Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Generating Realistic Interest-Driven Information Cascades

Federico Cinus, Francesco Bonchi, Corrado Monti, André Panisson
ISI Foundation, Turin, Italy
first name.last name@isi.it

Abstract

We propose a model for the synthetic generation of information cascades in social media. In our model the information "memes" propagating in the social network are characterized by a probability distribution in a topic space, accompanied by a textual description, i.e., a bag of keywords coherent with the topic distribution. Similarly, every user of the social media is described by a vector of interests defined over the same topic space. Information cascades are governed by the topic of the meme, its level of virality, the interests of each user, community pressure, and social influence.

The main technical challenge we face towards our goal is the generation of realistic interest vectors, given a known network structure and a tunable level of homophily. We tackle this problem by means of a method based on non-negative matrix factorization, which is shown experimentally to outperform non-trivial baselines based on label propagation and random-walk-based graph embedding.

As we showcase in our experiments, our model offers a small set of simple and easily interpretable "knobs" which allow to study, *in vitro*, how each set of assumptions affects the resulting propagations. Finally, we show how to generate synthetic cascades that have similar macro-statistics to the real-world cascades for a dataset containing both the network and the cascades.

1 Introduction

Modelling information diffusion through social media is an important task towards understanding the global phenomena that emerge from the basic mechanisms of human communication and interactions. Many questions in the field revolve around critical problems of present society. How can we help social media users to distinguish misinformation from legit news as they propagate? (Vosoughi, Roy, and Aral 2018) How interactions on social media affect opinion formation and dynamics? What is the role played by social bots in tampering political debates on social media (Ferrara et al. 2016)? How is it that small initial shocks can cascade to affect a large system, such as a communication network? (Watts 2002) These questions – often studied also in the context of *viral marketing* (Richardson and Domingos 2002; Bonchi 2011; Aslay et al. 2015) – have became central to our understanding of historical transformations of our times (Lane 2011).

The analysis of information cascades in social media revolves around two main themes: communication and its social network substrate. On the one hand, the network topology has proved itself to be an important factor that affects the information diffusion (Weng et al. ; Weng, Menczer, and Ahn 2014), and it can be described by several macroscopic characteristics, e.g., the level of homophily (Yuan, Alabdulkareem, and others 2018) (Weng, Menczer, and Ahn 2014) or the modular structure of the network (Barbieri, Bonchi, and Manco 2013a; Mehmood et al. 2013), as well as node-level characteristics, such as e.g., their centrality, or their capacity of spanning structural holes, thus bridging communities and facilitating, or blocking, the spread of information. On the other hand, the study of the diffusion processes happening over the network allows to analyse phenomena such as, e.g., the level of virality of the memes which are propagating, theirs topics, or their polarity (e.g., on a controversial debate), or if the propagation is driven by influential nodes or by group pressure. All these parts interoperate together in a complex way. Viewing them as parts of a single system allows us to ask ourselves new questions. Are influential nodes influential on all topics? Is propagation of highly viral items encouraged or discouraged by the presence of echo chambers?

The ingredients at play here are many: the structure of the social networks, the interests of each individual (which can exhibit more or less homophily w.r.t. the structure of the network), the strength of influence that nodes can exert on their peers, the items that propagate in the network, described explicitly by their bag-of-words representation or implicitly by their topic distribution. Having such richness of data from real-world interactions is not always easy or possible, due to the proprietary nature of social media data and to privacy regulations. Thus the driving research question we address in this work is the following: *can we devise a model able to coherently describe all these ingredients and use it to generate realistic information cascades?*

Besides the obvious benefits (i.e., availability, size control, no privacy issues) synthetic data generation allows to

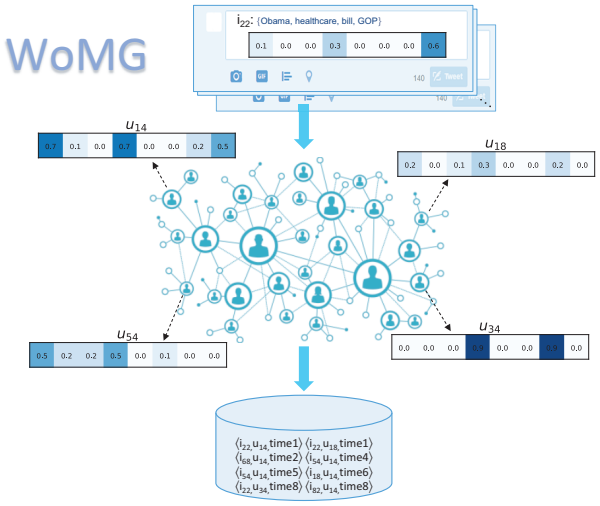


Figure 1: Bird's-eye view of WoMG: given a social network and a topic model, WoMG generates a database of propagation cascades.

study, *in vitro*, specific phenomena of interest by controlling the parameters of the model: e.g., having a more or less homophilic network, having a larger or smaller role of social influence in driving the cascades, and so on.

Overview of the model. Our model, dubbed WoMG (for Word-of-Mouth Generator), takes as input:

$(I1)$ a directed social graph structure $G = (V, E)$, and

$(I2)$ a topic model, as the one produced by running any topic-modeling algorithm (e.g. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003)) on a corpus of documents.

Let k be the number of topics. Each node gets labelled by WoMG with a k-dimensional vector representing how much it is interested in each topic. Similarly, each item that propagates in the network is described by a probability distribution on the topic-space. Besides its topic-distribution, each item will also have some content: i.e., a bag of keywords generated using the topic model. Alternatively, one can use real-world documents (as the ones in the corpus that generated the topic model) and feed them to the social network.

When a new item enters in the network, it starts propagating: nodes can *activate* on it based on their interests. Once a node u activates on the item (i.e., they like or repost the item), their followers become aware of the item and based on their interests and u's strength of influence, they might activate and propagate the item further. Such contagion process is governed by several parameters, such as, e.g., the level of virality of the item. The output of this process includes the following elements:

$(O1)$ the vector of interests for each node $v \in V$;

$(O2)$ a set of items I, where each item $i \in I$ is described by a bag-of-words and a distribution in the topic space;

$(O3)$ a propagation trace for each item $i \in I$, where a propagation trace is a relation (i, v, t) representing the fact that the item i was adopted by node v at time t.

Figure 1 provides a bird's-eye view of WoMG. A more formal and detailed description of our model is instead provided in Section 3.

Challenges. A key role in the topic-aware propagation model is played by the interests of each individual. The main technical challenge we had to solve is that of *generating realistic interest vectors, given a known network structure*. It is well known, in fact, that links in a social network are often shaped by homophily (McPherson, Smith-Lovin, and Cook 2001). Therefore, we need a way to generate vectors that respect this property – otherwise, any propagation model would fail to achieve sensible results. However, it has also been proved that the level of homophily is different among different networks (Bisgin, Agarwal, and Xu 2012). For this reason, we also wish for our method to achieve a *tunable level of homophily* in the generated interest vectors.

We tackle this problem by means of a method based on non-negative matrix factorization, which is shown experimentally to outperform non-trivial baselines: a method based on a simple iterative algorithm that exploits community structure, i.e., *label propagation* (Zhu and Ghahramani 2002), and one exploiting recent advances in random-walk-based graph embeddings, in particular using the *node2vec* (Grover and Leskovec 2016) method.

Beside interests, we needed to identify which others factor are important in shaping propagations in a multi-topic setting, and how to connect them. For instance: how to generate consistently the virality of the items, the influence capability of each node, or how to generate initial activations. For each of those choices, we devised different configurations of the model.

Combining these configurations with the tunable levels of homophily and the other key parameters we identify, we obtain a small set of simple and easily interpretable "knobs" to adjust the behavior of our generator. In this way, we are able to easily turn those knobs in order to study how each set of assumptions affects the resulting propagations. We found out, in fact, that there is a significant intertwining between these factors: for instance, the effect of virality on propagation dynamics is different depending on how items are introduced in the social network. Moreover, if nodes have different capabilities of influencing their peers, the presence or absence of echo chambers heavily affects the spread of highly viral content. We report our findings in the experimental section.

Roadmap. Next section briefly survey related literature. Section 3 describe in details WoMG. Section 4 discusses our solutions for the technical challenge of generating realistic interest vectors, given a known network structure and a tunable level of homophily. Finally, Section 5 presents our experimental findings.

2 Related Work

Studying how information propagates through a society has been a key question since the early days of social science. With the advent of the Web, and in particular with the rise of

social media, it has been possible to evaluate existing sociological theories on large scale real data, while many new questions have arisen about how information propagates on social media specifically. González-Bailón et al. (2011) studied the spread of protest movements in the Twitter network, finding evidence of social influence and complex contagion. In particular, they validated threshold models, by reproducing observed real behavior. Borge-Holthoefer and Moreno (2012) instead used a simulated, generative model to identify influential nodes. They find that the spreading capabilities of the nodes do not depend on their topological property (specifically, on their coreness).

A key question in many works is to clarify the relationship between social influence and homophily. When two friends propagate the same content, is it because they influenced each other, or because they appreciate similar content? Anagnostopoulos, Kumar, and Mahdian (2008) built a model that separates between social influence and homophily. They applied it to a data set collected from Flickr, and they found that information propagation on that web service is more likely to be caused by homophily than by social influence. On the same topic, Bakshy et al. (2012) conducted a large scale field experiment at Facebook using their users as test subjects. They report that not only social influence play a significant role, but that weak ties often help in spreading information that would not have otherwise spread, in particular for less viral content. Goyal, Bonchi, and Lakshmanan (2010) found, again on the Flickr data set, that by assuming social influence between individuals it is possible to predict several micro-level patterns of the cascades with high accuracy.

Many other works tried to approach information propagation as a prediction or inference task. For instance, Adar and Adamic (2005) reconstructed the path of individual cascades across political blogs. Goyal, Bonchi, and Lakshmanan (2011) and Cheng et al. (2014) tried to predict cascade size from initial characteristics of the cascade.

Plenty of efforts have been devoted to the characterization of information cascades and understanding which factors shape them. In (Cheng et al. 2016), authors studied the recurrence in time of content shared on social network. They found that homophily in some cases help content propagate at the beginning; but it may result then "in the content getting trapped in a local part of the network", thus explaining the emergence of echo chambers. They also find that content virality and network homophily are closely related, and that they are both driving factors shaping the information cascades. Their view on homophily is consistent with the findings of Sasahara et al. (2019). In this work, authors modeled and simulated the evolution of the topology of a social network. They find out that, even with minimal amounts of influence and unfriending, the network develops into segregated communities of similar nodes (i.e., echo chambers).

While many of the previous study focused on prediction and real data experiments, some studies in information propagation tried to explain observed patterns through generative models and simulations. In (Gleeson et al. 2014), authors compare different assumptions on the dynamics of the spreading by simulating a set of agents and observing their behavior. This type of analysis has been giving interesting results in related research fields, from classical examples such as the segregation model (Schelling 1969) to novel results on opinion dynamics (Del Vicario et al. 2017). Beside allowing to study *in vitro* the behavior of agents, generative models allow to build data sets that can be used as test bed for prediction tasks or inference algorithms. For instance, benchmark graphs (Lancichinetti, Fortunato, and Radicchi 2008) have been recognized as an important assessment for community detection algorithms.

3 Model

In this section we describe in details our model which builds on top of the *Topic-aware Linear Threshold* Model introduced in (Barbieri, Bonchi, and Manco 2013b). Such topic-aware influence propagation model stems from three main assumptions: (i) nodes have different interests, (ii) items have different topics, (iii) similar items are likely to activate similar nodes. These assumptions describe a propagation cascade through social influence, dependent on the topics of each item and the interests of each node.

We consider a directed social graph $G = (V, E)$ where the *nodes* V represent the set of individuals involved in the social network, and a directed link $(u, v) \in E$ represents the fact that v is a *follower* of u. As such, v receives is her timeline the bits of information shared by u and can be influenced by u to share further, thus allowing the propagation of information. We denote with $N^+(u) \subseteq V$ the out-neighborhood of node u, i.e., the set of followers of u. Let k be the number of topics under consideration. Each node $u \in V$ is labeled with a *vector of interests* $\mathbf{t}_u \in \mathbb{R}^k$. The quantity $t_{u,z} \geq 0$ represents how much node u is interested in topic z.

We also have a set I of items that propagate over the social network. Each item $i \in I$ is represented by a k-dimensional vector γ_i, that is a probability distribution over topics, i.e.,

$$\sum_{z=1}^{k} \gamma_i^z = 1,$$

where γ_i^z represents how much item i is of topic z. An item is also defined by a positive scalar that represents its propensity to propagate, which we denote as *virality* v_i.

At time t, a node u receives a *social pressure* $W_i^t(u)$ to activate on item i:

$$W_i^t(u) = \sum_{z=1}^{k} \left(\gamma_i^z \sum_{v \in F_i(u,t)} p_{v,u,z} \right) \tag{1}$$

where

- $p_{v,u,z}$ is the pressure exerted by v on u on topic z;
- $F_i(u,t)$ is the set of nodes already active on i at time t and that have u as follower:

$$F_i(u,t) = \{v \in V \| (v,u) \in E \wedge i \in D_t(v)\}$$

where $D_t(v)$ is the set of items adopted by v at time t.

Since it is a linear threshold model, activations happen when this pressure $W_i^t(u)$ exceeds a threshold. In our model,

the threshold is item-specific: it is the inverse of the virality of the item times a global constant r. Therefore, u activates when $W_i^t(u) \geq \frac{r}{v_i}$. This implies that items with higher virality v_i are more easily adopted and thus propagate more; at the same time, if r is higher the propagation encounter more resistance.

Social pressure. In Equation 1 the social pressure experienced by u to activate on item i depends on the set of other nodes, followed by u, which already activated on item i. As different individuals might have different level of influence on their followers we define

$$p_{v,u,z} = t_{u,z} + \rho_v \cdot t_{v,z}$$

where ρ_v is a positive scalar representing the capability of node v to influence other nodes. Therefore $p_{v,u,z}$ depends on how much u is interested in topic z (i.e., $t_{u,z}$), the overall strength of influence of v (i.e., ρ_v), and how much v is interested in topic z (i.e., $t_{v,z}$).

In this setting we can consider two different configurations:

- *propagation by interest only*: this is obtained by forcing $\rho_v = 0, \forall v \in V$, that is to say that a node will activate on a certain item only based on its interests on the item topics, and the number of neighbors that already activated;
- *propagation by influence*: this is the general case with $\rho_v \geq 0, \forall v \in V$, where a node will activate more easily when the item has been adopted by highly-influential neighbors.

Initial activations. The propagation model requires a set of nodes being active on an item $i \in I$ at time zero. We assume two possible mechanism to account for this:

- *Endogenous activation.* Each item is introduced by one specific node inside the network. Specifically, given an item with topic distribution γ_i, we assign it to the closest node in the network based on its interests:

$$\hat{u} = \arg\max_u \gamma_i \mathbf{t}_u$$

- *Exogenous activation.* Alternatively, we can think of items as something that is generated *outside* the network and not by one of its nodes. In this case, we model the external environment as *one dummy node which is followed by every other node* $v \in V$. Then, the set of initial activators is a natural consequence of the propagation model described above.

Generating items. We next describe how WoMG generates items—their level of virality and their topic distribution. Based on observations of real-world cascades, we generate virality levels by a power-law Pareto distribution with exponent λ with the following probability density:

$$P(v) = \frac{\lambda}{v^{\lambda+1}} \tag{2}$$

The exponent λ determines how likely it is to generate items with a high virality, and it is called *virality exponent*.

To generate the topic distribution of items $i \in I$ we use LDA as a purely generative model and therefore generate γ_i from a Dirichlet distribution with parameter α (the prior topic distribution). Note that the integration with LDA implies that, having a pre-trained LDA topic model (i.e., each topic is in turn defined by a distribution over a vocabulary of terms) we can easily generate also textual content, more precisely a bag of words, for each item $i \in I$ by sampling from the topic model according to the same γ_i we use for propagations.

Parameter	Explanation
k	Number of topics.
$G = (V, E)$	Directed social graph.
$\mathbf{t}_u$	Interest vector of node u.
ρ_u	Influence capability of u.
γ_i	Topic distribution of item i.
v_i	Virality of item i.
r	Virality resistance.
$p_{v,u,z}$	Strength of influence by v on u on topic z.
$W_i^t(u)$	Pressure on node u at time t for item i.

Table 1: Variables of the model.

Table 1 summarizes the key parameters and notation.

4 Generating Interest Vectors

In the propagation model described in the previous section, a key role is played by the interest vector of each node. In this section we discuss how to generate realistic interest vectors, given a known network structure and w.r.t. a tunable level of homophily. Our goal is two-folded: first, to generate interest vectors exhibiting realistic level of homophily; second, to be able to tune the level of homophily, in order to simulate different scenarios. This is the main technical challenge of our work.

4.1 Defining homophily

Traditionally, homophily is defined in terms of a single attribute (McPherson, Smith-Lovin, and Cook 2001), e.g. gender, ethnicity, age, etc. and it is measured as the number of outbound ties with users who share similar attribute values, divided by the overall number of outbound ties (Bisgin, Agarwal, and Xu 2012).

In our context, as we deal with a vector of continuous values, i.e., our definition of interests $\mathbf{t}$, we define homophily as the ratio between the average similarity among connected nodes and the average similarity among disconnected nodes. Since the similarity we wish to measure is based on continous-valued interest vectors, we choose to use the cosine similarity.

More formally, we define a metric measuring how much a given pair network-vectors $(G, \mathbf{t})$ is homophilic – that is, the extent to which the vectors $\mathbf{t}$ are similar among linked nodes and dissimilar in others. Let us consider the set of edges E and non-edges $\overline{E} = (V \times V \setminus E)$. Then, we need a measure of similarity between nodes $\delta : V \times V \to [0, 1]$.

$$\delta(u, v) = \frac{\mathbf{t}_u \cdot \mathbf{t}_v}{\|\mathbf{t}_u\| \|\mathbf{t}_v\|} \tag{3}$$

Hence, our homophily metric will be given by

$$h_\delta(E, \overline{E}) = \frac{|\overline{E}| \cdot \sum_{u,v \in E} \delta(u,v)}{|E| \cdot \sum_{u,v \in \overline{E}} \delta(u,v)} \quad (4)$$

Since our vectors will be non-negative, this measure ranges in $[0, \infty]$. In particular, when there is no homophily, $h_\delta(E, \overline{E}) = 1$, and the average similarity between nodes that have edges in E is the same as in $\overline{E}$. If $h_\delta(E, \overline{E}) > 1$, we are in the presence of homophily, which might be arbitrarily large, depending on the characteristics of the network. When $h_\delta(E, \overline{E}) < 1$, then similarity in non-edges is higher than similarity among connected nodes, and we enter the regimen of heterophily, which is not interesting for our purposes.

We next describe a method which is able to generate different level of homophily. First, we show how to achieve maximum levels of homophily, then we show how our method can generate interest vectors that go from minimum to maximum homophily.

4.2 Maximizing homophily

To generate a set of interest vectors $\mathbf{t}_u \in \mathbf{T}$ that maximizes homophily according to the measure defined in Eq. 4, we need to have a combination of high $\delta(u,v)$ if $(u,v) \in E$, and small $\delta(u,v)$ if $(u,v) \in \overline{E}$. Assuming that all vectors in $\mathbf{T}$ are normalized, and $\mathbf{A}$ is the adjacency matrix of G, then the homophily metric defined in Eq. 4 can be defined as:

$$\frac{\|\mathbf{A} \times \mathbf{T}\mathbf{T}^\intercal\|_1}{\|(1 - \mathbf{A}) \times \mathbf{T}\mathbf{T}^\intercal\|_1} \quad (5)$$

where $\|\|_1$ is the L_1 norm that in this case corresponds to the sum of all matrix elements. The value of this metric can be maximized by ensuring that $\mathbf{T}\mathbf{T}^\intercal \approx \mathbf{A}$. This can be solved through a symmetric non-negative matrix factorization (NMF) scheme (Lee and Seung 2001), where we minimize the following expression:

$$\min_T \|\mathbf{A} - \mathbf{T}\mathbf{T}^\intercal\|_F^2, \quad (6)$$

where $\|\|_F^2$ is the Frobenius norm, subject to the constraint that the values in $\mathbf{T}$ must be nonnegative.

Through this optimization problem, we ensure that the similarity among connected nodes $\mathbf{A} \times \mathbf{T}\mathbf{T}^\intercal$ is maximized (by approximating their values to 1), and the similarity among disconnected nodes $(1 - \mathbf{A}) \times \mathbf{T}\mathbf{T}^\intercal$ is minimized (by approximating their values to 0). However, in Eq. 5, the similarity among connected nodes has the same weight as the similarity among disconnected nodes, while in Eq. 6, if the number of edges in E is much smaller than the number of elements in $\overline{E}$ – which is the case for most real networks – the NMF optimization of Eq. 6 will give more weight to disconnected nodes (zero values in $\mathbf{A}$).

To balance the weights of connected and disconnected nodes, we increase the number of non-zero values in the factorized matrix with the following approach: In addition to the first-order proximity edges represented in $\mathbf{A}$, we consider also the second-order proximity (Tang et al. 2015; Wang et al. 2017), defined as the cosine similarity of the rows of the adjacency matrix:

$$\mathbf{S}_{i,j} = \frac{\mathbf{A}_u \cdot \mathbf{A}_v}{\|\mathbf{A}_u\| \|\mathbf{A}_v\|} \quad (7)$$

The second-order proximity matrix captures the amount of neighbors shared by each pair of nodes. Then, to compute the $|V| \times k$ matrix $\mathbf{T}$, a combination of $\mathbf{A}$ and $\mathbf{S}$ is used as input to the following NMF minimization problem:

$$\min_{T,U} \|\mathbf{A} + \eta\mathbf{S} + \beta\mathbf{R} - \mathbf{T}\mathbf{T}^\intercal\|_F^2, \quad (8)$$

where $\|\|_F^2$ is the Frobenius norm, subject to the constraint that the values in $\mathbf{T}$ must be nonnegative. $\mathbf{R}$ is a random matrix drawn uniformly in $[0, 1]$; finally, η is a parameter that controls the weight of second-order connections and β controls the degree of randomness to add in order to tune the level of homophily represented in $\mathbf{T}$. With β very large, the random matrix $\mathbf{R}$ will dominate as input to the NMF, so the resulting homophily value $h_\delta(E, \overline{E})$ is minimal. Both η and β can be modulated to tune the level of homophily and to create interest vectors that exhibit realistic cascading properties. The nonnegative factorization is achieved using the *projected gradient method* with sparseness constraints, as described in (Lin 2007; Hoyer 2004).

4.3 Other Methods

We also experimented with other methods for generating interest vectors: methods based on label propagation (Zhu and Ghahramani 2002), and node embedding methods, namely *Node2vec* (Grover and Leskovec 2016). However, since these baselines are not aimed to maximize homophily, the level of homophily achieved by them are not comparable to the factorization approach. In the next we define each of these alternative methods, and we report results in Section 5.

Label propagation. The *Continuous Label Propagation Algorithm* (CLPA) is an intuitive approach to assign similar interest vectors to nodes that are connected. We achieve this goal through an iterative method with the following steps:

1. Select a distribution $\mathcal{D}$ to assign interests vectors to each node (e.g. a Dirichlet distribution). Initialize the vector of interests $\mathbf{t}_u \in \mathbb{R}^k$ of each node u with a sample from $\mathcal{D}$. Vectors are normalized, since they are selected from $\mathcal{D}$.
2. Select from V a set of m "influencers", $\mathcal{M}$. For this we use a simple greedy heuristic approximating the set of m nodes that are maximally far from each other, thus avoiding influencers that are close to each other. We start adding a random node to $\mathcal{M}$, and greedily keep adding one node with the maximum shortest-path distance to all nodes already in $\mathcal{M}$, until we reach the prefixed size m.
3. For a given number of iterations, propagate the interests of each node u to their neighbors v. In each iteration, for each node $u \in G$, we update the interest of a neighbor v in the topic z as follows:

$$t_{v,z} = \frac{t_{v,z} + \alpha t_{u,z}}{\sum_k^K t_{v,k} + \alpha t_{u,k}} \quad (9)$$

where α is:

$$\alpha = \begin{cases} 0 & \text{if } v \in \mathcal{M} \\ 0.01 & \text{if } v \notin \mathcal{M} \wedge u \notin \mathcal{M} \\ 0.5 & \text{if } v \notin \mathcal{M} \wedge u \in \mathcal{M} \end{cases} \quad (10)$$

The denominator in Eq. 9 is just for normalization, and the values of α are chosen in such a way that influencers can change significantly the interests of their neighbors and never change their own interests, while the other nodes have a much smaller impact on their neighbors interests.

The parameters of this method are the number m of influencers and the number of iterations in Step 3. The different values assigned to α have lesser impact, but with smaller values, a higher number of iterations is needed to converge to the same degree of homophily, while larger values make the convergence unstable. With the given values for α, a low number of iterations results in lower homophily, and vice-versa, a high number of iterations results in higher homophily. The number of influencers has an impact in the converged homophily value, as discussed in Session 5.

Node2vec. *Node2vec* (Grover and Leskovec 2016) is a semi-supervised method which learns continuous feature representations in a k-dimensional subspace for nodes in a network, using second order random walks (Perozzi, Al-Rfou, and Skiena 2014). The random walk on a non-weighted graph is defined as follows. Consider a walk that just crossed edge $t \to v$; then, the probability of crossing the edge $v \to x$ is:

$$\pi_{vx} = \begin{cases} 1/p & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ 1/q & \text{if } d_{tx} = 2 \end{cases} \quad (11)$$

where d_{tx} is the shortest path distance between node t and x. Its parameters control depth and breadth of the walk, in particular: p controls the probability of immediately revisiting a node in the walk; q allows the walk to move further away from node t. *Noise-contrastive estimation* (NCE) (Mnih and Kavukcuoglu 2013) is then used to learn vectors that allow to distinguish between such random walks and random sets of nodes.

We modified the cost function of the algorithm in order to better suit our needs. First, since our goal is to generate *interests*, we want the embeddings to be non-negative, for interpretability. Second, we wish to be able to control the distribution of the resulting vectors. We solve these problems by introducing two additive terms in the loss function. For the former, we add a penalty factor if the embeddings are negative; for the latter, a KL-divergence term with a prior distribution to avoid components entanglement (Higgins et al. 2017). Hence, defining $\mathbf{T}$ as the $|V| \times k$ matrix of interests we want to generate, the new loss function is

$$L(\mathbf{T}) = J(\mathbf{T}) + s(\| - \min(0, \mathbf{T})\|) + \beta \cdot D_{KL}(P(\mathbf{T}) \| \pi)$$

where $J(\mathbf{T})$ is the NCE loss, s is the soft sign function $\frac{x}{|x|+1}$, $P(\theta)$ is the observed distribution of the embeddings and π is the prior distribution. We empirically tested several prior distributions and parameters, and we found the best results using a symmetric Beta distribution.

5 Experimental Assessment

In the previous sections, we presented WoMG our model for generating realistic interest-driven propagations, starting from a network structure and a topic model. WoMG is developed as a free, open-source, Python 3 library.[1] The software architecture is characterized by a division in blocks implemented using abstract classes. This architecture, together with the compatibility with other standard libraries, such as e.g., NetworkX[2], makes the library easily extensible, for instance, by implementing other propagation models. The library is available to the research community to be used both for generating synthetic datasets to test inference or prediction tasks on propagation data; but also as a basis for studying, *in vitro*, how different conditions result in different propagations, allowing users to implement their own assumptions into the model.

The *tunability* of the model is therefore of primary importance: in order to adapt to different real scenarios, the model needs to be able to generate datasets with different characteristics by controlling a small set of "knobs", in a way that can lead to predictable results. For instance, a user of WoMG might want to simulate a propagation dataset where (i) the level of homophily is low, and (ii) each cascade propagates in depth w.r.t. to its starting location. Our goal is to allow such a user to generate this scenario by tuning a small set of interpretable hyper-parameters.

To asses these abstract goals, we formalize three research questions to be explored in this section:

- **Q1**: Can we obtain both low and high levels of homophily (w.r.t. the interest vectors of the nodes) through a controllable parameter?
- **Q2**: Is there a small set of interpretable hyper-parameters that can tune the macroscopic characteristics of the generated data?
- **Q3**: Can the model generate propagations from a given real network that are similar in shape to real cascades?

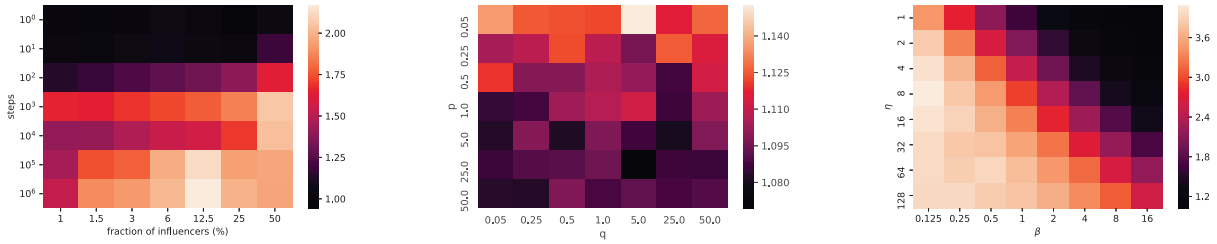
5.1 Generating interests

In many contexts, connected nodes can be very similar in their interests. In other kinds of networks, instead, there is no significant difference between connected or random pairs of nodes. We wish to be able to generate both scenarios through a single parameter that controls the generation of interests in our model. In other words, we want to obtain a high or a low homophily depending on a given parameter. Therefore, our goal is to find whether one of the presented techniques is able to obtain both low and high levels of homophily, in a simple and controllable way.

We evaluate the results of the methods mentioned in Section 4 with the metric defined in Equation 4.

[1] https://github.com/FedericoCinus/WoMG

[2] https://networkx.github.io/



(a) CLPA. Parameters: fraction of influencers, number of steps.

(b) Node2vec Parameters: p and q.

(c) NMF Parameters: η, β.

Figure 2: Homophily heat map for interest generation through the three different proposed methods. The two axis represent the parameters of each method. Color represents the obtained homophily measured through Equation 4. Each square in each heat map is the average of 10 realizations.

In the next, we show to which extent the parameters of each presented method allow to tune the resulting homophily. For each method (CLPA, node2vec and NMF) we vary their parameters to understand their relationship with the observed homophily, and compare the range of homophily that is achieved by each method. Finally, we choose the method with wider range, and define a parameter $H \in [0, 1]$ that ranges from a minimum to a maximum homophily.

To measure results, we generated a graph according to a generative model from (Klemm and Eguiluz 2002), that is able to generate graphs with high clustering coefficient and scale free degree distribution, two common characteristics of real networks. We set the graph to $N = 200$ nodes and $M = 400$ edges, a probability $\mu = 0.01$ of rewiring and we set the number of topics $k = 10$. Then, we tested the three methods by generating vectors from each one according to different parameters. For each combination of parameters, we generated ten experiments and took the average of the homophily metric we defined.

CLPA. For CLPA, we vary the fraction of nodes selected as influencers and the number of steps. We show the resulting homophily in Figure 2(a). Starting from a minimum number of two influencers (1% of 200 nodes), we observe that higher fractions result in higher homophily. It starts decreasing again when more than 25% of nodes are selected as influencers, due to the high fraction of nodes with fixed interests. Homophily also increases with a larger number of steps, clearly saturating after a certain number of steps. The average maximum homophily in 10 realizations is 2.15, achieved by fixing he fraction of influencers to 12.5% and the number of steps to 10^6, as shown in Figure 2(a).

Node2vec. For node2vec, we vary p and q, the parameters of the node2vec algorithm that define how breadth-first or how depth-first the random walks are. We chose to vary both in a wide scale from 0.05 to 50 – unusual for this algorithm – to search for variability in terms of observed homophily. We present the results in Figure 2(b). We observe that the range of homophily values for all combinations of p and q is very narrow: from a minimum of 1.08 to a maximum of 1.14. Also there is no clear trend in the level of homophily when varying p and q, although combinations of high q and low p tend to produce higher levels of homophily than combinations of high p and low q. Intuitively, when q is high and p is low, the random walks tend to involve a small community instead of exploring the whole network, and thus nodes belonging in the same community are more likely to be assigned similar interests.

Matrix factorization. For NMF method, we vary η and β, the coefficients of the two matrix that are summed to the adjacency matrix. While η defines the amount of higher-order paths and thus its effect is dependent on the specific network, the effect of β is very clear and allows to tune from randomness to homophilic vectors.

We present results in Figure 2(c). The behavior of β is as we expected, and allows to diminish the high amount of homophily that this method can generate. Instead the behavior of η is less clear, and we found that its effect might be different depending on the specific network under consideration.

We observe that the range of homophily levels achieved through this method is wider in comparison to the other two: a maximum of 4.2, against 2.15 using CLPA and 1.14 using Node2vec. With the right choices of η and β, this method allows for a better tunability of homophily to generate interests vectors. To tune the level of homophily using a single parameter, we reduce the parameter space to a single parameter $H \in [0, 1]$. Therefore, to define H, we fix $\eta = 8$ and we define a linear dependence on β; specifically, $\beta = 16 - 15.875H$. With $H = 0$ we achieve the lowest level of homophily, and with $H = 1$ the highest level.

5.2 Parameters analysis

We have defined a method for generating interests vector from a given network with a tunable parameter H. Now that we have fully defined our model, we turn to our second question: is there a small set of interpretable hyper-parameters that can tune the macroscopic characteristics of the generated data? In this subsection, we explore the range of properties for the synthetic propagation data that the model can generate, understanding more of their relationship with the input hyper-parameters.

We explore this relationship in the different configurations

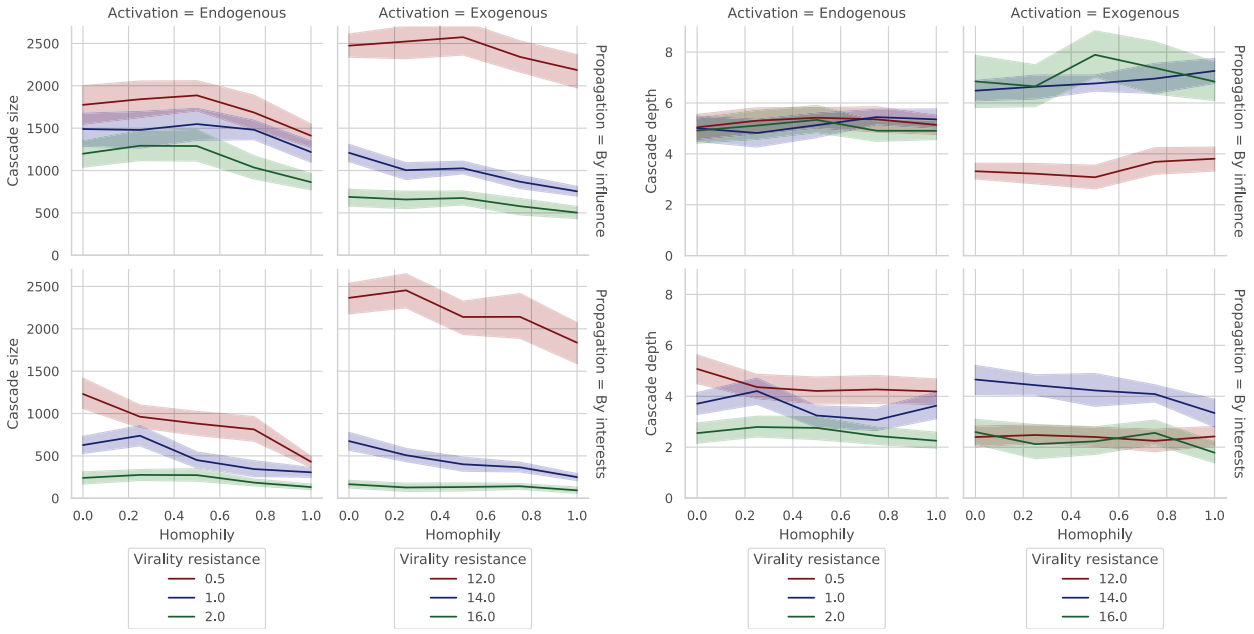


Figure 3: For each of the four configurations, average cascade size (left) and average cascade depth (right) obtained by the model with respect to the model hyper-parameter H that controls homophily. Different lines show different values for the virality-resistance (hyper-parameter r). Around each line, the shaded area indicates 95% confidence interval obtained by bootstrapping over 20 different experiments.

we presented in Section 3. Each configuration represents a set of assumptions on the propagation dynamics. We have two binary distinctions.

The first distinction is on the initial condition for each propagation. As discussed in Section 3, we have two different settings for initial activations. In the first one (*endogenous* activation) each item is initially created by one specific node in the network (the most interested one); in the second one (*exogenous* activation) each nodes receive the same stimulus at the beginning, and therefore each item is initially adopted by all the nodes sufficiently interested.

The second distinction represents whether the strength of influence is the same regardless of the spreader, or if different nodes have different influencing capabilities. In one case (propagation by interest, $p_{v,u,z} = t_{u,z}$), the propagation is guided only by the interests of the receiver. In the other (propagation by influence, $p_{v,u,z} = t_{u,z} + \rho_v \cdot t_{v,z}$), the propagation is guided *both* by the interests of the receiver and by the influencing capabilities of the spreader.

From these two binary choices, we obtain four different configurations for our model. Each configuration describe a different set of assumptions on the data. For instance, on the retweet network of Twitter, the initial activation must be represented as endogenous, since one node inside the network will be the single initial adopter.

Inside each configuration, we have two hyper-parameters to tune the properties of the generated cascades. The first parameter H controls the homophily of interests, defined in Section 4. The second one r (*virality-resistance*) controls the magnitude of the threshold and so the general resistance to propagation.

Observed propagation properties. Our goal is to show that by changing the two hyper-parameters H and r, we can obtain a wide range of scenarios in the generated propagations. Also, we aim at defining the relationship between hyper-parameters and the obtained properties in a clear, interpretable way.

To show this we need to define the macroscopic properties of the generated cascades being measured. To characterize the generated cascades, in fact, we choose the following two properties:

1. *Average cascade size*. The average number of activated nodes across all items:
$$\frac{1}{|I|} \sum_{i \in I} \sum_{v \in V} A(i, v)$$
where $A(i, v) = 1$ if v adopted item i and 0 otherwise.
2. *Average cascade depth.* The number of time steps between the last and the first activation of our model. This can also be seen as the depth of the propagation cascade, as in WoMG there are new activations at each timestamp, and the first timestamp in which there are no activations represents the end of the cascade.

Our goal is to show that in each configuration we can tune the resulting properties of the data through our two parameters. Therefore we run our model on a fixed graph (generated according to (Klemm and Eguiluz 2002)) of $N = 200$ nodes, with $k = 10$, and we varied H and r for each of the four configurations.

Results. Results are reported in Figure 3. These results show how one can obtain any wanted value for the macroscopic

properties we measure under any configuration by changing the two hyper-parameters of the model H and r. In particular, we observe the following:

- The relationship between the virality-resistance (defined by r) and the cascade size is very clear under each configuration: lower virality-resistance leads to larger cascades. This follows intuition and provides an easy way to control the obtained cascade size.
- Homophily have a clear cut effect on cascade size, and the relation appears to be monotonic: lower homophily leads to larger cascades; higher homophily leads to smaller cascades. This can be explained by noting that when the interests are distributed among nodes following homophily, the phenomenon of *echo chambers* arises. Inside echo chambers, an item will usually attract only nodes inside the echo chamber that it is interested to it, stopping as soon as it reaches the barrier. Hence, cascade sizes are limited.
- With propagation by interest, cascades tend to collapse on a few initial time steps—a *bursty* dynamics. Since the spreading depends only on individual interests, items cannot reach nodes that are far in the topic space. The different activation settings, instead, produce a divergence in the cascade size curves. In fact, for the exogenous configuration, cascades can collapse on a few initial time step when the virality resistance is low, otherwise they can reach a great depth when r is high. For the endogenous activation, instead, cascade depths are mostly stable.
- Symmetrically, while viralities affect more cascade size, they do not seem to have a profound effect on cascade depth. However, we can note that (a) in the exogenous activation, lower virality-resistance leads to short cascades: the bursty dynamics is amplified. All nodes get stimulated by the environment, and they all get activated as soon as they receive a viral item from it. Instead, when (b) the activation is endogenous, in a low virality-resistance setting (with influence-driven propagation) items spread along longer cascades. A node creates a viral content and this content is able to slowly reach far nodes in the network.

These observations, firstly, confirm that our model is able to reproduce a wide range of scenarios. It can be used with different assumptions on the spreading of items in the network. For each set of assumptions, its two main hyper-parameters (virality and homophily) allow to calibrate the macroscopic properties of the generated propagations. Secondly, these observations also show how our model can be used to better understand how phenomena such as echo chambers, viral contents, information spread behave.

5.3 Real data

In this section, we show how our model can be used to produce synthetically a data set similar to a given one (RQ3). For this experiment, we take a real-world propagation data set known as Digg 2009 (Lerman and Ghosh 2010). From this data set, we keep in the graph only the nodes with at least 100 activations; after this filtering step, we also remove singletons from the network. The resulting data set has 3482 nodes and 64519 edges (average degree 18.5). The number of items is 3553 as the original data set.

We considered the graph as input of our model and we compared the real cascades and the synthetic ones, derived as WoMG's outputs. From now on, we refer to the cascades extracted from the real dataset as Digg 2009 cascades, and to the cascades generated by WoMG as WoMG cascades.

Metrics. Since the time scale of Digg 2009 cascades is not comparable to the time scale of the synthetic propagations produced by WoMG cascades, we look only at the structure of the cascades. For this, we represent the cascades as directed acyclic graphs (DAGs). Given the sequence of nodes activations, the edges are created w.r.t. the temporal order. That is, a directed edge (u, v) will belong to the DAG_i of an item i, if (1) the edge (u, v) exists in the social graph (so that v has visibility of u's activity) and (2) u activated on i *before* v – so that v can see u's activation on i and be influenced, or in other terms, the item i can propagate from u to v. More formally, the nodes of the DAG_i of an item i are all the nodes that activated on that item; its edges are:

$$E_{DAG_i} = \{\forall (u, v) \in E_G : t_i(u) < t_i(v)\} \qquad (12)$$

Then, we can define the depth of a cascade i as the diameter of its DAG_i:

$$\text{depth}_i = \max_{d_{DAG_i}(u,v)<\infty} d_{DAG_i}(u, v)$$

where $d_{DAG_i}(u, v)$ is the distance between u and v on DAG_i.

Therefore, we look at these two metrics of a propagation data set: the size and the depth of its cascades. To better characterize the *shape* of cascades, we also look at the ratio between its size and its depth: a large size-to-depth ratio corresponds to "flat" cascades, with many nodes activating after a small set of influential nodes; a small size-to-depth ratio corresponds to "tall" cascades, with nodes activating one after the other in long chains.

Experimental setting. We choose the model configuration based on the data set semantics: an activation in this data set corresponds to a user voting a *story* on the social bookmarking website Digg. Users can see stories from their neighbors or from an external source: because of this, we choose the *exogenous* activation setting. Since users might have different influencing capabilities, we choose the *propagation by influence* configuration accordingly; for the influence capabilities, we draw each ρ_v at random from a Pareto distribution with exponent 2.

On this data set, external sources play a large role: nodes often activate without any of their neighbors activating before. For this reason, we set a large value (12.0) for the influence capability ρ_v of the dummy node representing the external environment.

Our hypothesis here is that our model allows to shape the resulting cascades by tuning the parameters we studied before, i.e. the *virality* of the items and the *homophily* of the interests of the nodes. In particular, we wish to discover

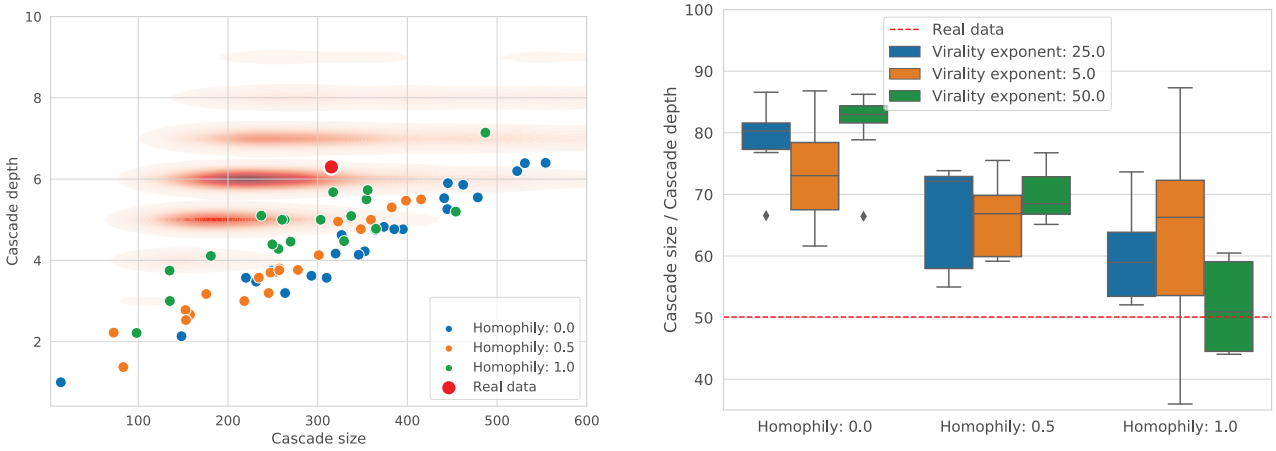


Figure 4: Comparison of the Digg 2009 data set with the simulated cascades for 3 values of homophily, 3 values of virality, and 3 run for each setting. Left: Mean cascade size and mean cascade depth for the DIGG 2009 data (in red) and each WoMG run; in red the KDE distribution of all the real cascades from the DIGG 2009 data set. Right: Distribution of the size-to-depth ratio of the cascades generated by each WoMG parameter setting; in red, we show the mean ratio of the DIGG 2009 cascades.

whether by tuning these two parameters we are able to generate a data set that is similar to the given one.

To do so, we chose three values for the virality exponent (5, 25 and 50) and three values for the homophily parameter H (0, 0.5, 1). For each of these 9 parameter settings, we realize 6 different experiment with 200 items each. In each experiment, we generate a different set of interests and of items at random.

Results. To analyze results, we first directly compare the average values for our metrics (cascade depth and size) on the generated data set against their real distribution. This is shown in Figure 4 (left). From this plot, we show how the parameter settings are able to obtain a variety of different cascade depth and cascade size, that largely overlap with the real distribution.

In particular, the effect of homophily is essential to change the cascade shape: a low homophily corresponds to *flat* cascades – large size and low depth – while a high homophily leads to *tall* cascades. This confirms previous experiments and respect the intuition that in environments with highly specialized communities the propagation follows longer paths; instead, if the links between nodes do not follow their interests, the propagation is mostly driven by the virality of the items, obtaining flat cascades.

The real data set average is close to the high homophily setting: we can therefore conjecture that on this data set, users follow each other mostly as a consequence of common interests. The role of the homophilic interests generation is therefore very important to achieve realistic results. In addition, to reach even closer results, further research would be needed to generate even more homophilic interests on a real given social network.

To investigate the *shape* of the cascades obtained by our model against real data, we analyze in detail how the size-to-depth ratio changes in the different parameter settings. In Figure 4 (right), we report the ratio obtained by each setting, along with the real average value of Digg 2009 cascades. Here, it is clear that the homophily plays an important role in shaping the cascades. This figure also confirms how the setting that most closely reproduces the Digg 2009 cascades corresponds to a high virality and a high homophily.

5.4 Propagations from real documents

We also used a real corpus of documents to produce a topic distribution of items and to check if similar items in the topic space produce cascades with similar set of nodes. In principle, since nodes have different interests and these depend on the topics, similar items in the topic space should activate similar sets of nodes. We considered the Associated Press data set: a corpus of 2246 document with a vocabulary of 10473 terms (Harman 1992). We inferred the topic distribution of each item using LDA (Blei, Ng, and Jordan 2003) model with 4 topics, and we generated a cascade for each item of the corpus on a synthetic graph created according to the generative model presented in section "Generating interests" (Klemm and Eguiluz 2002) ($N = 500$ nodes and $M = 1000$ edges, a probability $\mu = 0.01$ of rewiring). We set the parameters of WoMG to $homophily = 0.5$ and $virality - exponent = 32$ and we collected the sets of activated nodes.

For each pair of items we measure the cosine similarity of their topic distribution, and the Jaccard similarity of their set of activated nodes. In Figure 5, for each pair of items we plot the cosine similarity of the pair in the x-axis and the Jaccard similarity on the y-axis. Darker colors correspond to more nodes activated in WoMG on at least one of the two documents (minimum 2 and maximum $N = 500$).

The plot shows that items that are close in the topic space (i.e., cosine similarity greater than 0.5), have an average Jaccard similarity statistically greater than those with lower cosine similarity. We can observe that few items with high vi-

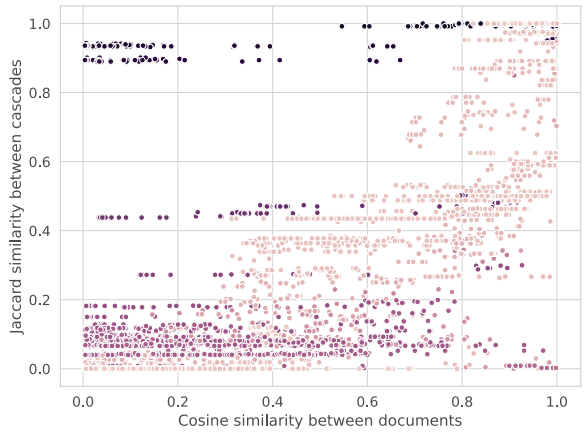


Figure 5: Similarity of documents in the topic space versus similarity of cascades. In this plot, each point represents a pair of real documents for the AP data set: on the X axis, we show the cosine similarity between their topic distributions; on the Y axis, the Jaccard similarity between the cascades generated by WoMG with these documents. Darker colors correspond to more nodes activated in WoMG on at least one of the two documents (minimum 2 and maximum $N = 500$). We note that there is a linear relationship between the topic similarities and the Jaccard similarities, albeit some documents were highly viral and were propagated by most nodes in the graph, resulting in a high Jaccard similarity despite having different topics.

rality activate almost the whole of the 500 nodes and as such, when paired, these items result to have Jaccard close to 1, regardless their distance in the topic space.

6 Conclusions and Future Work

We presented WoMG, a model for the synthetic generation of information cascades in social media. In our model the memes propagating in the social network are characterized by a probability distribution in a topic space, accompanied by a textual description. Similarly, every individual is described by a vector of interests defined over the same topic space. Information cascades are governed by the topic of the meme, its level of virality, the interests each node, community pressure, and social influence. By adjusting a small set of interpretable hyper-parameters our model can tune the macroscopic characteristics of the generated data and obtain realistic propagations and interests for a given network structure and topic model.

In our future work we plan to extend WoMG in several directions and specializations. The first extension is towards the analysis of how information propagation and debates on social media may affect people's opinion, strengthening or weakening echo-chambers, possibly leading to stronger polarization or cyberbalkanization (Chan and Fu 2017). This requires assigning to memes and to each person, besides topics, a polarity or an opinion, and integrating the propagation model with an opinion dynamics model. The second extension deals with the veracity of the information propagating in the network. The idea is to have different generative models for fake and genuine memes, as well as different tendency for users to spread or block the propagation of fake memes. By studying how high virality memes interact with the different echo chambers, we could understand more on the mechanisms of misinformation spreading and designing mitigation strategies.

References

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 207–214.

Anagnostopoulos, A.; Kumar, R.; and Mahdian, M. 2008. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 7–15.

Aslay, Ç.; Lu, W.; Bonchi, F.; Goyal, A.; and Lakshmanan, L. V. S. 2015. Viral marketing meets social advertising: Ad allocation with minimum regret. *PVLDB* 8(7):822–833.

Bakshy, E.; Rosenn, I.; Marlow, C.; and Adamic, L. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st International Conference on World Wide Web*, 519–528.

Barbieri, N.; Bonchi, F.; and Manco, G. 2013a. Cascade-based community detection. In *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013*, 33–42.

Barbieri, N.; Bonchi, F.; and Manco, G. 2013b. Topic-aware social influence propagation models. *Knowledge and information systems* 37(3):555–584.

Bisgin, H.; Agarwal, N.; and Xu, X. 2012. A study of homophily on social media. *World Wide Web* 15(2):213–232.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Bonchi, F. 2011. Influence propagation in social networks: A data mining perspective. *IEEE Intell. Informatics Bull.* 12(1):8–16.

Borge-Holthoefer, J., and Moreno, Y. 2012. Absence of influential spreaders in rumor dynamics. *Physical Review E* 85(2):026116.

Chan, C.-h., and Fu, K.-w. 2017. The relationship between cyberbalkanization and opinion polarization: Time-series analysis on facebook pages and opinion polls during the hong kong occupy movement and the associated debate on political reform. *Journal of Computer-Mediated Communication* 22(5):266–283.

Cheng, J.; Adamic, L.; Dow, P. A.; Kleinberg, J. M.; and Leskovec, J. 2014. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, 925–936. ACM.

Cheng, J.; Adamic, L. A.; Kleinberg, J. M.; and Leskovec, J. 2016. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web*, 671–681.

Del Vicario, M.; Scala, A.; Caldarelli, G.; Stanley, H. E.; and Quattrociocchi, W. 2017. Modeling confirmation bias and polarization. *Scientific Reports* 7:40391.

Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; and Flammini, A. 2016. The Rise of Social Bots. *Commun. ACM* 59(7):96–104.

Gleeson, J. P.; Cellai, D.; Onnela, J.-P.; Porter, M. A.; and Reed-Tsochas, F. 2014. A simple generative model of collective online behavior. *Proceedings of the National Academy of Sciences* 111(29):10411–10415.

González-Bailón, S.; Borge-Holthoefer, J.; Rivero, A.; and Moreno, Y. 2011. The Dynamics of Protest Recruitment through an Online Network. *Scientific Reports* 1:197.

Goyal, A.; Bonchi, F.; and Lakshmanan, L. V. 2010. Learning influence probabilities in social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, 241–250.

Goyal, A.; Bonchi, F.; and Lakshmanan, L. V. S. 2011. A data-based approach to social influence maximization. *PVLDB* 5(1):73–84.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* 855–864.

Harman, D. 1992. Overview of the first text retrieval conference (trec-1). In *TREC*, volume 1992, 1–20.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR* 2(5):6.

Hoyer, P. 2004. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research* 5:1457–1469.

Klemm, K., and Eguiluz, V. M. 2002. Growing scale-free networks with small-world behavior. *Physical Review E* 65(5):057102.

Lancichinetti, A.; Fortunato, S.; and Radicchi, F. 2008. Benchmark graphs for testing community detection algorithms. *Physical Review E* 78(4):046110.

Lane, S. N. 2011. The tipping point: How little things can make a big difference. *Geography* 96:34.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 556–562.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Lin, C. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19(10):2756–2779.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 27(1):415–444.

Mehmood, Y.; Barbieri, N.; Bonchi, F.; and Ukkonen, A. 2013. CSI: community-level social influence analysis. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013*, 48–63.

Mnih, A., and Kavukcuoglu, K. 2013. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in neural information processing systems*, 2265–2273.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710.

Richardson, M., and Domingos, P. 2002. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 61–70.

Sasahara, K.; Chen, W.; Peng, H.; Ciampaglia, G. L.; Flammini, A.; and Menczer, F. 2019. On the Inevitability of Online Echo Chambers. *arXiv:1905.03919 [physics]*.

Schelling, T. C. 1969. Models of segregation. *The American Economic Review* 59(2):488–493.

Tang, J.; Qu, M.; Wang, M.; Zhang, M.; Yan, J.; and Mei, Q. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, 1067–1077.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *Science* 359(6380):1146–1151.

Wang, X.; Cui, P.; Wang, J.; Pei, J.; Zhu, W.; and Yang, S. 2017. Community preserving network embedding. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Watts, D. J. 2002. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences* 99(9):5766–5771.

Weng, L.; Ratkiewicz, J.; Perra, N.; Gonçalves, B.; Castillo, C.; Bonchi, F.; Schifanella, R.; Menczer, F.; and Flammini, A. The role of information diffusion in the evolution of social networks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 356–364.

Weng, L.; Menczer, F.; and Ahn, Y.-Y. 2014. Predicting successful memes using network and community structure. In *Eighth international AAAI conference on weblogs and social media*.

Yuan, Y.; Alabdulkareem, A.; et al. 2018. An interpretable approach for social network formation among heterogeneous agents. *Nature communications* 9(1):4704.

Zhu, X., and Ghahramani, Z. 2002. Learning from labeled and unlabeled data with label propagation. Technical report.