geq \theta_u$$

where $P(z|u) = \vartheta^z_u$, while $P(i|u,z,t)$ is the following logistic selection function:

$$P(i|u,z,t) = \frac{\exp\left\{\sum_{v\in V} p^z_v f_v(i,u,t) + \varphi^z_i f(i,u,t)\right\}}{1 + \exp\left\{\sum_{v\in V} p^z_v f_v(i,u,t) + \varphi^z_i f(i,u,t)\right\}} \tag{7}$$

The *selection scaling factors* $f_v(i,u,t)$ and $f(i,u,t)$ are used to distinguish potential influencers from non influencers ($f_v(i,u,t) = 0$ if $v \notin \mathcal{F}_i(u,t)$) and to potentially relate influence to time. As observed in [21], the likelihood of an item propagating is likely to decay proportionally to time. In particular, it decays at two different levels: *locally* the influence exerted by v on u for item i decays with the time elapsed from the moment in which v adopted i; *globally* the interest in the item i decays as i gets older. The adoption of the selection scaling factors in Eq. 7 allows to directly model both local and global temporal decay, e.g. by ensuring that $f_v(i,u,t) \propto (t_i(v) - t)$, and $f(i,u,t) \propto (\underline{t_i} - t)$.

Compared to the models presented in the previous section, the AIR influence propagation model has only $K(|V| + |\mathcal{I}|)$ parameters. As a result, AIR is a simpler model, more robust to overfitting and still capable of describing influence propagation in an effective way.

A. AIR: learning the parameters

The problem of learning the parameters of the AIR model, has the same input of the learning the parameters for TIC, presented in Section III-A.

Within the generative process, we can assume that for each given item i, a user u picks a topic z by drawing from her own characteristic distribution over the topics $\vec{\vartheta}_u$ (representing her prior interests). Then, for each timestamp t, u activates on i with probability $P(i|u,z,t)$ defined as in Eq. 7. Given the model parameters, we can compute the likelihood of the data as in Eq. 3. Recall that $D_i(t)$ denotes the set of users who selected the item i at time t, while $C_i(t)$ denotes the set of users who selected i by time t. For sake of notation compactness we use the binary indicators $d_i^u(t) = 1$ if $u \in D_i(t)$, and zero otherwise, and $c_i^u(t) = 1$ if $u \in C_i(t)$, and zero otherwise.

Then the Complete-Data Expectation Likelihood is:

$$\mathcal{L}(\Theta; Q) = \sum_i \sum_u \sum_z Q(z;u,i) \{\log \vartheta_u^z + \sum_{\underline{t_i}}^{\overline{t_i}} d_i^u(t) \log P(i|u,z,t) + (1 - c_i^u(t)) \log (1 - P(i|u,z,t))\} \quad (8)$$

Each observation $\langle u, i\rangle$ is associated with a state z of the latent variable, modeling the preference of u for i. Also, for the sake of simplicity, we assume that the hidden topic variable is independent from time. This modeling trick simplifies the formulation of the expected likelihood, and provides the following definition for the expected value:

$$Q(z;u,i) = \frac{\hat{\vartheta}_z^u \prod_{\underline{t_i}}^{\overline{t_i}} (P(i|u,z,t))^{d_i^u(t)} \cdot (1 - P(i|u,z,t))^{(1-c_i^u(t))}}{\sum_{z'} \hat{\vartheta}_{z'}^u \prod_{\underline{t_i}}^{\overline{t_i}} (P(i|u,z,t))^{d_i^u(t)} \cdot (1 - P(i|u,z,t))^{(1-c_i^u(t))}}$$

Within the EM framework, the $\vec{\vartheta}$ component can be obtained using standard optimization. The remaining parameters are difficult to solve in a closed form, due essentially to the non-linearity of Eq.7. We overcome this limitation by combining the *Improved Iterative Scaling* algorithm [22] and the *Generalized Expectation-Maximization (GEM)* procedure [23].

Rather than maximizing $\mathcal{L}(\Theta, Q)$, we look for an upgrade Γ of Θ that guarantees

$$\mathcal{L}(\Theta + \Gamma, Q) \geq \mathcal{L}(\Theta, Q)$$

In practice, this corresponds to find, for each p_v^z an upgrade δ_v^z and for each item i an upgrade η_i^z such that the M-step can be defined as $p_v^z \leftarrow p_v^z + \delta_v^z$ and $\varphi_i^z \leftarrow \varphi_i^z + \eta_i^z$.

We can express a lower bound on the difference $\mathcal{L}(\Theta + \Gamma, Q) - \mathcal{L}(\Theta, Q)$ through the inequality $-\log x \geq 1 - x$ and Jensen's inequality (by constraining $\sum_{v \in V} f_v(i,u,t) + f(i,u,t)$ to a constant value). By maximizing this lower bound, we can obtain a closed formula for the updates δ_v^z and η_i^z:

$$\delta_v^z = \log \left\{ \frac{\sum_{i,u} Q(z;u,i) f_v(i,u,t_i(u))}{\sum_{i,u} Q(z;u,i) \sum_{\underline{t_i}}^{t_i(u)} P(i|u,z,t) \cdot f_v(i,u,t)} \right\}$$

$$\eta_i^z = \log \left\{ \frac{\sum_u Q(z;u,i) f(i,u,t_i(u))}{\sum_u Q(z;u,i) \sum_{\underline{t_i}}^{t_i(u)} P(i|u,z,t) \cdot f(i,u,t)} \right\}$$

Algorithm 2 summarizes the overall learning scheme.

Algorithm 2: EM inference of parameters for AIR

Input : Social graph $G = (V, E)$, data $\mathbb{D}$, and $K \in \mathbb{N}^+$.
Output: The set of all parameters of AIR Θ, that are $p_u^z(\mathsf{A}), \vartheta_u^z(\mathsf{I}), \varphi_i^z(\mathsf{R})$, forall $u \in V, z \in [1,K], i \in \mathcal{I}$.

$init(p_u^z, \vartheta_u^z, \varphi_i^z)$; //Random initialization of parameters
repeat
  E-step:
  forall the $i \in \mathcal{I}$ **do**
    forall the $u \in V$ **do**
      forall the $z = \{1, \cdots, K\}$ **do**
        $Q(z;u,i) \leftarrow \frac{\vartheta_u^z \prod_{\underline{t_i}}^{\overline{t_i}} P(i|u,z,t)^{d_i^u(t)} \cdot (1 - P(i|u,z,t))^{(1-c_i^u(t))}}{\sum_{z'} \vartheta_u^{z'} \prod_{\underline{t_i}}^{\overline{t_i}} P(i|u,z',t)^{d_i^u(t)} \cdot (1 - P(i|u,z',t))^{(1-c_i^u(t))}}$
      end
    end
  end
  M-step:
  forall the $z = \{1, \cdots, K\}$ **do**
    forall the $v \in V$ **do**
      $\vartheta_v^z \leftarrow \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} Q(z;v,i)$
      $\delta_v^z \leftarrow \log \left\{ \frac{\sum_{i,u} Q(z;u,i) f_v(i,u,t_i(u))}{\sum_{i,u} Q(z;u,i) \sum_{t_m(i)}^{t_i(u)} P(i|u,z,t) \cdot f_v(i,u,t)} \right\}$
    end
    forall the $i \in \mathcal{I}$ **do**
      $\eta_i^z \leftarrow \log \left\{ \frac{\sum_u Q(z;u,i) f(i,t_i(u))}{\sum_u Q(z;u,i) \sum_{t_m(i)}^{t_i(u)} P(i|u,z,t) \cdot f(i,t)} \right\}$
    end
    forall the $u \in V$ **do**
      $p_u^z \leftarrow p_u^z + \delta_u^z$
    end
    forall the $i \in \mathcal{I}$ **do**
      $\varphi_i^z \leftarrow \varphi_i^z + \eta_i^z$
    end
  end
until convergence;

B. AIR: dealing with new items

Modeling unobserved items follows the general guidelines exposed in Sec. III-B, with some variations. The selection probability for a new item can be simplified as:

$$P(i|u,t) = \sum_z P(z|u) P(i|z,u,t) = \sum_z \vartheta_u^z \frac{\exp\{\varphi_i^z f(i,u,t)\}}{1 + \exp\{\varphi_i^z f(i,u,t)\}} \quad (9)$$

In general relevance parameter φ_i^z for a new item i is not bound to an optimal value. However, when tag information is available, we can assume a prior tendency of the item to be selected, according to its likelihood to be associated with the topic. That is, we can model new item by assuming a prior probability $p(\varphi_i^z)$, defined as a gaussian distribution with constant variance σ and mean γ_i^z (as defined in Eq. 6). As a consequence, the log-likelihood can be reformulated to

comprise the prior probabilities, resulting into

$$\mathcal{L}(\Theta;\mathbb{D}) = \sum_{i\in\mathcal{I}} \log \mathcal{L}(\Theta; D_i) + \log P(\Theta)$$

A more thorough *maximum a posteriori estimation* (MAP) treatment for the whole parameter set Θ is omitted here for lack of space. Without loss of generality, we assume uniform prior probabilities for all the parameters other than new items. As a consequence, $P(\Theta)$ can be simplified as:

$$\log P(\Theta) = \sum_{i:D_i=\emptyset} \sum_{z} \log p(\varphi_i^z) + C$$

Combining the above equations and Eq. 8 finally yields

$$\mathcal{L}(\Theta;Q)' = \mathcal{L}(\Theta;Q) + \sum_{i:D_i=\emptyset} \sum_{z} \log p(\varphi_i^z) + C$$

Optimizing the latter with respect to a parameter φ_i^z relative to a new item i yields the straightforward solution $\varphi_i^z = \gamma_i^z$.

C. Influence Maximization in AIR

We next discuss the problem of influence maximization in AIR. Given a generic item i that we want to promote, we assume that its AIR parameters are known. The problem is to select a set S of k nodes such that the expected spread of influence of S under the AIR model, denoted $\sigma_{\mathsf{AIR}}(S)$, is maximal.

Although AIR is a *general threshold model*, the fact that user authoritativeness can be negative makes σ_{AIR} not submodular and not even monotone. Therefore the standard greedy algorithm cannot provide any approximation guarantee, as it does for the classic IC and LT models, and for their topic-aware versions TIC and TLT.

Even without any provable guarantee, it is reasonable to consider the greedy algorithm a reasonable candidate also for the AIR model, given that in any case we shall naturally avoid users with negative authoritativeness. Therefore, in the next section we compare the spread $\sigma_{\mathsf{AIR}}(S)$ achieved by the following two methods:

- **Greedy:** at each iteration greedily add to the set of seeds S the node x that brings the largest marginal gain, i.e., $\sigma_{\mathsf{AIR}}(S \cup \{x\}) - \sigma_{\mathsf{AIR}}(S)$ is maximal. Estimate $\sigma_{\mathsf{AIR}}(S)$ for a given S by Monte Carlo simulations [2].
- **Top-k authorities:** given the new item i and its distribution over topics γ_i^z, select the top-k users v w.r.t.

$$\sum_{z=1}^{K} \gamma_i^z p_v^z.$$

Recall that all over the paper K is the number of topics, while here k is the size of the required seed set.

Studying alternative approaches to influence maximization under the AIR model will be part of our future investigation.

	FLIXSTER		DIGG	
	Training	Test	Training	Test
Users	6,572	4,686	16,297	14,061
Items	7,158	7,138	3,553	3,547
Actions	1,432,716	340,495	1,160,428	264,066
Avg # actions (user)	218	72	71	18
Avg # actions (item)	200	47	326	74
Min # actions (user)	6	1	6	1
Min # actions (item)	9	1	90	3
Max # actions (user)	5,525	1,786	2,640	1,912
Max # actions (item)	3,173	778	4,995	828
Avg lifetime (item)	952 days		14 days	
	Avg time between two actions			
per user	94 hours		66 hours	
per item	22 days		38 minutes	

TABLE II: Summary of the propagation data.

V. EXPERIMENTAL EVALUATION

The goal of our experiments is twofold. At a high level, we want to evaluate the impact of introducing a topic-based estimation of the influence probabilities. That is, we are interested in evaluating whether topic-aware propagation models can better predict the activation of a user on a specific item. The expected result is that the combined adoption of both influence and topic modeling exhibits an improvement over the single contributions. We also aim aim at assessing whether considering the topic model of the item can bring any benefit in a viral marketing campaign. That is to say, to compare topic-aware models against models that ignore the topic distribution of the item, in the influence maximization problem.

Datasets. We use two real-world and publicly available datasets, both containing a social graph $G = (V, E)$ and a log of past propagations $\mathbb{D} = \{(\texttt{User}, \texttt{Item}, \texttt{Time})\}$: the datasets come from Digg (www.digg.com) and Flixster (www.flixster.com). Digg is a social news website, where the users vote stories. In this case $\mathbb{D}$ contains information about which user voted which story (item) at which time. If we have user v vote a story about the new iPhone, and shortly later v's friend u does the same, we consider the story as having propagated from v to u, and v as a potential influencer for u. Flixster is one of the main players in the mobile and social movie rating business. Here, an item is a movie, and the action of the user is rating the movie.

In both cases we started from the publicly available dataset[1] [2] and we performed some standard consistency cleaning and removal of all users and items that do not appear at least 20 times in $\mathbb{D}$. The final DIGG social graph contains 11,142 users and 99,846 directed arcs, while FLIXSTER contains 6,353 users and 84,606 directed arcs: in both cases we do not consider the disconnected nodes, i.e., users that appear actively in $\mathbb{D}$ but which have no friends in G.

Moreover, for our purposes we performed a chronological split of $\mathbb{D}$ in both datasets into training (80%) and test (20%). Table II summarizes the main properties of $\mathbb{D}$.

Experiments settings. We start by noticing that there is a direct relationship between the the scaling factors $f_v(i, u, t)$ of the AIR model and the size of influence window Δ used in the parameters learning of the IC and TIC models. We studied two alternative definitions for $f_v(i, u, t)$.

[1] www.isi.edu/~lerman/downloads/digg2009.html
[2] http://www.cs.sfu.ca/~sja25/personal/datasets/

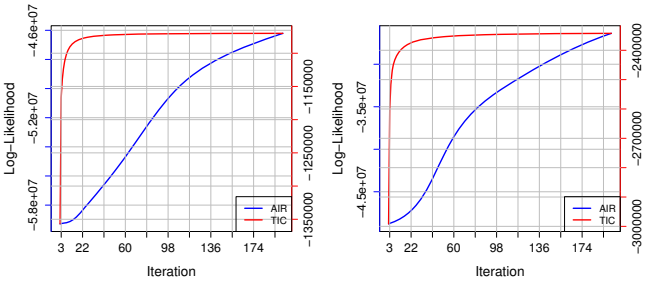


Fig. 1: Convergence rate: AIR vs. TIC on DIGG (left) and FLIXSTER (right).

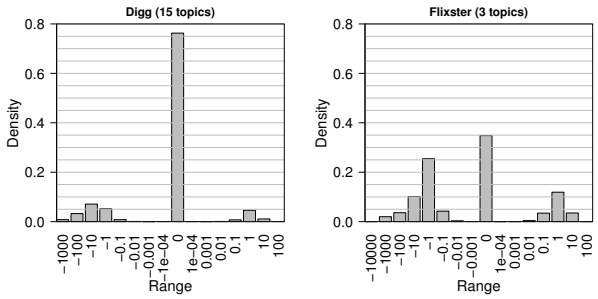


Fig. 2: Distribution of p_v^z in the AIR model.

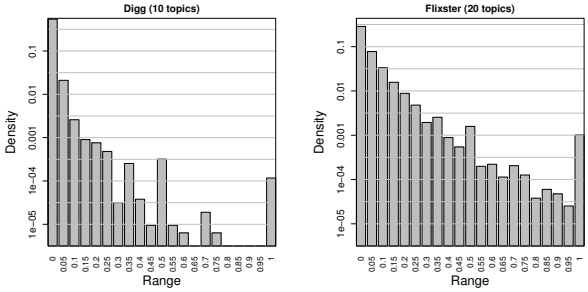


Fig. 3: Distribution of $p_{v,u}^z$ in the TIC model (reporting only the non-null values).

The first one assumes that propagation can degrade following an exponential decay:

$$f_v(i,u,t) \propto \begin{cases} \exp(t_v(i) - t) & \text{if } v \in \mathcal{F}_i(u,t) \\ 0 & \text{otherwise} \end{cases}$$

This definition of scaling factor corresponds to a very short influence threshold Δ (typically, 3 to 5 timestamps). The second option we explored is to keep $f_v(i,u,t)$ constant. This corresponds to adopting a value $\Delta = \infty$ within the IC and TIC models. As a matter of fact, the statistics on the average time between two actions involving the same item, and the average time-life for an item in Table II suggest for a large Δ. Our empirical analysis determined that, at least in these two datasets, the best results are achieved by considering all the influencers up to the considered time: i.e., $\Delta = \infty$ and consequently $f_v(i,u,t)$ constant.[3] Therefore in the experiments reported here we always adopt these settings.

Learning. In Figure 1 we compare the learning rate of the TIC and the AIR model in the first 200 iterations. As expected, TIC exhibit an faster convergence rate than AIR: this is due to the difference in their respective M-step. AIR relies on a *GEM* procedure which clearly affects the number of iterations needed to achieve convergence. Notably, the TIC parameter estimation phase provides a good estimation of the model parameters after about 60 iterations on both the datasets, whereas the AIR model requires approximately 1400. Also, both algorithms are initialized randomly, but the likelihood increase for AIR is slower. We plan to investigate ways to speed up the parameter estimation phase of the AIR model, as well as better initialization strategies in future works.

Finally, Figure 2 plots the distributions of the influence weights of the AIR model for the number of topics achieving the best performances on the two considered datasets (as described later in this section). Values are distributed according to two log-normal distributions centered in the positive and negative quadrants, with relatively slow values and relatively few extreme values. The graphs show that negative influencers also play a significant role in the learning phase. The Digg dataset exhibits a lower level of influence among users, as witnessed by the high number of weights set to 0. Figure 3 plots of the influence probabilities for the TIC models, and confirm such a trend. Here, values are exponentially distributed; however a percentage of users exhibit highest influence.

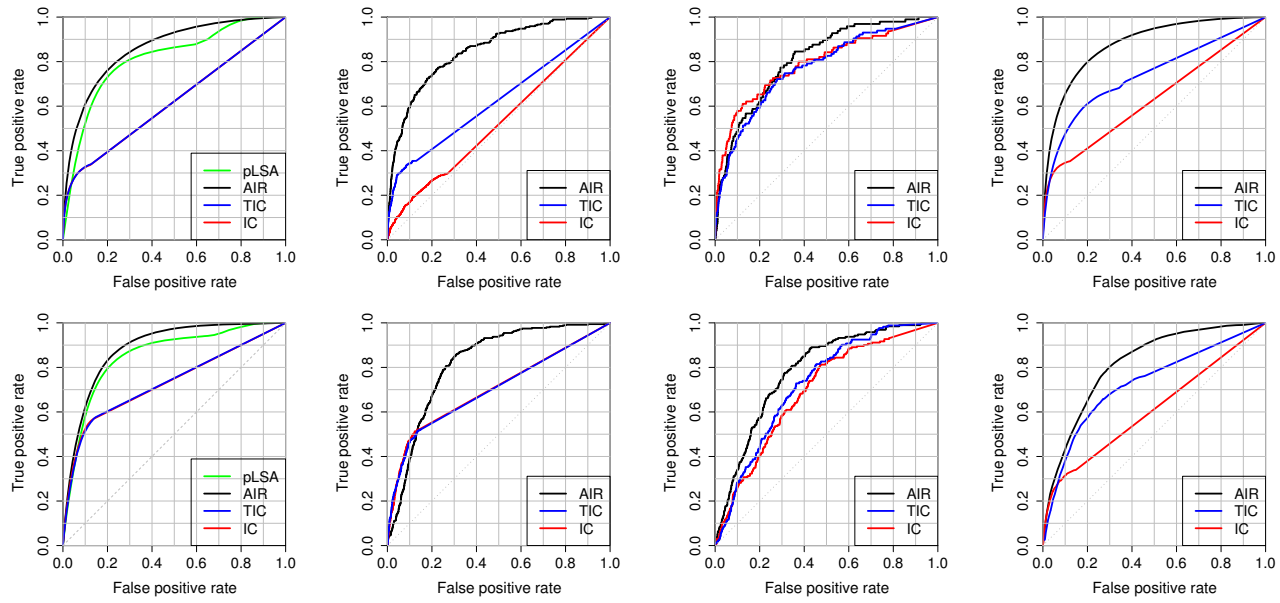
[3]This is in accordance with the experiments in [20], that firstly introduced the Δ influence window.

A. Predictive accuracy

In the following we compare IC, TIC and AIR: the parameters of the model are learned using the EM method in [10], the method in Section III-A, and in Section IV-A respectively. The basic principle guiding our evaluation can be summarized as follows. Given the training propagation data $\mathbb{D}_T$ and a test propagation data $\mathbb{D}_{Test}$, a generic model, whose parameters have been learned on $\mathbb{D}_T$, provides a suitable estimation of influence and behavior if its application to unobserved data $\mathbb{D}_{Test}$ provides accurate predictions, which can be measured through the following tests.

Activation Test (General). The idea is to measure whether a diffusion model can predict the overall user's activations. This is basically a binary prediction task: for a given user-item pair $\langle u,i\rangle \notin \mathbb{D}_T$, we try to predict whether $\langle u,i\rangle \in \mathbb{D}_{Test}$. Since this test is time-independent, we also use as a baseline for comparison the *Probabilistic Latent Semantic Analysis (pLSA)* model [15]. Although not originally aimed at modeling influence, the latter also relies on topic modeling and occurrences of user actions. Hence, its inclusion in the test allows us to evaluate the contribution of topic modeling on the activation prediction.

Selection Probabilities (General). For each pair $\langle u,i\rangle$ we measure the degree of responsiveness of the model at the actual



(a) activation test (general) (b) selection probs. (general) (c) selection probs. (inf. episodes) (d) activation time

Fig. 4: ROC analysis: DIGG (first row) and Flixster (second row).

activation time $t_i(u)$ (if it exists). A good model should assign high probability of activation to a positive case $\langle u, i\rangle \in \mathbb{D}_{Test}$, and low probability (relative to all the possible timestamps) to a negative case $\langle u, i\rangle \notin \mathbb{D}_{Test}$.

Selection Probabilities (Influence Episodes). The previous test strongly penalizes pure influence-based diffusion models, as they assign zero probability to episodes $\langle u, i, t\rangle$ for which the set of influencers is empty. This is not true, of course, for the AIR model which is able to capture the relevance of an item for a given topic. In order to measure the effects of influence, in this test we focus on those episodes $\langle u, i, t\rangle$ for which $\mathcal{F}_i(u, t) \neq \emptyset$.

Activation Time (Influence Episodes). A final test measures the precision of activation at a given timestamp, only considering episodes with non-empty influencers set. Each pair $\langle u, i\rangle \notin \mathbb{D}_T$ is evaluated by comparing the true activation time (if any) with the predicted activation time. Let $t'_i(u)$ represent the predicted activation timestamp, i.e., the minimal timestamp t where $P(i|u, t)$ is greater than a given activation threshold (with $t'_i(u) = \infty$ if the model does not indeed predict any activation for the given item). We can devise the following confusion matrix:

	$\langle u, i\rangle \in \mathbb{D}_{Test}$	$\langle u, i\rangle \notin \mathbb{D}_{Test}$
True Positive	$t'_i(u) = t_i(u)$	-
False Positive	$t'_i(u) < t_i(u)$	$t'_i(u) \neq \infty$
True Negative	-	$t'_i(u) = \infty$
False Negative	$t'_i(u) > t_i(u)$	-

For all the above mentioned tests we plot the *Receiver Operating Characteristic* (ROC) curves relative to varying activation thresholds. The results are given in Figure 4, while in Table III we report the *Area Under the Curve* (AUC) values.

Evaluation. We experimentally found that the optimal number of topics on Digg is 15 topics for AIR, and 20 on TIC. Also, Flixster settles 3 topics on AIR, and 10 on TIC.

The AIR models achieve the best results in detecting the activations, with a consistent gain over the other models (including the runner-up pLSA model). Independent cascade models (IC and TIC) exhibit partial curves on this test, limiting the upper bound of FPR to 0.1. This is due to the fact that negative cases $\langle u, i\rangle$ are a vast majority, and when the case exhibit no active influencers the IC and TIC models assign 0 probability, which eventually results in a True Negative in the test. For this reason, the extension to topics does not provide a significant improvement: both IC and TIC overlap, and the difference in AUC is marginal. Things change when activation time is taken into account: in the remaining plots, TIC outperforms IC, an evidence of a substantial contribution of the topic modeling in increasing the accuracy of time-oriented predictions. Again, AIR achieves the best accuracy among all the models under investigation.

Tests in Fig.4(d) are the most fine-grained: here underestimation of influence (resulting in retarded activation prediction) as well as overestimation (resulting in anticipated activation prediction) are paid as errors. Clearly, topic modeling plays a crucial role in this test, as it allows to better correlate the estimation phase to the actual activation time.

B. Influence Maximization

We now turn our attention to the influence maximization problem and to the following questions: (1) how important is it to consider the topic-distribution of the item while selecting the seed sets? (2) how good are the greedy algorithm and the top-k-authorities heuristic on the AIR model? (3) how much does the item "popularity" affect the overall spread?

In Figure 5(left) we compare the expected spread achieved on the AIR propagation model by the greedy algorithm and the top-k-authorities heuristic. The experiment is performed on FLIXSTER, using 50 different items, and averaging the

Model	DIGG	FLIXSTER	DIGG	FLIXSTER
	Activation Test (General)		Selection Probs. (General)	
AIR	**0.8585511**	**0.8857634**	**0.8484368**	**0.8201586**
TIC	0.6190136	0.731208	0.6256339	0.7000218
IC	0.6189209	0.730694	0.5256555	0.702175
	Selection Probs. (Inf. episodes)		Activation Time	
AIR	**0.8123432**	**0.7834864**	**0.8784483**	**0.8150082**
TIC	0.7714797	0.7253222	0.7377654	0.7377654
IC	0.7916101	0.6940882	0.6294611	0.6089646

TABLE III: Summary of the evaluation: AUC values.

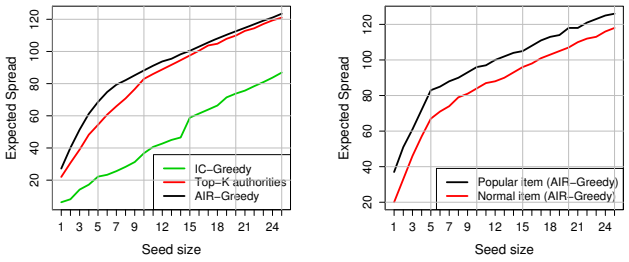


Fig. 5: Influence maximization experiments.

results. Items are described by their relevance over 3 topics. We also add to the comparison a seed set selected by the greedy algorithm on the IC model: i.e., without considering the topics. Being topic-blind, the IC experiment is run only for one generic item. All the greedy algorithms use 1000 Monte Carlo simulations to estimate the expected spread.

Although (as discussed in Section IV-C) the greedy algorithm does not provide approximation guarantee, it outperforms the top-k-authorities heuristic. The latter still performs very well: it achieves a spread quite close to those of the greedy approach, and in addition it is much faster to compute. More importantly, both topic-aware strategies largely outperform the topic-blind IC-greedy strategy.

In Figure 5 (right) we compare a "popular" item, i.e., an item which has a rather high relevance (a value of 10) in all three topics, with a normal item having relevance 10 in one topic, and relevance 1 in the other two topics. Not surprisingly, we can observe that the popular item achieves a larger spread. The difference tends to decrease with larger seedset: apparently, popular items are tolerant to smaller seedsets, whereas general items require more seeds.

VI. Conclusions and Future work

We provided a topic-modeling perspective over social influence, by introducing novel topic-aware propagation models. We devised methods to learn model parameters from a log of past propagations. We experimentally found the proposed models more accurate in describing real-world influence-driven propagations than the state-of-the-art approaches: as a matter of fact, the two proposed models exhibit an average 28% (AIR) and 7% (TIC) improvement on AUC over the baseline IC approach. The tests show that the models provide accurate predictions of both activations and activation times, and they provide robust estimates of influence parameters. Finally, we showed that by considering the characteristics of the item we can obtain larger spread in influence maximization.

There are several ways to extend the main results of this paper. First of all, we plan to investigate ways to speed up the parameter estimation phase of the AIR model, as well better initialization. Also, from a modeling perspective, a full bayesian treatment of the topic models introduced here can help with model generalization and overfitting avoidance.

We also plan to study influence maximization methods based on the AIR model. Finally, we plan to extend the focus of this paper to further application domains, by investigating how to combine influence maximization with topic modeling for recommender systems.

Acknowledgments. This research was partially supported by the Torres Quevedo Program of the Spanish Ministry of Science and Innovation, and partially funded by the European Union 7th Framework Programme (FP7/2007-2013) under grant n. 270239 (ARCOMEM).

References

[1] F. Bonchi, "Influence propagation in social networks: A data mining perspective," *IEEE Intelligent Informatics Bulletin, Vol.12 No.1*, 2011.
[2] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *KDD*, 2003.
[3] P. Domingos and M. Richardson, "Mining the network value of customers," in *KDD*, 2001.
[4] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *KDD*, 2002.
[5] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions - i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.
[6] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *PKDD*, 2006.
[7] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in *KDD*, 2007.
[8] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *KDD*, 2010.
[9] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *PVLDB*, vol. 5, no. 1, 2011.
[10] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *KES*, 2008.
[11] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *KDD*, 2009.
[12] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *WSDM*, 2010.
[13] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *WWW*, 2010.
[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.
[15] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *UAI*, 1999.
[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[17] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *CIKM*, 2010.
[18] X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The joint inference of topic diffusion and evolution in social communities," in *ICDM*, 2011.
[19] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *WSDM*, 2010.
[20] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen, "Sparsification of influence networks," in *KDD*, 2011.
[21] D. Ienco, F. Bonchi, and C. Castillo, "The meme ranking problem: Maximizing microblogging virality," in *Workshops of ICDM*, 2010.
[22] K. Nigam, "Using maximum entropy for text classification," in *In IJCAI Workshop on Machine Learning for Information Filtering*, 1999.
[23] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, MIT Press 1998.