ORIGINAL ARTICLE

Modeling cascade formation in Twitter amidst mentions and retweets

Soumajit Pramanik[1] · **Qinna Wang**[2] · **Maximilien Danisch**[3] · **Jean-Loup Guillaume**[4] · **Bivas Mitra**[1]

Received: 30 November 2016 / Revised: 17 July 2017 / Accepted: 13 August 2017

Abstract This paper presents an analytical framework for cascade formation considering both retweet and mentioning activities into account. We introduce two mention strategies (a) random mention and (b) smart mention to model the mention preferences of the users. The proposed framework $\mathcal{C}_F^M$ analytically computes the cascade size, depicting tweet popularity and discovers the presence of a critical retweet rate, under which mentioning in a tweet significantly helps in cascade formation. We validate the proposed framework with the help of Monte Carlo simulation; we demonstrate the generality of the framework taking both empirical and synthetic follower networks into consideration. This framework proves the elegance of smart mention strategy in boosting tweet popularity, specially in the low retweeting environment.

Keywords Multiplex network · Epidemic models · Information diffusion · Twitter

1 Introduction

In recent times, Twitter has become one of the most influential micro-blogging systems for spreading and sharing breaking news, personal updates and spontaneous ideas (González-Bailón et al. 2011). In Twitter, propagation of tweets or hashtags from one user to another occurs mainly via two modalities: “retweeting” and “mentioning” (Kato et al. 2012). In case of retweet, information is simply relayed to all the followers of the retweeting user. However, mention utility allows to spread an information far beyond the neighborhood and improves its visibility by making it available to the appropriate set of users. Furthermore, as mentions get listed in a separate tab, they gain higher attention than regular posts. Admittedly, mention utility has a potential to play a significant role in the cascading behavior of tweets and hashtags. For instance, in our dataset, we observe that the probability that a mentioned user retweets a post is on average 32% higher than the one of a follower.

Retweet cascades play a central role in tweet popularity, measured as the retweet count (Kupavskii et al. 2012; Cheng et al. 2014). Figure 1 demonstrates the heterogeneity in tweet popularity in the ‘Algeria’ dataset (described in the next section). This heterogeneity stems from the fact that cascade formation in Twitter is a complex process; the popularity of a tweet widely depends on the interplay of two propagation modalities, *retweet* and *mention*. Mentioning suitable users in a post may expose the tweet to a larger population, hence mention strategy surely plays an important role behind tweet popularity

✉ Soumajit Pramanik
soumajit.pramanik@cse.iitkgp.ernet.in

Qinna Wang
qinna.wang@gmail.com

Maximilien Danisch
maximilien.danisch@gmail.com

Jean-Loup Guillaume
jean-loup.guillaume@univ-lr.fr

Bivas Mitra
bivas@cse.iitkgp.ernet.in

[1] Department of Computer Science and Engineering, IIT Kharagpur, Kharagpur, India

[2] Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu, 75005 Paris, France

[3] Telecom Paris Tech, Paris, France

[4] L3I, University of La Rochelle, La Rochelle, France

Published online: 28 August 2017

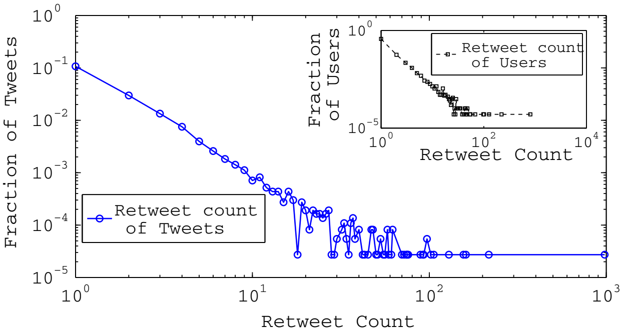


Fig. 1 Retweet count distribution of all tweets in 'Algeria' dataset; *Inset* depicts the retweet count distribution of all users in 'Algeria' dataset

(Wang et al. 2013; Tang et al. 2014; Gong et al. 2015; Pramanik et al. 2016). On the other side, social network of followers is a dynamic communication medium that facilitates the propagation of tweets. Precisely, the *retweeting activity* of users, through the underlying follower network, relays the post to a new population. Study shows that the retweet activity is heavily skewed across the users (see inset of Fig. 1). Notably, retweeting a post depends on its importance to the viewer; the importance can be characterized as high contextual similarity, mentioning the right user in that post etc. In a nutshell, the *retweet rate* and the underlying *follower network* may appear as a decisive factor in the cascade formation. In this context, investigating the role of retweet and mention activities in promoting tweet propagation is an interesting research question. Modeling the cascade formation process, considering these two modalities into account, can be the first step towards this direction.

The dynamics of the cascade formation in Twitter have been studied in the light of the classical diffusion models such as linear threshold (Granovetter 1978; Kempe et al. 2003) and independent cascade models (Goldenberg et al. 2001). Recently, the evolution of text mining in Twitter led to deep analysis of the topical retweet cascades and virality prediction (Cheng et al. 2014; Kwak et al. 2010; Suh et al. 2010). There exists stochastic models (epidemic models (Li et al. 2013) and Markov chains (Dickens et al. 2012a)), machine learning (Xu and Yang 2012; Gupta et al. 2012) and point process-based models (Zhao et al. 2015; Bao et al. 2015) to explain the retweeting activities and modeling the cascade formation. However, all these endeavors have only focused on the tweet propagation through retweets; they failed to shed any light on the role of other modalities, such as *mentioning*. Moreover, they lack the proper understanding of the underlying dynamics, such as retweet rate and formation of follower links, behind the tweet diffusion.

Recently, diffusion in multiplex network has received considerable attention where different modes of information propagations coexist and contribute to the spreading process (Boccaletti et al. 2014; Kivelä et al. 2014). Indeed, multiplex network exhibits an excellent platform for explaining tweet propagation through both retweet and mentioning modalities. Several attempts have been made in bits and pieces to investigate the epidemic outbreak across multiple layers (Cozzo et al. 2013; Li et al. 2015), competing epidemic propagations in multilayer networks (Darabi Sahneh and Scoglio 2014), epidemic vs. awareness propagation (Granell et al. 2013) etc. Notably, there exists a subtle difference between the propagation of tweets via retweet activity (where tweet propagates through the underlying follower network) against the propagation via mentioning activity (where propagation relies on the user's mention preferences). The absence of underlying network in case of mention prohibits the direct adaptation of the aforementioned multiplex epidemic models.

In this paper, we propose an analytical framework ($\mathcal{C}_F^M$) to model the cascade formation, considering both retweet and mentioning activities into account. We introduce a *'generic'* mention strategy to model the mention preferences in a tweet, with two special cases—random and smart mentioning. The key idea behind the $\mathcal{C}_F^M$ model is the multilayer representation of the tweet propagation, where the tweet propagates via retweets in one layer and via mentioned users in another layer. Our data study highlights the importance of mentions on tweet popularity and clearly demarcates the retweeting behavior of two different types of users (a) normal followers (b) mentioned users in the tweet (Sect. 2). We claim that the influence of mentioning on cascade formation directly reflects in the (a) cascade size, interpreted as retweet count and (b) critical retweet rate for cascade formation. Subsequently, the proposed model analytically computes the cascade size and reveals the complementary effect of followers and mentioned users in cascade formation; it derives the critical retweet rate under which mentioning in a tweet significantly helps in cascade formation (Sect. 3). We then develop a Monte Carlo simulation platform for model analysis and validation (Sect. 4). Moreover, we establish the correctness of this simulation platform from the empirical data and show that the cascade size and critical retweet rate, computed from the analytical model, exhibits a nice agreement with simulation (Sect. 5). We also observe that proper mentioning proves to be more beneficial in propagating an information in case of a less active retweeting environment (which is commonly the case for Twitter). Furthermore, we demonstrate the generality of the framework for both empirical (Algeria, Egypt datasets) and synthetic (scale-free, Kronecker) follower networks and study the influence of mention strategies on cascade formation (Sect. 6).

2 Dataset and representation

In this section, we introduce the dataset and propose a multiplex framework to demonstrate the flow of tweets. Additionally, we perform few motivational experiments to establish the importance of mention utility on the spread of tweet.[1]

2.1 Dataset

Arab-Spring Movement was a wave of anti-government demonstrations and protests which took place in a substantial number of middle-east countries in 2011. During this events, Twitter was used extensively to propagate news and opinions. We collected two publicly available tweet datasets[2] connected to these events—(1) "Algeria" Dataset is a collection of around 60K tweets (tweet-ids) and 20K users who posted them during the 'Algeria movement'. We also crawled the tweet content, user profile and the corresponding follower network. (2) "Egypt" dataset is a collection of around 2.6 million tweets (tweet-ids) posted during 'Egypt uprising'. We crawled the tweet content of 0.2 million posts and the profile details and follower network of around 60K users who posted them.

2.2 Multiplex representation

For a given hashtag '#h', the multiplex representation contains two layers: the bottom one represents tweet propagation via follow links, the top one via mention links (Fig. 2). More precisely, all users who tweet '#h' appear as a node in the bottom (follow) layer. A directed link connects user 'A' to 'B' if 'A' (re)tweets '#h' before 'B' further retweets and 'B' is a follower of 'A'. In the top (mention) layer a directed link connects 'C' to 'D' if 'C' tweets '#h' before 'D' further retweets and 'C' *mentions* 'D' in her post ('D' may or may not be a follower of 'C'). Any user can appear in both layers.

A closer look reveals that both layers are essentially collections of directed acyclic graphs (DAG). We denote the root of each DAG as an initiator since they are responsible for initiating the spreading process. We can identify two classes of initiators, the 'true initiators' and the 'dummy initiators'. A *true initiator* of '#h' is a user who is a root in a Follow or a Mention DAG but never appears as non-root member of any DAG. These users have actually started the spreading process (for '#h') as a result of some external influences. A *dummy initiator* is a user who is a root in a follow DAG but a non-root member of a mention DAG. Basically, a dummy initiator gets the information from some other user via mention and subsequently initiates the spreading process to its followers.

2.3 Dissecting mentioning

We demonstrate the importance of mention utility from three different perspectives; first, the boost it provides to the tweet popularity, second, the higher attention a mentioned post receives for retweeting and finally, the distribution of number of mentions in a tweet.

1. *Boost in tweet popularity*: In the multiplex representation, we measure the impact of mentioning on the popularity of hashtags. Let us define the popularity of a hashtag as the number of (re)tweets it receives. We select few popular hashtags (tweets) for which we estimate the popularity reduction by dropping mentions. In this estimation, first we find the dummy initiators (set D) for a hashtag '#h' and all the users (set S) who only belong to the DAGs rooted by dummy initiators. Obviously the retweet activity of the $S \cup D$ users is dependent on the mention layer. If hashtag '#h' is tweeted by total n users, then *mention dependency* of '#h' can be measured as $\frac{(|S \cup D|)}{n}$. Looking at the most popular hashtags (tweets) in Fig. 3a, b we observe that such hashtags (tweets) are heavily mention dependent.
2. *Higher retweet rate*: Mention utility improves the tweet visibility by making it available to the appropriate set of users. Since mentioned posts get enlisted in a separate tab, they gain higher attention than regular posts. Figure 4 demonstrates the fact that most of the people tend to respond to mentioning with higher propensity than the normal retweet. Hence, this result exhibits the presence of two different retweet rates, one for mentioned users (mean 0.048 for 'Algeria' dataset) and another for normal followers (mean 0.006 for 'Algeria' dataset).
3. *Frequency of mentions*: Figure 5 shows that the number of mentions per post exhibits a skewed distribution. Almost 95% of the tweets are limited to just one or two mentions whereas the average number of mentions per tweet is limited to 2.

The aforementioned observations point to the key features that highlight the importance of mentioning on

[1] Notably, in this section, we perform experiments on both hashtags and tweets. This is because both 'hashtag' and 'tweet' are commonly used as an unit of information in Twitter and the propagation dynamics can be assumed to be mostly identical for both of them.

[2] http://dfreelon.org/2012/02/11/arab-spring-twitter-data-now-available-sort-of/.

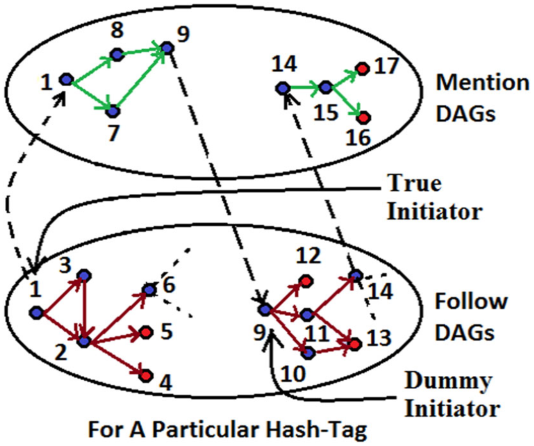


Fig. 2 Example of mention-follow multiplex

cascade formation. Henceforth, we concentrate on the following two indices to characterize mentioning, (a) cascade size (b) critical retweet rate for cascade formation. Cascade size captures the impact of mentioning on the overall retweet count. On the other hand, interplay of critical retweet rate for normal followers and mentioned users reveal their complementary effect in cascade formation. The analytical framework proposed in this paper and the subsequent validations revolve around these two indices.

3 Developing $\mathcal{C}_F^M$ framework

In this section, we leverage on the multiplex network representation and propose an analytical framework to model the cascade formation considering both mention and retweet activities. We introduce the mention strategies used to develop the framework. As a direct outcome of the framework, we compute the cascade size and critical condition for the cascade formation for different mentioning scenarios.

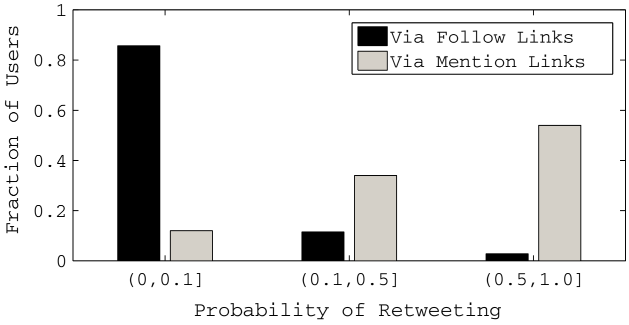


Fig. 4 Distribution of users' probabilities of retweeting a tweet received via follow links and mention links in 'Algeria' dataset

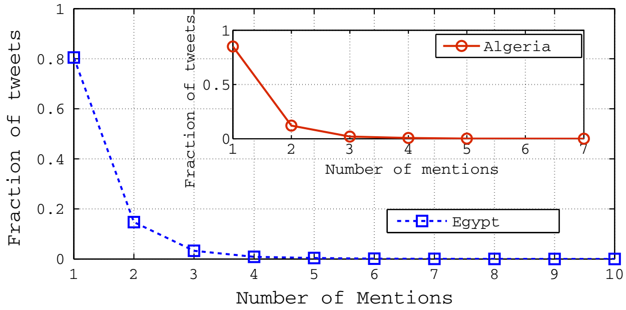


Fig. 5 Distribution of number of mentions for tweets containing at least one mention in 'Egypt' dataset; *Inset* shows the same for 'Algeria' dataset

3.1 Model description

The data study and representation proposed in Sect. 2.2 exhibit the diffusion of tweets following two modalities, via retweets and mentioning. In order to develop $\mathcal{C}_F^M$ to model this dynamics, we rely on the SIR (Susceptible-Infected-Recovered) epidemic model where tweets spread

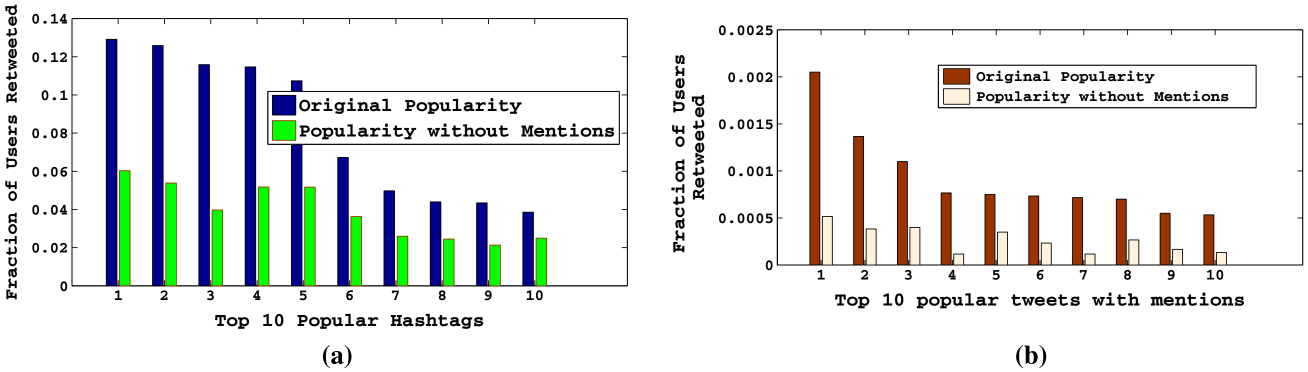


Fig. 3 Mention dependency for tweets and hashtags in 'Algeria' and 'Egypt' datasets. **a** Popularities (number of times posted) of top 10 popular hashtags in 'Algeria' dataset with and without Mentions. **b** Popularities (retweet counts) of top 10 popular tweets (containing mentions) in 'Egypt' dataset with and without Mentions

through both mention and follow links of the mention-follow multiplex network (see Table 1). The infection of a node v by tweet T is governed by three factors, (a) v has to be exposed to T. A node v may be exposed to T by a node u in two different modalities: if u posts T and v is a follower of u, or if v is not a follower of u but u mentions v while posting T. This determines the structure of the multiplex network (see Fig. 2) (b) v has to show interest in T. The interest of v in T depends on whether it has been exposed through mention or follow. (c) v must have a suitable retweet (or activity) rate, i.e., even being exposed to an interesting tweet, v can decide not to retweet it.

Initially, all the nodes are in the susceptible state. A node v is said to be infected by a tweet T if it retweets T in the next timestep. A user once infected gets recovered instantaneously in the next timestep. In our system, we assume the propagation of multiple posts (tweets) to be independent of each other and consider the propagation of only one tweet at a time which any node can tweet / retweet only once (due to Twitter restrictions). If in case the same tweet is posted multiple times, each of them is considered separately as independent tweets. At time t, if a susceptible user v on the top (mention) layer gets mentioned by an infected user, it switches its state to mentioned-susceptible (*MS*); rest of the susceptible nodes on the follow (bottom) layer remain at the unmentioned-susceptible (*US*) state. In the next timestep $(t+1)$, both kinds of nodes are allowed to retweet and switch to infected (*I*) state. However, there is a possibility that at time $t+1$ one user v does not get infected and hence switch back to susceptible (*S*) state. This event enables v to be mentioned or unmentioned in the next subsequent steps.

In Fig. 6a we explain the states of our proposed model with an example. Suppose, getting influenced from some external source, user 1 posts a tweet T and mentions users 4 and 5 in her tweet. As a result of this event, user 1 changes her state to infected (*I*), users 2 and 3 become unmentioned-susceptible (*US*) as they received the tweet only from the follow link and users 4 and 5 become mentioned-susceptible (*MS*). Among the unmentioned and mentioned-susceptible users, only user 5 decides to retweet and becomes infected (*I*) in the next timestep. Others return back to their original susceptible (*S*) state. User 1 gets transformed into a recovered (*R*) node who will not get infected (retweet the same tweet *T*) anymore. We use the Microscopic Markov Chain Approach (Arenas et al. 2010; Granell et al. 2013) to mathematically represent our $\mathcal{C}_F^M$ model. The state transition equations are derived from the scheme shown in Fig. 6b which describes the transition probability of the states in the multiplex framework (Fig. 6a).

Table 1 Mapping the terminologies of epidemic propagation and tweet propagation

Epidemic propagation	Tweet propagation
Susceptible	Users yet to post any tweet or retweet
Getting infected	Tweeting/retweeting a post
Infected individual	User who tweets/retweets a post

3.2 Model parameters

Let $A=(a_{vz})$ be the adjacency matrix that describes the follow network in a population of size N. For a user v, f_v be the follower count and α_v and β_v denote the probabilities of getting infected via mention and follow links, respectively;[3] α and β denote the respective average probabilities over the population N. λ_v^T denotes the number of mentions used by user v while posting tweet T whereas λ is the average number of mentions per tweet. On the follow layer, each individual v has a certain probability of being in one of the three states at time t, denoted by $p_v^S(t)$, $p_v^I(t)$ and $p_v^R(t)$, respectively. Similarly, on the mention layer, for each susceptible individual v, the probabilities of being in mentioned-susceptible (*MS*) and unmentioned-susceptible (*US*) states are denoted by $p_v^M(t)$ and $1-p_v^M(t)$, respectively. On the other hand, for the unmentioned (US) and mentioned (MS) individuals, the probabilities of staying at the susceptible state without being infected are expressed as $q_v^{US}(t)$ and $q_v^{MS}(t)$. Figure 6a provides an overview of the state transitions.

3.3 Mention strategies

We model mention strategies following which a user u can be chosen for mentioning in a tweet. We introduce a *'generic'* mention strategy where user u is chosen preferentially to her $(f_u \times \alpha_u)^\theta$ score where $(\theta \in [0,1])$ is a tunable parameter, f_u is the follower count and α_u is the retweet rate of u. The intuition behind this strategy stems from the fact that mentioning a person in a tweet can only be beneficial if the mentioned user retweets the post and enables the tweet to be exposed to a large population (via followers). This motivates us to preferentially mention a user u with high retweet rate α_u as well as a high follower count f_u. Notably, the extent of preference can be regulated with parameter θ as the mentioning strategy followed by one individual may vary widely, from $\theta=0$ ('random' mention) to $\theta=1$ ('smart' mention). This immediately leads to the following two special cases

[3] As inspired from the datastudy (Sect. 2.3), we keep two separate retweet rates for follow and mention.

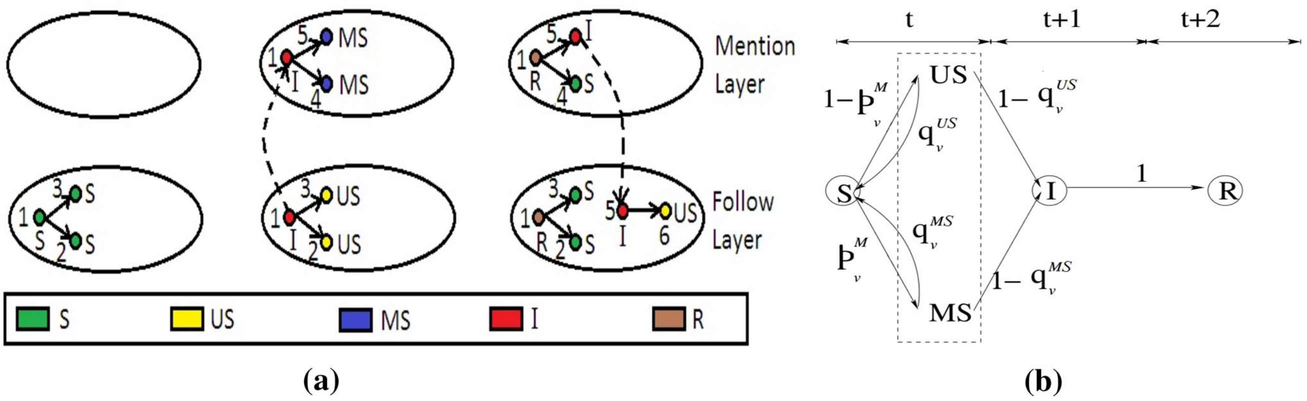


Fig. 6 Analytical representation of the $\mathcal{C}_F^M$ model—an example in (**a**) and the schematic diagram in (**b**). **a** State transition in the multiplex representation. **b** Transition probabilities of the states in the $\mathcal{C}_F^M$ model.

A. *Random mention*: The user u is chosen uniformly at random from the set of all susceptible users.

B. *Smart mention*: The user u is chosen preferentially to her $f_u \times \alpha_u$ score. Evidently, the main objective of the 'smart' strategy is to maximize the expected number of users exposed to that tweet.

In an alternate interpretation, $0 < \theta < 1$ captures the situation where one fraction of the population applies random mention strategy, and the rest of the population relies on the smart mention.

3.4 Analytical representation of $\mathcal{C}_F^M$ framework

In this section, we analytically represent the framework and compute the cascade size and critical condition of cascade formation. For the sake of generality, we first consider the 'generic' mention strategy for computation, and subsequently derive the equations for the specific cases of 'random' and 'smart' mention strategies.

3.4.1 Computing the cascade size

We first estimate the cascade size following generic mention strategy for a population of N users. Let the total number of individuals infecting neighbors at time t be denoted by $I(t)$, that is, $I(t) = \sum_v p_v^I(t)$. The users to be mentioned in a post are chosen following the generic mention strategy. Thereby, the probability that a (susceptible) individual v is in mentioned state at time t is

$$p_v^M(t) = 1 - \left[1 - \frac{(f_v\alpha_v)^\theta}{\sum_v (f_v\alpha_v)^\theta}\right]^{\lambda I(t)} \tag{1}$$

In order to keep the calculations tractable, we approximate α_v as α; we note the error introduced due to this approximation in Sect. 5. Therefore

$$p_v^M(t) = 1 - \left[1 - \frac{(f_v\alpha)^\theta}{\sum_v (f_v\alpha)^\theta}\right]^{\lambda I(t)} \tag{2}$$

Subsequently,

$$p_v^M(t) = 1 - \left[1 - \frac{f_v^\theta}{\sum_v f_v^\theta}\right]^{\lambda I(t)} \tag{3}$$

Using linear approximation, we obtain

$$p_v^M(t) \approx \lambda I(t) \frac{f_v^\theta}{\sum_v f_v^\theta} \tag{4}$$

Since all susceptible users are either mentioned or unmentioned, given an individual v at time t, the probability to be in the mentioned-susceptible state $p_v^{MS}(t)$ is the product of the probabilities of being mentioned and susceptible. The same applies for probability of being in the unmentioned-susceptible state $p_v^{US}(t)$. Hence,

$$\begin{aligned} p_v^{MS}(t) &\approx \lambda I(t) \frac{f_v^\theta}{\sum_v f_v^\theta} p_v^S(t) \\ p_v^{US}(t) &\approx \left(1 - \lambda I(t) \frac{f_v^\theta}{\sum_v f_v^\theta}\right) p_v^S(t) \end{aligned} \tag{5}$$

The transition probability for a susceptible individual v not to be infected by any neighbor through a follow link is $r_v(t) = \Pi_z(1 - a_{zv} p_z^I(t)\beta)$ and the probability for v not to be infected through mention link is $(1 - \alpha)$. Hence the transition probabilities for an individual v, at unmentioned-susceptible $(q_v^{US}(t))$ or mentioned-susceptible $(q_v^{MS}(t))$ state, not to be in infected state are

$$\begin{aligned} q_v^{US}(t) &= r_v(t) \\ q_v^{MS}(t) &= r_v(t)(1 - \alpha) \end{aligned} \tag{6}$$

By using Eqs. 5 and 6, we can develop the Microscopic Markov Chains for the epidemic spreading process for each node v:

$$p_v^S(t+1) = p_v^{US}(t)q_v^{US}(t) + p_v^{MS}(t)q_v^{MS}(t) \tag{7}$$

$$p_v^I(t+1) = p_v^{US}(t)\left(1 - q_v^{US}(t)\right) + p_v^{MS}(t)\left(1 - q_v^{MS}(t)\right) - p_v^I(t) \tag{8}$$

$$p_v^R(t+1) = p_v^R(t) + p_v^I(t) \tag{9}$$

Applying iterative numerical methods we calculate the fraction of infected individuals at any point of time; this eventually provides us the final cascade size. We denote the total fraction of all individuals infected during the diffusion process as R_U.

3.4.2 Critical condition for epidemic outbreak

Epidemics are defined as outbreaks that infect a nonzero fraction of the population in the limit of large-scale system. Epidemic behavior is usually described as a phase transition from a regime without an outbreak to one with an outbreak, with the parameters of the model (Newman 2002). In this subsection, we derive the critical values of retweet rate α and β during this phase transition. Let us first assume the existence of a critical point β_c^g for fixed parameters α and λ, i.e., the epidemic will die out if $\beta < \beta_c^g$. The calculation of this critical point is performed by considering that when $\beta \rightarrow \beta_c^g$, the probability of nodes being infected $p_v^I \approx \epsilon_v \ll 1$. The smaller the probability of nodes being infected, the faster the epidemic dies out.

Consequently, $q_v^{US} \approx 1 - \beta\sum_z a_{vz}\epsilon_z$, $q_v^{MS} \approx 1 - \alpha - \beta\sum_z a_{vz}\epsilon_z$. Inserting them in Eq. 8, we then obtain

$$\begin{aligned}\epsilon_v &= \left(p_v^{US} + p_v^{MS}\right)\beta\sum_z a_{vz}\epsilon_z + \alpha p_v^{MS}\\ &= p_v^S\beta\sum_z a_{vz}\epsilon_z + \alpha\lambda I(t)\frac{p_v^S f_v^\theta}{\sum_v f_v^\theta}\end{aligned}$$

when $p_v^I(t) \rightarrow 0$, then $p_v^S(t) \rightarrow 1$, then we obtain

$$\sum_v\left[a_{vz} - \left(1 - \alpha\lambda\frac{f_v^\theta}{\langle f^\theta\rangle}\right)\frac{\delta_{vz}}{\beta}\right]\epsilon_v = 0 \tag{10}$$

where $\langle f^\theta\rangle$ is the average of follower-counts to the power θ, δ_{vz} are the elements of the identity matrix such that $\delta_{vz} = 1$ if $v = z$; otherwise, $\delta_{vz} = 0$. We can rewrite the solution of Eq. 10 into the form: $A = (1 - \frac{\alpha\lambda f_v^\theta}{\langle f^\theta\rangle})\frac{I_N}{\beta}$ where $I_N = (\delta_{vz})$. According to the Perron–Frobenius theorem (Pillai et al. 2005), the vector A is equal to the vector I_N only if $(1 - \frac{\alpha\lambda f_v^\theta}{\langle f^\theta\rangle})\frac{1}{\beta}$ is equal to the maximum eigenvalue of A denoted by $\Lambda_{\max}(A)$. Hence,

$$\beta \times \Lambda_{\max}(A) + \alpha\lambda\frac{f_v^\theta}{\langle f^\theta\rangle} = 1 \tag{11}$$

Equation 11 is the key equation of our analysis which can be used to derive critical values of both parameters α and β. Considering that f_v in range $[f_{\min}, f_{\max}]$, the critical value β_c^g lies in range

$$\left(\frac{1 - \alpha\lambda\frac{f_{\max}^\theta}{\langle f^\theta\rangle}}{\Lambda_{\max}(A)}, \frac{1 - \alpha\lambda\frac{f_{\min}^\theta}{\langle f^\theta\rangle}}{\Lambda_{\max}(A)}\right) \tag{12}$$

Similarly, using the same equation, the ranges of critical value α_c^g can be determined as

$$\left(\frac{1 - \beta\Lambda_{\max}(A)}{\lambda\frac{f_{\max}^\theta}{\langle f^\theta\rangle}}, \frac{1 - \beta\Lambda_{\max}(A)}{\lambda\frac{f_{\min}^\theta}{\langle f^\theta\rangle}}\right) \tag{13}$$

Note that, for $\beta > \frac{1}{\Lambda_{\max}(A)}$, α_c^g does not exist and for $\alpha > \frac{\langle f^\theta\rangle}{\lambda f_{\min}^\theta}$, β_c^g does not exist.

Intuitive Justification of Eq. (11): Basically, the terms in the left-hand side of Eq. 11 represents the contribution of follow links and mention links on the total number of infected users. The infection via follow links depends on two factors—(1) the density of the follow network (represented by $\Lambda_{\max}(A)$) and (2) the probability of getting infected via follow links, i.e., β. Similarly, the infection via mention links depends on the average number of mentions per tweet, i.e., λ and the probability of getting infected via mention links, i.e., $\alpha\frac{f_v^\theta}{\langle f^\theta\rangle}$. For both the terms, if any of the factor be 0, the contribution of that part will be nullified regardless of how big the other factor is. For example, if β is 0, even if $\Lambda_{\max}(A)$ is very high, there will not be any infection via follow links and vice versa.

3.5 Special case: random mention

Next, we deal with the specific case of random mention strategy ($\theta = 0$), leveraging on the aforementioned framework.

3.5.1 Computing the cascade size

For random mention strategy, the users to be mentioned in a post are chosen at random. Therefore, substituting $\theta = 0$ in Eq. 4, the probability that a (susceptible) individual v is in mentioned state at time t is

$$p_v^M(t) = \frac{\lambda I(t)}{N} \tag{14}$$

Similarly, from Eq. 5, the probabilities of being at mentioned-susceptible and unmentioned-susceptible states are the following,

$$\begin{aligned} p_v^{MS}(t) &= p_v^M(t) p_v^S(t) = \frac{\lambda I(t)}{N} p_v^S(t) \\ p_v^{US}(t) &= (1 - p_v^M(t)) p_v^S(t) = \frac{N - \lambda I(t)}{N} p_v^S(t) \end{aligned} \tag{15}$$

We substitute $p_v^{MS}(t)$, $p_v^{US}(t)$ in Eqs. 7, 8 and 9 and numerically calculate the final cascade size R_U.

3.5.2 Critical condition for epidemic outbreak

The key equation Eq. 11 gets reduced to a much simpler form (as $\theta = 0$) for random mention, which is,

$$\beta \times \Lambda_{\max}(A) + \alpha\lambda = 1 \tag{16}$$

Accordingly, the critical α and critical β can be derived as

$$\alpha_c^r = \frac{1 - \beta \Lambda_{\max}(A)}{\lambda} \tag{17}$$

$$\beta_c^r = \frac{1 - \alpha\lambda}{\Lambda_{\max}(A)} \tag{18}$$

3.6 Special case: smart mention

In the following, we derive the cascade size and the critical conditions for epidemic outbreak in case of smart mention strategy ($\theta = 1$). Evidently, unlike random mention, here the probability of getting mentioned is not identical for all the susceptible users.

3.6.1 Computing the cascade size

In smart mention, the probability that a susceptible node v is mentioned at time t is (replacing θ by 1 in Eq. 4),

$$p_v^M(t) \approx \lambda I(t) \frac{f_v}{\sum_v f_v} \tag{19}$$

Hence, the probability that any individual v is at mentioned-susceptible state at time t, can be approximated as,

$$p_v^{MS}(t) \approx \lambda I(t) \frac{f_v}{\sum_v f_v} p_v^S(t) \tag{20}$$

Similarly, the probability of being at unmentioned-susceptible state at any point can be approximated as

$$p_v^{US}(t) \approx \left(1 - \lambda I(t) \frac{f_v}{\sum_v f_v}\right) p_v^S(t) \tag{21}$$

We substitute $p_v^{MS}(t)$, $p_v^{US}(t)$ in Eqs. 7, 8 and 9 and numerically calculate the final cascade size R_U.

3.6.2 Critical condition for epidemic outbreak

In case of smart mention, the key equation Eq. 11 takes the following form,

$$\beta \times \Lambda_{\max}(A) + \alpha\lambda \frac{f_v}{\langle f \rangle} = 1 \tag{22}$$

where $\langle f \rangle$ is the average follower count. Considering that f_v is in range $[f_{\min}, f_{\max}]$, the critical value β_c^s thus in range

$$\left(\frac{1 - \alpha\lambda \frac{f_{\max}}{\langle f \rangle}}{\Lambda_{\max}(A)}, \frac{1 - \alpha\lambda \frac{f_{\min}}{\langle f \rangle}}{\Lambda_{\max}(A)} \right) \tag{23}$$

Similarly, the ranges of critical value α_c^s can be determined as

$$\left(\frac{1 - \beta\Lambda_{\max}(A)}{\lambda \frac{f_{\max}}{\langle f \rangle}}, \frac{1 - \beta\Lambda_{\max}(A)}{\lambda \frac{f_{\min}}{\langle f \rangle}} \right) \tag{24}$$

Clearly, unlike random mention, here we provide upper and lower bounds for the critical retweet rates.

4 Experimental setup

In this section, we develop the Monte Carlo simulation setup to validate the $\mathcal{C}_F^M$ model and describe the parameter settings we adopt for different experiments. Additionally, we introduce the empirical and synthetic follow networks to prove the versatility of the framework.

4.1 Simulation setup and parameter settings

Initially, all the nodes are in the susceptible state. A node v is said to be infected if it tweets/retweets in the next timestep. Every Monte Carlo simulation result presented here is an average of 500 simulations. Each simulation starts with a single infected node and stops when no more users can be infected. The model parameters used in the simulation can be summarized as follows: for a node v while posting a tweet T, we consider the number of mentions as λ_v^T and infection rate (probability of retweeting) through mention and follow links as α_v and β_v, respectively (see Table 2).

In this paper, we primarily conduct two different kinds of experiments (A) Correctness of simulation: evaluates the correctness of the simulation from the empirical data (Sect. 5.1) (B) Validation experiment: validates the analytical model with simulation (Sects. 5.2 and 5.3). We fix the parameter values from the empirical observations

Table 2 Summarizing the model parameters and their corresponding interpretations

Model parameters	Interpretation
Infection probability (via mention) α_v	Probability that v retweets a tweet T after being mentioned in it
Infection probability (via follow) β_v	Probability that v retweets a tweet T after receiving it from a followee
λ_v^T	Number of users mentioned in tweet T of user v

reported in Sect. 2.3; however, the exact parameter setting also depends on the kind of experiment (A and B) we perform. Detail follows.

4.1.1 *Number of mentions* λ_v^T

In correctness experiment (A), we assign the λ_v^T directly from the tweets posted in the empirical dataset. However, for the validation experiments (B), we rely on the distribution of number of mentions observed in Fig. 5 and assign the values for λ_v^T accordingly.

4.1.2 *Retweet rates* α_v *and* β_v

Following Cerchiello and Giudici (2016), we model the retweet rate α_v and β_v following two Poisson distributions with mean μ_1 and μ_2, respectively for mentioned and regular posts, normalized between [0,1]. Since in Fig. 4, we empirically observe that in general α_v is greater than β_v for all users, here we keep $\mu_1 \geq \mu_2$. In the correctness experiment (A), we assign the mean retweet rates μ_1 and μ_2 from the empirical dataset, as reported in Sect. 2.3. However, in the validation experiments (B), we vary the values of μ_1 and μ_2 depending on the experiment, to reveal the impact of retweet rates.

4.2 Follower networks

The follower network is a dynamic communication medium which facilitates the tweet propagation. In order to validate the generality of the analytical framework, we implement a wide variety of follower networks; from empirical to synthetic.

4.2.1 *Empirical networks*

We implement two real follower networks from ‘Algeria’ and ‘Egypt’ datasets. In the ‘Algeria’ network, we have 21,141 users and 19,802,923 directed follow links (avg. indegree 1118.1 and avg. outdegree 772.1). The largest strongly connected component of the network contains 71% of all users. Similarly, in the ‘Egypt’ network, there are 59,776 users, 5,521,949 follow links (avg. indegree 116.5 and avg. outdegree 92.4) and its largest strongly connected component consists of 74% of all users. The indegree and outdegree distributions of both networks are shown in Fig. 7.

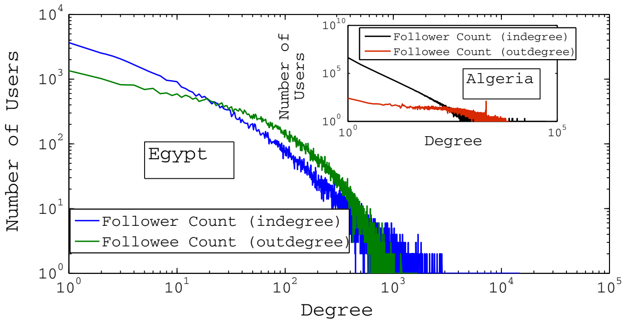


Fig. 7 Follower (indegree) and followee (outdegree) count distributions for ‘Egypt’ and ‘Algeria’ datasets

4.2.2 *Synthetic networks*

We consider the following two synthetic graphs to model the follower networks.

1. *Scale-free Networks*
   Scale-free or power law network is a popular topology to model the real social networks (Barabási and Albert 1999; Barabâsi et al. 2002; Dező and Barabási 2002). We generate power law degree distributions ($p_k \sim k^{-\gamma}$) with exponent (γ) varying as 1.3, 1.8 and 2.3. To be able to observe the effect of the exponent (γ) on the cascade size and critical retweet rates, we fix the total number of nodes as 16384 and the total edge count around 98,000 in all three networks so that the average degrees of this networks get fixed around 6. It is not very trivial to generate scale-free networks with same average degree but different exponents. Here, we utilize the generalized Barabasi–Albert’s method (Barabási and Albert 1999) for generating this scale-free networks. In this method, at each step a node enters the network with a constant outdegree and gets attached to existing nodes with probabilities proportional to $k + k_0$ where k is the indegree of an existing

node and k_0 is a constant. By varying k_0, we vary the exponent of the obtained scale-free network.

2. *Kronecker Networks*
Recently Kronecker networks (Leskovec et al. 2010; Leskovec and Faloutsos 2007) have attracted a huge attention for modeling social networks (Gomez-rodriguez and Schlkopf 2011; Du et al. 2013) as they are simple to generate, mathematically tractable, and exhibit important social network characteristics such as heavy-tailed degree and eigen-distributions, high clustering, small diameter, and network densification (Bodine-Baron et al. 2010). We generate the following three types of Kronecker graphs (Gomez-rodriguez and Schlkopf 2011; Du et al. 2013; Valera et al. 2014) with the average degree fixed around 9 (by fixing the number of nodes as 16,384 (2^{14}) and the number of edges around 145,000):

   (1) Core-periphery Kronecker network (Leskovec et al. 2008) encodes a very commonly seen core-periphery structure, in which the network consists of a large and well-connected core and a set of small peripheral communities loosely connected to the core. We generate a core-periphery Kronecker network with seed matrix [0.9, 0.5; 0.5, 0.3].
   (2) Hierarchical Kronecker network (Clauset et al. 2008) can be thought of as a structure which encodes the organizational hierarchy. For instance, in case of encoding the social links of a university with the following hierarchy—laboratories, departments and schools, hierarchical structure ensures the dense social links within the people of the same laboratory, whereas the density of links decays across different laboratories, and further falls across different departments and schools. One such hierarchical Kronecker network is generated with seed matrix [0.9, 0.1; 0.1, 0.9].
   (3) Random Kronecker network (Erdös and Rényi 1959) is generated with seed matrix [0.5, 0.5; 0.5, 0.5]—This is just simple Erdos–Renyi random network structure.

5 Evaluation of the framework

In this section, first we evaluate the correctness of the simulation setup in the light of the empirical dataset. Next, we rely on this simulation to validate the analytical framework. This is important to note that while developing the analytical model $\mathcal{C}_F^M$ in Sect. 3, we simplified few modeling assumptions through mean field approximation. For instance, instead of considering individual retweet rates α_v and β_v for each user v, we introduced mean retweet rate α and β in the model development. Similar assumption has been made for the number of mentions λ in a tweet. In the validation experiment, we demonstrate the agreement between simulation with analytical model, amidst the mean field approximations. We concentrate on two different indices—cascade size and critical retweet rates. In this validation process, we consider both empirical and synthetic follower networks and include both random and smart mention strategies.

5.1 Correctness of the simulation setup

We implement the generic mention strategy and simulate the model for each tweet on the follower network obtained from the "Algeria" and "Egypt" datasets. In order to run the simulation, we fix the model parameters α_v and β_v from empirical dataset following Sect. 4.1. Furthermore, we simulate each tweet (say, T) diffusion by starting with the same set of initiators (say, v) and keeping the same number of mentions (λ_v^T) as in the real data. We adjust θ in the setup to compute the total infected population and fix θ which results to the best agreement with empirical R_U, estimated from the dataset. We consider the retweet counts (R_U) of the tweets containing mentions in the "Algeria" and "Egypt" datasets as the evaluation metric.

In Fig. 8, we observe a nice agreement between the infected population and the real retweet count R_U estimated for both the "Algeria" and "Egypt" datasets. For most of the tweets, we fix $\theta \approx 0$, indicating that in reality, random mention strategy mostly gets followed. Nevertheless, Fig. 8 demonstrates the fact that there is ample scope to boost the retweet count R_U by choosing the users to be mentioned, smartly.

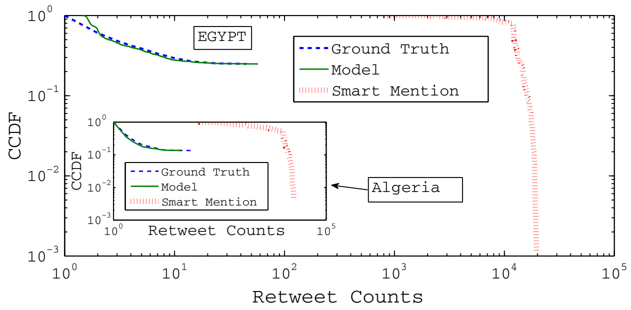


Fig. 8 Matching ground truth tweet popularities with the simulation setup (with same α_v, β_v, λ_v^T and initiator) and comparing with smart mention strategy for 'Algeria' and 'Egypt' datasets

5.2 Validation with cascade size R_U

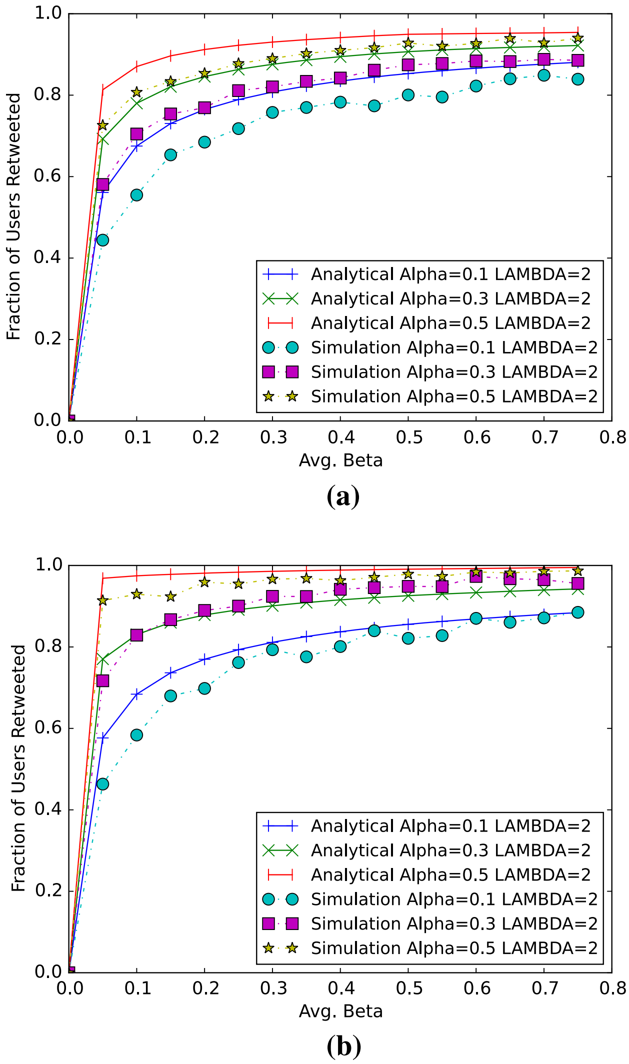
We have analytically estimated the cascade size R_U in Sect. 3. Here we first validate R_U with the empirical networks, followed by the synthetic networks. As we know, each simulation starts with a single infected node that is randomly chosen among the individuals, while in the analytical model, each individual v is initially infected with probability $p^I_v(0) = \frac{1}{N}$ (where N is the total number of individuals). Each simulation stops when no susceptible node gets infected and R_U is the average number of retweeting users during the simulations; while in the analytical calculations, $R_U = \sum p^R_v(t)$ when $I(t) < 10^{-7}$ ($p^I_v(t), p^R_v(t), I(t)$ are defined in Sect. 3).

5.2.1 *Empirical follow network*

In order to validate the cascade size R_U, first we compare it with the results obtained from the simulations on the 'Egypt' follower network. One can observe in Fig. 9 that the analytical and simulation results are in good agreement for both random and smart mentioning strategies. Importantly, in the low β region, a small increase in β results in sharp hike in R_U. This is due to the fact that an increase in β lowers $q^{US}_v(t)$ and $q^{MS}_v(t)$ in Eq. 6 which results in an increase in the fraction of infected individuals $p^I_v(t+1)$ in Eq. 8. However, in high β region, R_U attains a saturation. Side by side, the role of α to boost R_U, also becomes predominant in the low β region. Basically, as $r_v(t)$ is higher (close to 1) in that regime, a little change in the value of α impacts the value of $q^{MS}_v(t)$ significantly (see Eq. 6) which gets reflected in the fraction of infected individuals $p^I_v(t+1)$ (see Eq. 8). We note that the analytically calculated R_U upper-bounds simulations for a large range of β. This is caused by the mean field theory which assumes that events are independent (Youssef and Scoglio 2011).

5.2.2 *Synthetic follow network*

Next, we validate the cascade size on synthetic networks. We use a random Kronecker network to realize the synthetic follower network. We simulate both random and smart mention strategies for validation. In Fig. 10a, we show that the analytically estimated R_U exhibits a good agreement with simulation for random mention. However, in case of smart mention (see Fig. 10b), even if the nature of both the curves are same, the analytical estimations seem to overshoot the simulation results a bit for the high α–low β region. This is possibly due to the approximation we make in Sect. 3.4 by assuming all α_v's same as the average α.



(a)

(b)

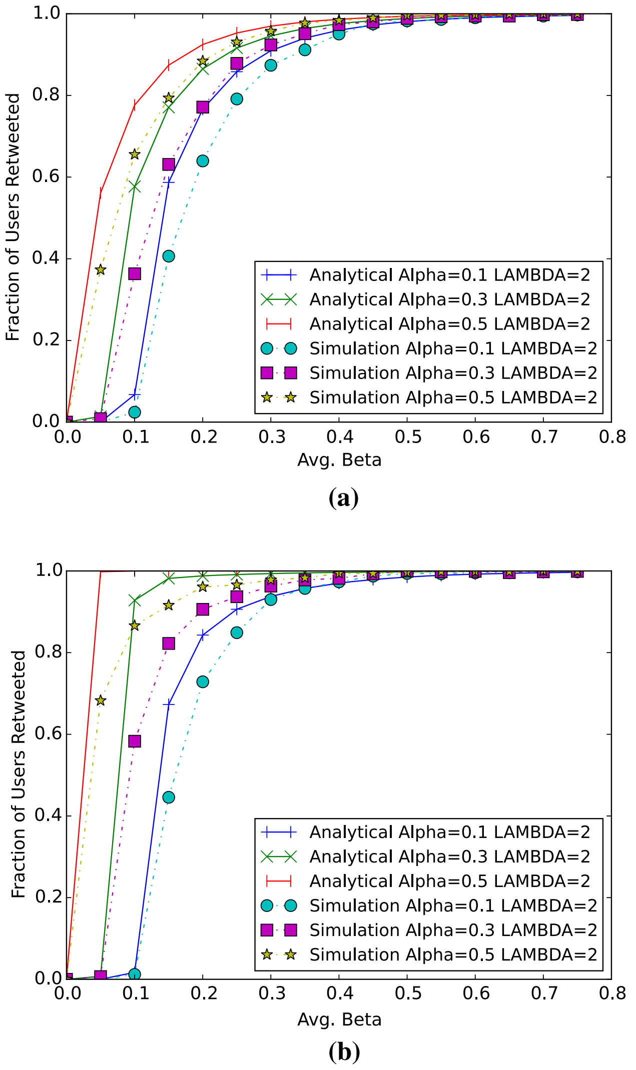
Fig. 9 Comparison of analytical $\mathcal{C}^M_F$ model and Monte Carlo simulation w.r.t. R_U for 'Egypt' network. **a** Random Mention. **b** Smart Mention

5.3 Validation with critical retweet rate

First we show the existence of the critical retweet rate for cascade formation, by simulating the model on both empirical and synthetic networks. Next, we validate this rate with analytical results derived in Sect. 3.

5.3.1 *Empirical follow network*

We implement the follower network from the 'Algeria' dataset and simulate the Monte Carlo tweet propagation on top of it. In Fig. 11, we show the presence of a critical α (and β as well) beyond which the retweet count R_U increases sharply. Next, we validate the critical retweet rate with the analytically computed threshold. In the inset of Fig. 12, we show that for random mention, the analytically



(a)

(b)

Fig. 10 Comparison of analytical C_F^M model and Monte Carlo simulation w.r.t. R_U for random Kronecker network. **a** Random Mention. **b** Smart Mention

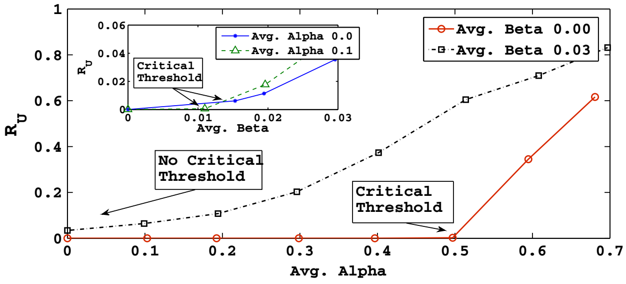


Fig. 11 Effect of varying average α and average β (see *Inset*) on R_U (random mentioning) for ‘Algeria’ dataset

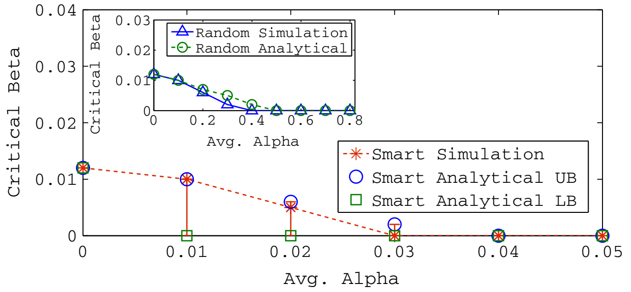


Fig. 12 Matching the epidemic thresholds obtained analytically and from simulation for random and smart mention strategies in ‘Algeria’ dataset. ‘UB’ and ‘LB’ in the figure represent the corresponding estimated upper and lower bounds, respectively

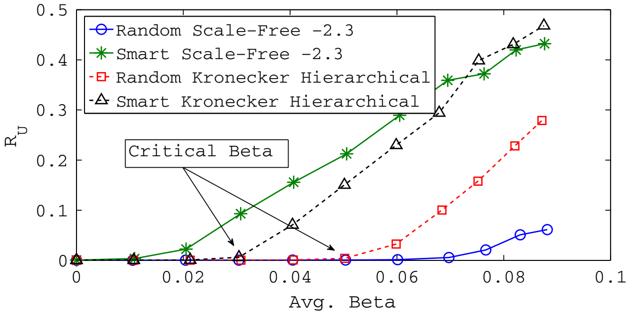


Fig. 13 Critical β for different synthetic topologies for random and smart mentions

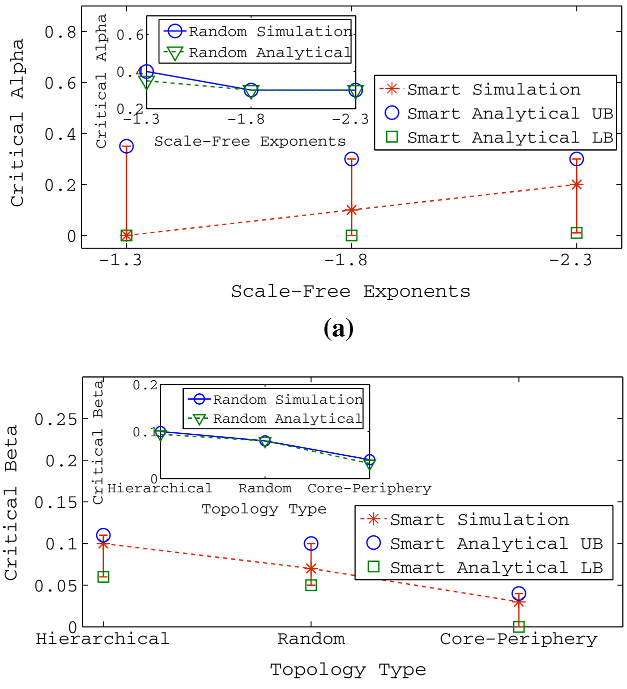
computed threshold (see Eq. 18) β_c^r shows a good agreement with simulation. We observe that β_c^r decays with increasing α. The intuition is, both the retweet rates complement each other in cascade formation. In other words, if α is high, then phase transition can occur with a lower β. In smart mention, we observe (Fig. 12) that β_c^s obtained from the simulation lies within the analytically estimated bounds (see Eq. 23). Similar to random mention, here also we observe that β_c^s decreases with increasing α. But, smart mention attains $\beta_c^s = 0$ at a much lower value of α (0.03) whereas β_c^r becomes 0 only after α becomes 0.4.

5.3.2 *Synthetic follow network*

Finally, we implement the synthetic follower networks using scale-free and Kronecker graphs. Figure 13 demonstrates the presence of a critical β_c for cascade formation for both random and smart mention. Next, we analytically compute this critical rates using Eqs. 18 and 23, respectively. The first step is to estimate the largest eigenvalue $\Lambda_{\max}(A)$, of the network A, directly from the network model parameters. In scale-free network $p_k \sim k^{\gamma}$, the largest eigenvalue can be approximated as the square root of the maximum degree (Chung et al. 2003). Maximum degree of the scale-free network can be estimated as $N^{\frac{1}{\gamma-1}}$ (Boguñá et al. 2004) where N is the number of nodes in the network. Hence, the largest eigenvalue of the scale-free network is approximated as $\sqrt{(N^{\frac{1}{\gamma-1}})}$. Similarly, in Kronecker network, $\Lambda_{\max}$ is approximated from the seed matrix following

Leskovec and Faloutsos (2007). In Table 3, we show the agreement between this approximated eigenvalues and the true eigenvalues computed directly from adjacency matrix A; this approximation shows a nice agreement in terms of computed β_c as well.

Finally, in Fig. 14a, we plot the critical α_c obtained from Eq. 17 (random mention) and Eq. 24 (smart mention), respectively, for scale-free networks and validated with simulation. We observe a contrasting behavior for smart and random mention; for random mention α_c^r decreases with increasing γ whereas the opposite happens for smart mention. Increase in γ decreases the skewness of the degree distribution (improves network homogeneity) which in turn reduces the number of high degree nodes. Since smart mention leverages on the high degree nodes, high γ penalizes the same by increasing the corresponding critical retweet rate. In Kronecker network (see Fig. 14b), core-periphery structure shows lower β_c compared to random and hierarchical structure for both smart and random mention. This happens due to the strongly connected core in the core-periphery Kronecker network.



(a)

(b)

Fig. 14 Critical α for scale-free networks (for $\beta = 0.04$) and critical β for Kronecker networks (for $\alpha = 0.1$) along with their corresponding analytically estimated values. 'UB' and 'LB' in the figures represent the corresponding estimated upper and lower bounds, respectively. **a** Scale-free Networks. **b** Kronecker Networks

6 Observations

This section studies the influence of mention strategies and follower networks on the cascade formation.

6.1 Comparing smart versus random mention

(a) *Impact of retweeting environment*: Figures 9 and 10 shows that the impact of mentioning (α) on R_U is more significant in the low β regime. Naturally, in Fig. 15, smart mention is also observed to be more beneficial in low activity environment (low β). However, increase in β reduces the gap of R_U between the two mention strategies. Essentially, as β increases, mentioning (and hence, mention strategies) becomes less important as most of the

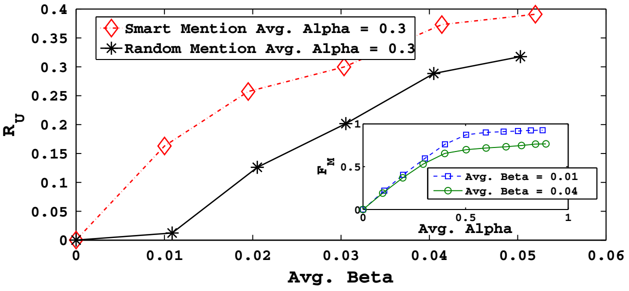


Fig. 15 Smart mentioning versus random mentioning w.r.t. R_U for 'Algeria' dataset. *Inset*: effect of varying average α on F_M for 'Algeria' dataset

Table 3 Estimating maximum eigen values and critical β thresholds for random mention with $\alpha = 0.3$ and $\lambda = 2$ from network's structural parameters

Network	Parameters	$\lambda_{\max}$		β_c^r using	
		Actual	Estimated	Actual $\lambda_{\max}$	Estimated $\lambda_{\max}$
Kronecker	Core-periphery	25.01	25.53	0.01	0.01
	Random	9.71	9.05	0.06	0.06
	Hierarchical	8.43	8.19	0.07	0.07
Scale-free	exp = −2.5	22.81	25.40	0.02	0.02
	exp = −3.0	12.13	11.31	0.05	0.05
	exp = −3.5	6.35	6.96	0.09	0.08
	exp = −4.0	4.58	5.04	0.12	0.11

population starts getting infected due to only follow links (retweeting).

(b) *Retweets by mentioned users*: We denote the average fraction of all the retweets done by the mentioned users as F_M. This is basically the fraction of users who has received the tweet via mention links and retweeted it. We note that F_M increases almost linearly with α upto a point and then converges. In the inset of Fig. 15, we observe that for same α, retweet fraction by the mentioned users (F_M) is lower for higher β values. This is intuitive because if β is high, more people retweet due to follow links which in turn lowers the fraction F_M.

(c) *Impact of* α *on critical* β: In Fig. 12, we observe that smart mention attains critical β $(\beta_c^s) = 0$ for a much lower α $(\alpha = 0.03)$ than random mention $(\alpha = 0.4)$. Therefore, even a small value of α (> 0.03), is able to guarantee the epidemic outbreak (irrespective of β values) in case of smart mention. This happens due to the fact that in comparison to random mentioning, smart mentioning strategy is able to optimally mention suitable users who retweet and expose the tweet content to a much wider population. Hence, in case of smart mention, a larger fraction of the infection spreads through the mention layer which helps the retweet count to attain phase transition even at a low value of α.

(d) *Bounds on critical retweet rates for smart mention*: In Sect. 3.6, we provide the bounds on α_c^s and β_c^s for smart mention. This bounds can be further improved if we assume that smart mention's worst case scenario is equivalent with random mention. In fact, for same parameter values and network configuration, critical retweet rates for smart mention is always less than the critical retweet rates estimated for random mention. This (intuitive) assumption can be established from our simulation results shown in Fig. 13. In that case, we can use the critical retweet rates obtained for random mention as the new (tighter) upper-bounds for critical retweet rates for smart mention. Therefore, the updated estimated ranges for critical α and β for smart mention becomes

$$\left(\frac{1-\beta\Lambda_{\max}(A)}{\lambda\frac{f_{\max}}{\langle f\rangle}},\frac{1-\beta\Lambda_{\max}(A)}{\lambda}\right) \tag{25}$$

and

$$\left(\frac{1-\alpha\lambda\frac{f_{\max}}{\langle f\rangle}}{\Lambda_{\max}(A)},\frac{1-\alpha\lambda}{\Lambda_{\max}(A)}\right) \tag{26}$$

, respectively. We leave the analytical justification of this assumption as future work.

6.2 Importance of follower network

Finally, we investigate the effect of network parameters on the cascade properties. In scale-free networks (see Fig. 16a), we observe that the cascade size R_U increases with γ for random mention. In scale-free network, increase in γ makes the degree distribution more homogeneous, benefiting random mention strategy. On the contrary, in smart mention, two different regimes can be identified depending on the β. In low β region (Fig. 16c), mentioning plays a major role in cascade formation. Notably, increase in γ reduces the fraction of hub nodes in the network, deteriorating the effectiveness of smart mention. This in turn reduces R_U. However, in high β region (Fig. 16b), the tweet propagation mostly relies on the retweeting via follow links (less importance to the mentions), hence R_U increases with γ.

In Kronecker network, Fig. 17 illustrates that core-periphery structure exhibits superior performance against hierarchical structure. Essentially in core-periphery structure, the core plays a pivotal role in propagating tweets across the network. On the contrary, in hierarchical structure the propagation remains confined within a single hierarchical tree like structure due to the poor interaction across different modules.

7 Related work

The state of the art literature in this area can be summarized in three different segments (a) First, modeling information diffusion via retweets, next (b) recent endeavors incorporating mentions in tweets and finally (c) analytical attempts to model epidemics in multiplex networks. The details follows.

7.1 Information propagation in Twitter

Diffusion on social network classically involves the following two propagation models—linear threshold (Granovetter 1978) and independent cascade (Goldenberg et al. 2001). Linear threshold model associated a threshold with each node; a node gets infected if the number of infected neighbors exceeds that threshold. On the other hand, the independent cascade model associates a fixed spreading probability per graph edge and allows each node to attempt infecting another node only once. Further studies (Galuba et al. 2010; Dickens et al. 2012b) have generalized these models. In continuation, Kwak et al. (2010) treated retweet trees as communication channels of information diffusion and analyzed the tweets of top trending topics in whereas Lerman and Ghosh (2010) studied the distribution of retweet cascades on Twitter. Side by side, popular

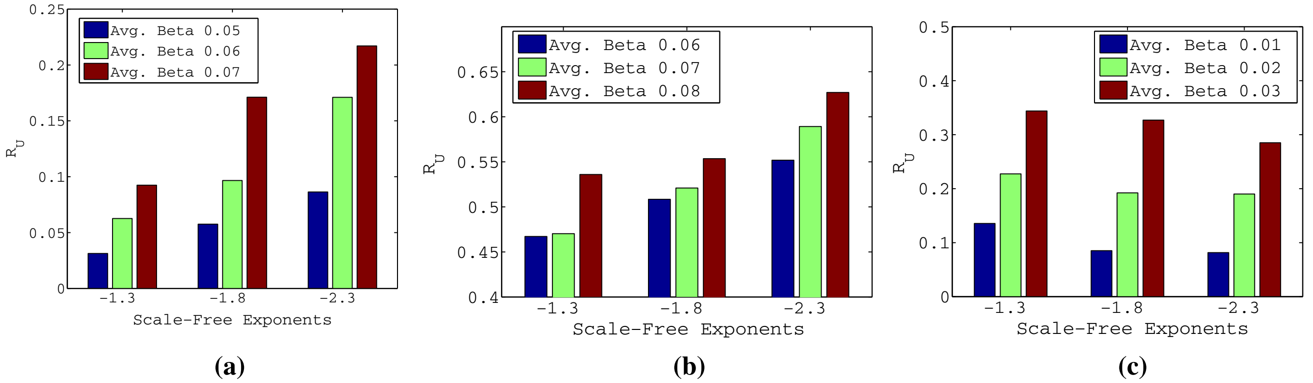


Fig. 16 In scale-free networks, effect of different exponents on R_U for random and smart mention strategies with $\alpha = 0.4$ and $\lambda = 2$. **a** Random mention. **b** Smart mention high avg. β. **c** Smart mention low avg. β

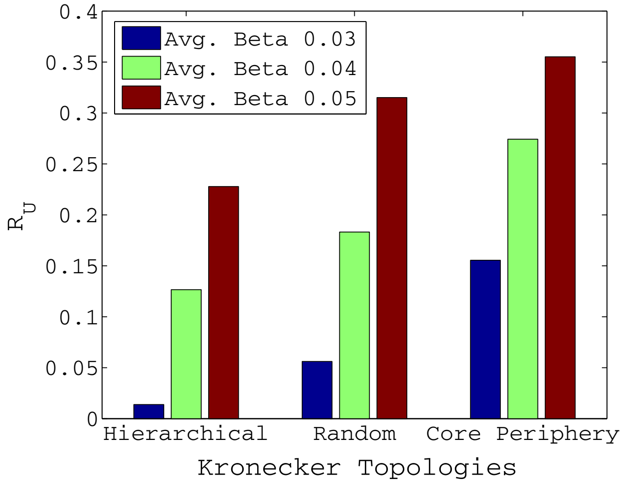


Fig. 17 In Kronecker networks, effect of different topologies on R_U for random mention strategy with $\alpha = 0.4$ and $\lambda = 2$

epidemic like models such as Susceptible-Infected-Susceptible (SIS), Susceptible-Infected-Recovered (SIR) are also explored to model information contagion in Twitter (Li et al. 2013; Abdullah and Wu 2011; Jin et al. 2013).

With the advent of text mining, research in this domain started progressing in two clearly distinguishable tracks. On one hand, several studies have been carried out in understanding the dynamics behind the popularity of tweets. For example, Suh et al. (2010) and Malhotra et al. (2012) investigated the role of content and contextual features of tweets and identified factors that are significantly associated with retweet rate and tweet popularity. In similar line, Petrovic et al. (2011) developed an automated system to predict the popularity of individual tweets. On the other hand, several studies have been made on influence models (Chen et al. 2009; Bakshy et al. 2011) and different recommendation systems have been proposed. For example, Uysal and Croft (2011) proposed methods to recommend useful tweets that users are really interested in and more likely to retweet: given a tweet, they rank users based on retweet probability. Subsequently, considering the influential users in Twitter as potential information brokers, researchers proposed models to identify them and maximize the information propagation (Borge-Holthoefer et al. 2012; Chen et al. 2009). Notably, all the aforementioned models consider retweets as the only mode of tweet propagation.

7.2 Mentioning activities in Twitter

Mentioning is mainly considered as a medium of sending tweet to influential people so that the popularity of the corresponding content increases. However, mentioning one influential user does not ensure that she retweets the post. This later part depends on several factors including information content of the post, profile of the tweeting user, etc. which are not considered while computing the influence. This motivates the community for the development of mention recommendation algorithms to identify the suitable users to mention. For instance, Wang et al. (2013) proposed *Whom-to-Mention* heuristics that uses features (such as user interest match, content-dependent user relationship and user influence) and trains a machine learned ranking function to extract the best users to mention. Similar recommendation heuristics can be found in Gong et al. (2015); Zhou et al. (2015); Pramanik et al. (2016) and Tang et al. (2014). However, most of these recommendation heuristics are based on empirical observations and lacks the analytical understanding of the role of mentions on tweet diffusion.

7.3 Epidemic models in multiplex networks

Diffusion of tweets via both follow and mention modalities can be effectively modeled in the multiplex network framework. Initially a few simplistic models such as Buono et al. (2014); Zhao et al. (2014); Cozzo et al. (2013)

analyzed the epidemics in multilayer networks which started as a single contagion process spreading across multiple layers. Subsequently, recent endeavors model the simultaneous epidemic spread across the multiple layers of the network. For example, Darabi Sahneh and Scoglio (2014) proposed a simple extension of standard SIS framework to model competitive spreading over a two-layer network. The major contribution of this paper is identifying and quantifying extinction, coexistence, and absolute dominance of the competitive epidemic process via defining survival thresholds and absolute dominance thresholds. Guo et al. (2015) and Granell et al. (2013) demonstrated the interplay between simultaneous spread of an epidemic disease and awareness against it, in the framework of multiplex networks. Utilizing a microscopic Markov chain approach, they identified a phase transition and allowed us to capture the evolution of the epidemic threshold depending on the topological structure of the multiplex and the interrelation with the awareness process. However, the absence of network structure realizing mentioning activities makes the aforesaid multiplex models inadequate for Twitter.

8 Conclusion

In this paper, we have proposed a SIR-based analytical framework $\mathcal{C}_F^M$ to model the cascade formation in Twitter, incorporating both retweet and mention activities. Our data study has revealed that the influence of mention directly impacts on the retweet count of a post and retweeting behavior of users. We have introduced a 'generic' mention strategy to model the mention preferences in a tweet, with two special cases—random and smart mentioning. Subsequently, with the help of our proposed framework $\mathcal{C}_F^M$, we have modeled the propagation of tweets on the mention-follow multiplex network and analytically computed the cascade size as well as the critical retweet rate for cascade formation. In random mention, we have been able to obtain the exact critical retweet rate; however for smart mention, only the (upper and lower) bounds can be estimated from the derived open form equations. A close inspection has uncovered the complementary role of followers and mentioned users in cascade formation under different retweeting environment β. Finally, we have validated the $\mathcal{C}_F^M$ framework with Monte Carlo simulation considering both empirical and synthetic follow networks; we have verified the correctness of the simulation setup in the light of real datasets. Interestingly, in scale-free networks, the performance of random mention improves with increasing power law exponent γ irrespective of the retweeting environment β, whereas for smart mention, β plays an important role along with the exponent γ. In a nutshell, the proposed framework may work as the first step toward developing a model driven mention recommendation system.

Acknowledgements This work has been partially supported by the SAP Labs India Doctoral Fellowship program, DST—CNRS funded Indo—French collaborative project 'Evolving Communities and Information Spreading' and French National Research Agency contract CODDDE ANR-13-CORD-0017-01.

References

Abdullah S, Wu X (2011) An epidemic model for news spreading on twitter. In: 2011 IEEE 23rd international conference on tools with artificial intelligence, pp 163–169. doi:10.1109/ICTAI.2011.33

Arenas A, Borge-Holthoefer J, Meloni S, Moreno Y et al (2010) Discrete-time markov chain approach to contact-based disease spreading in complex networks. EPL (Europhys Lett) 89(3):38,009

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining. ACM, pp 65–74

Bao P, Shen HW, Jin X, Cheng XQ (2015) Modeling and predicting popularity dynamics of microblogs using self-excited hawkes processes. In: Proceedings of the 24th international conference on world wide web, WWW '15 Companion. ACM, New York, pp 9–10. doi:10.1145/2740908.2742744

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Barabâsi AL, Jeong H, Néda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. Phys A Stat Mech Appl 311(3):590–614

Boccaletti S, Bianconi G, Criado R, del Genio C, Gmez-Gardees J, Romance M, Sendia-Nadal I, Wang Z, Zanin M (2014) The structure and dynamics of multilayer networks. Phys Rep 544(1):1–122. doi:10.1016/j.physrep.2014.07.001

Bodine-Baron E, Hassibi B, Wierman A (2010) Distance-dependent Kronecker graphs for modeling social networks. IEEE J Sel Top Signal Process 4(4):718–731

Boguñá M, Pastor-Satorras R, Vespignani A (2004) Cut-offs and finite size effects in scale-free networks. Eur Phys J B Condens Matter Complex Syst 38(2):205–209. doi:10.1140/epjb/e2004-00038-8

Borge-Holthoefer J, Rivero A, Moreno Y (2012) Locating privileged spreaders on an online social network. Phys Rev E 85(6):066,123

Buono C, Alvarez-Zuzek LG, Macri PA, Braunstein LA (2014) Epidemics in partially overlapped multiplex networks. PloS ONE 9(3):e92,200

Cerchiello P, Giudici P (2016) How to measure the quality of financial tweets. Qual Quant 50(4):1695–1713. doi:10.1007/s11135-015-0229-6

Chen W, Wang Y, Yang S (2009) Efficient influence maximization in social networks. In: KDD '09. ACM, New York, pp 199–208. doi:10.1145/1557019.1557047

Cheng J, Adamic L, Dow PA, Kleinberg JM, Leskovec J (2014) Can cascades be predicted? In: Proceedings of the 23rd international conference on World wide web. ACM, pp 925–936

Chung F, Lu L, Vu V (2003) Eigenvalues of random power law graphs. Ann Comb 7(1):21–33. doi:10.1007/s000260300002

Clauset A, Moore C, Newman ME (2008) Hierarchical structure and the prediction of missing links in networks. Nature 453(7191):98–101

Cozzo E, Baños RA, Meloni S, Moreno Y (2013) Contact-based social contagion in multiplex networks. Phys Rev E 88(050):801. doi:10.1103/PhysRevE.88.050801

Darabi Sahneh F, Scoglio C (2014) Competitive epidemic spreading over arbitrary multilayer networks. Phys Rev E 89(062):817. doi:10.1103/PhysRevE.89.062817

Dezső Z, Barabási AL (2002) Halting viruses in scale-free networks. Phys Rev E 65(5):055,103

Dickens L, Molloy I, Lobo J, Cheng PC, Russo A (2012a) Learning stochastic models of information flow. In: 2012 IEEE 28th international conference on data engineering, pp 570–581. doi:10.1109/ICDE.2012.103

Dickens L, Molloy I, Lobo J, Cheng PC, Russo A (2012b) Learning stochastic models of information flow. In: 2012 IEEE 28th international conference on data engineering. IEEE, pp 570–581

Du N, Song L, Gomez-Rodriguez M, Zha H (2013) Scalable influence estimation in continuous-time diffusion networks. In: Advances in neural information processing systems, pp 3147–3155

Erdös P, Rényi A (1959) On random graphs, i, vol 6. Publicationes Mathematicae, Debrecen

Galuba W, Aberer K, Chakraborty D, Despotovic Z, Kellerer W (2010) Outtweeting the twitterers-predicting information cascades in microblogs. WOSN 10:3–11

Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. Mark Lett 12(3):211–223. doi:10.1023/A:1011122126881

Gomez-rodriguez M, Schlkopf DBB (2011) Uncovering the temporal dynamics of diffusion networks. In: Proceedings of the 28th international conference on machine learning (ICML11)

Gong Y, Zhang Q, Sun X, Huang X (2015) Who will you "@"? In: Proceedings of the 24th ACM international on conference on information and knowledge management, CIKM '15. ACM, New York, pp 533–542. doi:10.1145/2806416.2806458

González-Bailón S, Borge-Holthoefer J, Rivero A, Moreno Y (2011) The dynamics of protest recruitment through an online network. Scientific reports 1

Granell C, Gómez S, Arenas A (2013) Dynamical interplay between awareness and epidemic spreading in multiplex networks. Phys Rev Lett 111(12):128,701

Granovetter M (1978) Threshold models of collective behavior. Am J Soc 83(6):1420–1443

Guo Q, Jiang X, Lei Y, Li M, Ma Y, Zheng Z (2015) Two-stage effects of awareness cascade on epidemic spreading in multiplex networks. Phys Rev E 91(1):012,822

Gupta M, Gao J, Zhai C, Han J (2012) Predicting future popularity trend of events in microblogging platforms. Proc Am Soc Inf Sci Technol 49(1):1–10

Jin F, Dougherty E, Saraf P, Cao Y, Ramakrishnan N (2013) Epidemiological modeling of news and rumors on twitter. In: Proceedings of the 7th workshop on social network mining and analysis, SNAKDD '13, vol 8. ACM, New York, pp 1–8:9. doi:10.1145/2501025.2501027

Kato S, Koide A, Fushimi T, Saito K, Motoda H (2012) Network analysis of three twitter functions: favorite, follow and mention. In: Richards D, Kang B (eds) Knowledge management and acquisition for intelligent systems, Lecture notes in computer science. Springer, Berlin, pp 298–312. doi:10.1007/978-3-642-32541-0_26

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '03. ACM, New York, pp 137–146. doi:10.1145/956750.956769

Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. J Complex Netw 2(3):203–271

Kupavskii A, Ostroumova L, Umnov A, Usachev S, Serdyukov P, Gusev G, Kustarev A (2012) Prediction of retweet cascade size over time. In: Proceedings of the 21st ACM international conference on information and knowledge management, CIKM '12. ACM, New York, pp 2335–2338. doi:10.1145/2396761.2398634

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. ACM, pp 591–600

Lerman K, Ghosh R (2010) Information contagion: an empirical study of the spread of news on digg and twitter social networks. ICWSM 10:90–97

Leskovec J, Faloutsos C (2007) Scalable modeling of real graphs using kronecker multiplication. In: Proceedings of the 24th international conference on machine learning, ICML '07. ACM, New York, pp 497–504. doi:10.1145/1273496.1273559

Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008) Statistical properties of community structure in large social and information networks. In: Proceedings of the 17th international conference on World Wide Web. ACM, pp 695–704

Leskovec J, Chakrabarti D, Kleinberg J, Faloutsos C, Ghahramani Z (2010) Kronecker graphs: an approach to modeling networks. J Mach Learn Res 11:985–1042

Li W, Tang S, Fang W, Guo Q, Zhang X, Zheng Z (2015) How multiple social networks affect user awareness: the information diffusion process in multiplex networks. Phys Rev E 92(4):042,810

Li Y, Feng Z, Wang H, Kong S, Feng L (2013) ReTweet p: modeling and predicting tweets spread using an extended susceptible-infected- susceptible epidemic model. Springer, Berlin, pp 454–457. doi:10.1007/978-3-642-37450-0_35

Malhotra A, Malhotra CK, See A (2012) How to get your messages retweeted. MIT Sloan Manage Rev 53(2):61–66

Newman ME (2002) Spread of epidemic disease on networks. Phys Rev E 66(1):016,128

Petrovic S, Osborne M, Lavrenko V (2011) Rt to win! predicting message propagation in twitter. In: ICWSM

Pillai SU, Suel T, Cha S (2005) The Perron–Frobenius theorem: some of its applications. IEEE Signal Process Mag 22(2):62–75. doi:10.1109/MSP.2005.1406483

Pramanik S, Wang Q, Danisch M, Bandi S, Kumar A, Guillaume JL, Mitra B (2016) On the role of mentions on tweet virality. In: The 3rd IEEE international conference on data science and advanced analytics (DSAA)

Suh B, Hong L, Pirolli P, Chi EH (2010) Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: 2010 IEEE second international conference on social computing (socialcom). IEEE, pp 177–184

Tang L, Ni Z, Xiong H, Zhu H (2014) Locating targets through mention in twitter. World Wide Web, pp 1–31. doi:10.1007/s11280-014-0299-8

Uysal I, Croft WB (2011) User oriented tweet ranking: a filtering approach to microblogs. In: CIKM '11. ACM, pp 2261–2264

Valera I, Gomez-Rodriguez M, Gummadi K (2014) Modeling diffusion of competing products and conventions in social media. arXiv preprint arXiv:14060516

Wang B, Wang C, Bu J, Chen C, Zhang WV, Cai D, He X (2013) Whom to mention: expand the diffusion of tweets by @ recommendation on micro-blogging systems. In: Proceedings of the 22nd international conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, WWW '13, pp 1331–1340

Xu Z, Yang Q (2012) Analyzing user retweet behavior on twitter. In: Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012), ASONAM '12. IEEE Computer Society, Washington, pp 46–50. doi:10.1109/ASONAM.2012.18

Youssef M, Scoglio C (2011) An individual-based approach to sir epidemics in contact networks. J Theor Biol 283(1):136–144

Zhao D, Li L, Peng H, Luo Q, Yang Y (2014) Multiple routes transmitted epidemics on multiplex networks. Phys Lett A 378(10):770–776

Zhao Q, Erdogdu MA, He HY, Rajaraman A, Leskovec J (2015) Seismic: a self-exciting point process model for predicting tweet popularity. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '15. ACM, New York, pp 1513–1522. doi:10.1145/2783258.2783401

Zhou G, Yu L, Zhang CX, Liu C, Zhang ZK, Zhang J (2015) A novel approach for generating personalized mention list on micro-blogging system. In: 2015 IEEE international conference on data mining workshop (ICDMW), pp 1368–1374. doi:10.1109/ICDMW.2015.51