Estimating the Spreading of Viral Threads on Twitter

Luigi Corvacchiola[1], Eleonora Iotti[2], and Michele Tomaiuolo[2]

[1] Dipartimento di Scienze Matematiche, Fisiche e Informatiche
Università di Parma, I-43124 Parma (Italy)
luigi.corvacchiola@studenti.unipr.it
[2] Dipartimento di Ingegneria e Architettura
Università di Parma, I-43124 Parma (Italy)
eleonora.iotti@studenti.unipr.it
michele.tomaiuolo@unipr.it

Abstract. Microblogging and social news web sites like Twitter are largely used as an important source of up-to-date information. Consequently, organizations and firms have interest in using those platforms to diffuse their own news and updates. The dynamics of information or rumor spread in online social networks depends mainly on network characteristics and is currently a critical topic in Social Network Analysis (SNA). The diffusion through 'retweets' of such information occurs in a time lapse immediately after the publication of the original tweet, and it is internal to some hashtag-based 'channel'. In this study, the retweet count of a given tweet is assumed as an index of its diffusion. For analyzing the statistical features of viral tweets, we have selected five tweets. Our model is based on the hypothesis that it is highly probable that a user decides to retweet a tweet if he/she is following either the tweet author, or another retweeter of the tweet. Therefore, we choose as main features of a tweet its number of retweets and the number of followers of retweeting users.

Keywords: Social Network Analysis · Twitter · MLE

1 Introduction

Social platforms involve billions of people all around the world, attracting users from several social groups, regardless of age, gender, education, or nationality. These systems blur the distinction between the private and working spheres, and users are known to use such systems both at home and on the work place, both professionally and with recreational goals. In particular, microblogging and social news web sites like Twitter are largely used as an important source of up-to-date information. On the other hand, firms and agencies are interested in using those platforms to diffuse their own news and updates, related to specific campaigns or for their daily operation.

In Social Network Analysis (SNA), the study of information spreading processes is a critical topic [1]. In fact, understanding the dynamics of information

or rumor spread in social networks is very important for many different purposes, such as marketing campaigns, political influence, news diffusion and so on. The way a piece of information reaches people and how much time it takes to do it depend mainly on network characteristics, on the influence of the source of information, and on the meaning of the information content, which deserves a special attention and may depends on the context. Examine such information content is out of the scope of this paper, and has to be analyzed, for example, with Speech Act Theory, which studies linguistic expressions that aim at performing some functions.

Thus, information spreading is based on the analysis of the underlying social graph and its users' motives and patterns of participation. At its core, SNA is the process for studying social networks and understanding the behaviors of their members. Graph theory provides the basic foundations for representing and studying a social network. In fact, each member of the social network can be mapped onto a node of a graph and each relationship between two members onto an edge that connects two nodes. In real life, it is very common to find examples of social networks: groups of friends, a company's employees, contributors with different aims, etc. In fact, SNA is currently used in many research fields including anthropology, biology, economics, geography, information science, organizational studies, political science, social psychology.

One of the most important application of SNA is to find subgroups of strongly interconnected users, i.e., to perform community detection [13]. Many users can be considered a community if the existing connections, internal to the community, are many more than the ones with outside users (this situation is similar to a dense graph). Detecting the presence of a community allows analysts to recognize the paths followed by information for reaching the network users, on the basis of different metrics. For example, Degree Centrality measures the capability to spread information directly to other users. Instead, Betweenness Centrality is gauge of how much a user could be able to diffuse information from a community to another, especially if he/she belongs to many communities. Finally, Closeness Centrality provides information about how far a user is from all other members of the community; thus, it provides information about the probability of his/her own posts to reach all those fellow members.

Other important kinds of analysis regard the behavior of a certain user [19], who can be classified for example as "active" (when he produces contents, sends videos and photos, comments posts of other users, reports original texts and documents) or "passive" (when he is only a consumer of other users' contents, limiting himself to liking or unliking those contents). But it is also important to study the dynamics of a social network structure during time [3], to discover for example its lead users, who can be distinguished as the best connected and stable nodes in the social graph.

In the following sections, the paper will first discuss the state of the art about the analysis of rumor spreading, also in relation to the nature of the underlying social network; then it will present some theoretical tools to analyze the phenomenon of viral information spreading; afterwords, it will describe the

methodology of analysis and finally it will provide some experimental results, obtained by comparing the mathematical model with some real world cases of viral tweets.

2 Related work

Several models have been developed in order to study the phenomenon of information spreading, but there is not an unique standard option, due to the heterogeneity of social networks [23], from real-world ones to online social networks, such as micro-blogging services or forums. Despite those diversities, social networks share common features that are taken as basis for further analysis. First of all, a network is often viewed as a graph $G = (V, E)$, where V is a discrete finite set of nodes (or vertices) that represents the people or users involved, and E is a binary relation on V, that represents relationships among users. The neighborhood of a node is the set of other nodes directly connected to him/her.

Depending on network, the topological characteristics of the graph change; several models have been investigated to match the correct shape of a network, such as complete graph [24], hypercubes [11], random graphs [25] and evolving random graphs [5], preferential attachment graphs [2, 9], power-law degree graphs [14] and so on. Complete graphs are graphs where all nodes are connected to each others, i.e., where each individual has a complete view of the social network and can communicate with all other users. Hypercube graphs are graphs whose nodes and edges are the vertices and edges of a n-dimensional hypercube. Random graphs refer to the Erdős-Rényi model, i.e., graphs where edges appear independently with a certain probability p, thus connecting nodes randomly. Such random graphs were discovered to not effectively model social networks, while evolving random graphs, i.e. random graphs which changes as functions of time, show more realistic behaviors. Finally, preferential attachment graphs and power-law degree graphs are variants of evolving random graphs, and are currently studied in SNA, mainly because they produce scale-free networks. As a matter of fact, real social networks have often the shape of scale-free networks, i.e., their degree distribution follows a power-law.

In literature, rumor spreading on a graph (thus, a social network) has been studied by means of two types of distributed mechanisms [20, 21]: the push protocol and the flooding protocol. Both protocols are synchronous, i.e., time steps, or rounds, are used to describe the behavior of a node, and the piece of information or rumor originates by a single source node. In the flooding protocol, starting from the source at the first time step, each node forwards the information to all nodes in its neighborhood. In the push protocol, instead, at every time step, each informed node in the social network chooses uniformly at random another node, and shares with it the piece of information. Behavior of such protocols are widely investigated for several types of graphs [17], and their performance, time of completion [4,6] or other measures, such as conductance [15], are well-known. In [5], a formal argument is provided, for demonstrating the robustness of the

push protocol also against network changes, using the model of edge-markovian dynamic graphs.

The actual challenge is to understand when and how such protocols, or their variants, are suitable in order to describe information spreading in a certain social network with its own topological model. Answers to such problem differ according to social network characteristics and platforms, taking account of the peculiar communication patterns of certain online social networks, e.g., the Twitter retweet mechanism [22, 28], or the way Facebook users share posts [10]. In particular, in [18] the simplicial model is applied to the study of higher dimensional social groups, where opinion leaders play an important role in information spreading. Members of such groups are characterized by: sharing a world-view and a sense of identity; open in-group communication climate; and a shared life story. All these features can be mapped to various Facebook types of information and activities: overlapping profile data and liked pages; multiple interactions through messages and comments; tags in common pictures and participation to events.

Another, recent, approach is the study of network metrics, such as degree centrality, closeness centrality, and betweenness centrality, by means of Semantic SNA [7,8]. This approach takes into account contents of topics, in order to obtain different understanding of the information flow in a social network.

The study of information diffusion often gave rise to other inherent questions, such as how a topic becomes popular and which methods can make it viral [30]. Those matters are analyzed by means of statistical models that aim to predict the future impact of a new information released within the social network. Currently, "*little is known about factors that could affect the dissemination of a single piece of information*" [27], and several predictive models have been proposed. Each model have to face two main issues: the impact of the topology of the underlying social network—with all the related formalizations—, the influence of the individual behavior of users and, finally, the communication patterns of the community (online or not).

A common approach is to assign a score to such features [26, 29, 31]. In some networks, the underlying graph model is very important because diffusion is subordinated to connection among users, for example if the piece of information is visible only to a user's neighborhood. In other networks, messages or posts are public, and this fact overcome topological limits, bypassing relationship to address wide audience. Moreover, the propagation speed depends on the context in which the piece of information is introduced. All those considerations are useful to gain the correct score of a feature, and then the scores are put together to obtain an estimation of the diffusion probability of a single topic.

3 Background and Notations

In this section we define and formalize some main features of tweets, in order to model the information diffusion phenomenon of Twitter. Tweets representation as key-value dictionary (in particular, JSON objects) can be obtained by using

the Twitter APIs. The information contained in these objects may vary from user personal details to the text of the message, or the number of times the tweet was retweeted.

In this study, tweets information are grouped under so-called 'channels', i.e. lists of tweets identified by the presence of the same hashtag in their text, or the same keyword. The diffusion of such information occurs in a time lapse immediately after the publication of the tweet, and it is internal to the channel. This diffusion consists in the re-publication of the same content of the original tweet, possibly commented. Such a practice is called 'retweet' and it is widely employed by Twitter users. We assume the retweet count of a given tweet as an index of its diffusion inside the online social network, and, in particular, inside its channel. Extremely popular tweets are retweeted thousand times, but inside a channel, a tweet can become popular with few dozens of retweets. Deciding how many retweets make a tweet 'viral' depends on the underlying social network and on the topic of the tweet, and is out of the scope of this paper. Any user who reads a tweet, in a channel or on its Twitter feed, can retweet that tweet. We assume as a hypothesis that is highly probable that a user decides to retweet a tweet if he/she is following either the tweet author or another retweeter of the tweet. Therefore, we choose as main features of a tweet its number of retweets and the number of followers of retweeting users.

In the following, some useful definitions are given, which will be used in the rest of the paper. Definitions are taken from [12], where details and properties on Beta and mixed distributions are largely explained.

Definition 1 (Beta Distribution). *The* Beta distribution *is a continuous probability distribution, which has two positive parameters* $\alpha, \beta \in \mathbb{R}^+$ *and that takes values in the* $[0, 1]$ *interval. Its probability density function is*

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \tag{1}$$

where $B(\alpha, \beta)$ *defined below is called* Beta function*, and it acts as a normalization constant.*

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{2}$$

$\Gamma(z)$ *denotes the Euler's Gamma function. A random variable* X *beta-distributed is denoted by* $X \sim \text{Beta}(\alpha, \beta)$.

Definition 2 (Beta-binomial Distribution). *The* Beta-binomial distribution *of parameters* $n \in \mathbb{N}$ *and* $\alpha, \beta \in \mathbb{R}^+$ *is a compound distribution of the binomial and the beta distributions, where the parameter* p *of the binomial distribution is drawn from a beta distribution. Hence, the beta-binomial distribution is a discrete distribution with probability mass function*

$$\varphi(k) = \binom{n}{k} \frac{B(\alpha + k, \beta + n - k)}{B(\alpha, \beta)}. \tag{3}$$

The proposed model consists in the assumption that each user who reads a given tweet can retweet that tweet independently with probability $p \in [0, 1]$, or not, with probability $q = 1 - p$. As a central design decision, we assume that this probability p is not a constant, but it has a Beta distribution. The Beta distribution is used because of its peculiarities. As a matter of fact, by varying the parameters α and β, the distribution adopts different shapes. Hence, it is suitable for describing various and distinct phenomena. Because we do not have any information which helps in defining the distribution of p a-priori, the Beta distribution acts as an indicator that models the behavior of the random variable.

Fixed a tweet T, for each of the N_T Twitter users who retweeted T (including the author of the tweet), we consider two parameters, namely n_i and x_i, where $i = 1, \ldots, N_T$ denotes the user. The n_i parameter is the followers count of the i-th user who retweeted T. It can be observed directly on Twitter. The x_i parameter is the number of users who follow i and who also retweeted T.

As an observation, x_i is a known numeric value, but under the assumption of the proposed model, is a random variable $X_i \sim \text{Beta-Bin}(n_i, \alpha_T, \beta_T)$ following the Beta-binomial distribution. The parameters α_T and β_T are unknown and one of the targets of this work is to find a way to estimate them.

3.1 Maximum Likelihood Estimation

The parameters α_T and β_T of our problem are estimated by using the *Maximum Likelihood Estimation* (*MLE*) method. As a matter of fact, given a random variable X distributed with a certain law $f(\cdot)$, which belongs to a family parameterized by unknown parameters θ, the MLE is a method for estimating such parameters. There are other statistical methods useful in this case, cited, e.g., in [16], such as the method of moments. Not all methods are suitable for investigating the Beta-binomial distribution parameters.

The MLE, in details, starts from a given sample of N independent and identically distributed (iid) observations $(x_1, x_2, \ldots, x_N)$ of X, of which the joint probability density function $f(x_1, \ldots, x_N, \theta)$ is unknown. Hence, it is desirable to estimate the joint density $f(x_1, \ldots, x_N|\theta)$ of the observations, parameterized by θ. Because the observations are iid, such a joint density is equal to $f(x_1|\theta)f(x_2|\theta)\cdots f(x_N|\theta)$. Then, we can obtain some prediction $\hat{\theta}$ of the parameter θ. The cost of a prediction $\hat{\theta}$ is called *loss function* and is denoted by $L(\hat{\theta}, \theta)$. Thus, the optimum value for θ is obtained by minimizing the loss function.

$$\hat{\theta}_{\text{OPT}} = \arg\min_{\hat{\theta}} \mathbb{E}\left(L(\hat{\theta}, \Omega)|X = (x_1, x_2, \ldots, x_N)\right). \tag{4}$$

The equation above can be rewritten as

$$\hat{\theta}_{\text{OPT}} = \arg\max_{\theta} f(\theta|x_1, x_2, \ldots, x_N). \tag{5}$$

Following Bayes Theorem and the Law of Total Probability, it is sufficient to maximize the joint function $f(x_1, \ldots, x_N|\theta)f(\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_N|\theta)f(\theta)$.

The estimation

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f(x_1, x_2, \ldots, x_N|\theta). \tag{6}$$

is called *Maximum A Posteriori* (*MAP*).

Definition 3 (Likelihood). *The* likelihood *is defined as follows*

$$\mathcal{L}(\theta; x_1, \ldots, x_N) = \prod_{i=1}^{N} f(x_i|\theta) \tag{7}$$

where θ is an array of parameters, $f(\cdot|\theta)$ is a family of probability density functions, and $(x_1, x_2, \ldots, x_N)$ is a sample of N iid observations under the law $f(\cdot)$.

Following the equation (6), the MLE consists in maximizing the likelihood function (7) evaluated in $(x_1, x_2, \ldots, x_N)$:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \mathcal{L}(\theta; x_1, \ldots, x_N). \tag{8}$$

Hence, in the following section, the likelihood of the observed number of retweets is computed, using the Beta-binomial distribution and the MLE method to find the missing parameters.

4 Experimental Results

In the proposed model, we fix a tweet T. Then, the parameter θ of the equation (8) is the pair (α_T, β_T) that represents the parameter of the Beta distribution. As a matter of fact, for each observation x_i, the parameter n_i of the Beta-binomial is also known. Since the iid observations are occurrence of a Beta-binomial random variables, the likelihood has the following equation:

$$\mathcal{L}(\alpha, \beta) = \prod_{i=1}^{N} \binom{n_i}{x_i} \frac{B(\alpha + x_i, \beta + n_i - x_i)}{B(\alpha, \beta)} \tag{9}$$

where the subscript T is omitted for the sake of clarity. Let $l(\alpha, \beta) = \log \mathcal{L}(\alpha, \beta)$ be the log-likelihood. Hence,

$$\begin{aligned} l(\alpha, \beta) &= \sum_{i=1}^{N} \left[\log \binom{n_i}{x_i} + \log B(\alpha + x_i, \beta + n_i - x_i) - \log B(\alpha, \beta) \right] \\ &= \sum_{i=1}^{N} \log \binom{n_i}{x_i} - N \log B(\alpha, \beta) + \sum_{i=1}^{N} \log B(\alpha + x_i, n_i - x_i + \beta). \end{aligned} \tag{10}$$

Recalling the property of the Γ function:

$$\log \Gamma(x + y) = \log \Gamma(x) - \sum_{k=0}^{y-1} \log(x - k) \tag{11}$$

and by the equation (2) which defines the Beta function,

$$\begin{aligned} l(\alpha,\beta) &= K - N\left(\log\Gamma(\alpha)+\log\Gamma(\beta)-\log\Gamma(\alpha,\beta)\right) \\ &\quad + \sum_{i=1}^{N}\left(\log\Gamma(\alpha+x_i)+\log\Gamma(\beta+n_i-x_i)-\log\Gamma(\alpha+\beta+n_i)\right) \\ &= K + \sum_{i=1}^{N}\left[\sum_{k=0}^{x_i-1}\log(\alpha+k)+\sum_{k=0}^{n_i-x_i-1}\log(\beta+k)\right. \\ &\quad \left. - \sum_{k=0}^{n_i-1}\log(\alpha+\beta+k)\right] \end{aligned} \tag{12}$$

where $K=\sum_{i=1}^{N}\log\binom{n_i}{x_i}$ because it is constant with respect to α and β. Denoted with $\#A$ the cardinality of the set A, the log-likelihood can be rewritten as follows.

$$\begin{aligned} l(\alpha,\beta) &= K + \sum_{k=0}^{\infty}\log(\alpha+k)\,\#\{i|x_i>k\} + \sum_{k=0}^{\infty}\log(\beta+k)\,\#\{i|n_i-x_i>k\} \\ &\quad - \sum_{k=0}^{\infty}\log(\alpha+\beta+k)\,\#\{i|n_i>k\}. \end{aligned} \tag{13}$$

It is worth noting that each summation contains a finite number of addend, because the sets $\{i|x_i>k\}, \{i|n_i-x_i>k\}$ and $\{i|n_i>k\}$ are definitely empty. Using the log-likelihood form of the equation (13), we obtain

$$\frac{\partial l(\alpha,\beta)}{\partial\alpha} = \sum_{k=0}^{\infty}\frac{\#\{i|x_i>k\}}{\alpha+k} - \sum_{k=0}^{\infty}\frac{\#\{i|n_i>k\}}{\alpha+\beta+k} \tag{14}$$

$$\frac{\partial l(\alpha,\beta)}{\partial\beta} = \sum_{k=0}^{\infty}\frac{\#\{i|n_i-x_i>k\}}{\beta+k} - \sum_{k=0}^{\infty}\frac{\#\{i|n_i>k\}}{\alpha+\beta+k}. \tag{15}$$

For maximizing the log-likelihood, and finding the MLE of α and β, the following non-linear system has to be solved

$$\begin{cases} \displaystyle\sum_{k=0}^{\infty}\frac{\#\{i|x_i>k\}}{\alpha+k} = \sum_{k=0}^{\infty}\frac{\#\{i|n_i>k\}}{\alpha+\beta+k} \\ \displaystyle\sum_{k=0}^{\infty}\frac{\#\{i|n_i-x_i>k\}}{\beta+k} = \sum_{k=0}^{\infty}\frac{\#\{i|n_i>k\}}{\alpha+\beta+k} \end{cases} \tag{16}$$

In order to test the model, five viral tweets were selected. Their data were obtained by using Twitter APIs. Four tweets were chosen among the most popular tweets of a trend topic of the 2016, namely 'blizzard2016'. Such tweets have a high number of retweets and have an high rating within the channel. The

Table 1. Total number N_{T_i} of retweets, and α_{T_i}, β_{T_i} values of tweets $T_1, \ldots, T_5$.

i	N_{T_i}	α_{T_i}	β_{T_i}
1	185564	0.45071028	26.14814566
2	16990	0.54793266	173.70447388
3	11271	0.20797788	32.33433353
4	9707	0.75284294	366.08739169
5	47	0.14291630	42.61378355

Table 2. Perturbations of tweet T_5.

Perturbation	$\tilde{\alpha}_{T_5}$	$\tilde{\beta}_{T_5}$	err_α	err_β
$d_i + U_i$	1.34595852	376.71443709	1.20304222	334.10065354
$\text{round}(d_i + U_i)$	0.39128181	94.79466674	0.24836551	52.18088319
$d_i * M_i$	0.31015776	126.54103851	0.16724146	83.92725496
$\text{round}(d_i * M_i)$	0.25690555	105.04234006	0.11398925	62.42855651

last tweet belongs to the 'macron' channel, which become popular after French elections in 2017. It has a very low number of retweets if compared to the others.

The datasets are indexed by the Twitter user id i for each of the retweeters of the fixed tweet T, and they contains the followers count n_i and the number of followers of i who retweet T, i.e. x_i. It is worth noting that many users have a high number of followers, but the number of retweeters is often very low or zero. As a matter of fact, the summation of x_i for $i = 1, \ldots, N_T$ must be greater or equal the total number of retweet N_T of the tweet T.

The datasets were used for evaluating the summations in (16). Fixed the tweet T_i, the solutions α_{T_i} and β_{T_i} of the system were obtained by using the Newton-Raphson (NR) method for finding roots of non-linear equations. The solutions have precision of 10^{-12}. The NR method was used due to its velocity in convergence, that is quadratic. Moreover, an analytical expression for the derivative of the system (16) is easily obtainable, so the NR method was suitable for effectively calculating the solutions.

In Table 4, the results of the parameters estimation is shown for each tweet. The resulting parameters vary significantly from tweet to tweet, especially β_{T_i}. The only notable characteristics of such results is that $0 < \alpha_{T_i} < 1$ and that $\beta_{T_i} >> \alpha_{T_i}$. In Figure 4, the plot of the Beta distributions of parameters α_{T_i} and β_{T_i} for $i = 1, \ldots, 5$ is shown. Because $\alpha_{T_i} < 1$ and $\beta_{T_i} \geq 1$, all lines have the same shape: skewed, decreasing and convex. As shown in the figure, the probability p of retweeting a content from a friend (i.e., a user you follow) tends to be very low, as expected.

Nevertheless, the model has to be refined to correctly gain a score for the retweeting probability, because the estimated solutions are very diverse among each others. Moreover, the problem itself is ill-conditioned: little variations on data may change significantly the results. In Table 4, some examples of artificial perturbation were given. We denote with d_i both the followers and retweets

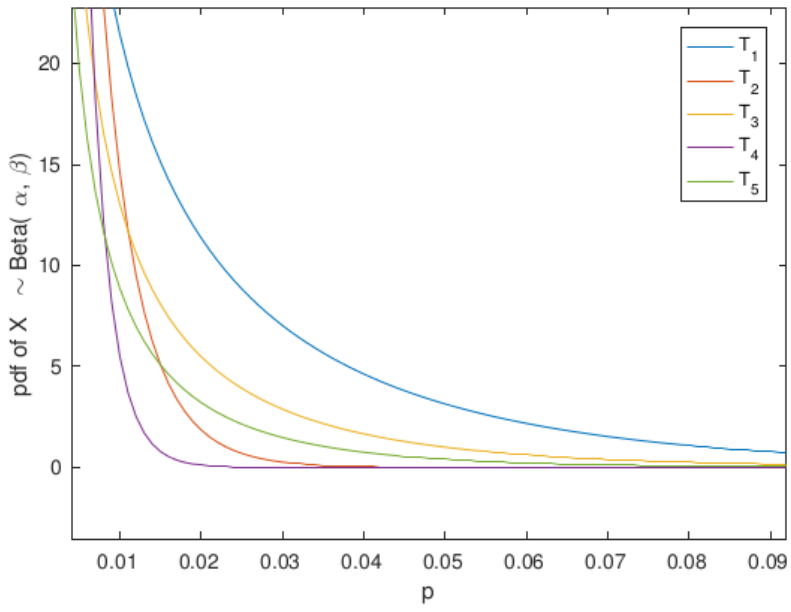

Fig. 1. Plot of the Beta distributions of parameters α_{T_i} and β_{T_i} for $i = 1, \ldots, 5$.

counts. The first perturbation was obtained by adding to d_i the occurrence of a uniform random variable between 0 and 1: $U_i \sim U(0, 1)$. The second is the first rounded to the closest integer, because n_i and x_i are integers by the definition of the problem. The third perturbation was obtained by multiplying to d_i the occurrence of a uniform random variable $M_i \sim U(0, 1)$ (scaling the values of each d_i). As the second, the forth perturbation is the rounded value of the third.

Table 4 shows also the errors $\text{err}_\alpha = |\alpha_{T_5} - \tilde{\alpha}_{T_5}|$ and $\text{err}_\beta = \left|\beta_{T_5} - \tilde{\beta}_{T_5}\right|$ in estimating the perturbed parameters.

5 Conclusions

In this paper, a model for predicting the retweet probability of a given tweet T was proposed. First, a discussion on the state of the art in SNA and the analysis of rumor spreading was given, in order to highlight problems and developments. The key issue addressed in this paper is the searching for the score of a feature, in particular, the retweet count of a viral tweet.

We assume that a user who reads a tweet can retweet it with a probability p, which is unknown. Each user retweets a tweet independently from the

other users. Thus, a binomial probability distribution is suitable to model the phenomenon. As another assumption, we state that p is the occurrence of a Beta-distributed random variable. Such an assumption is crucial because the Beta distribution density function has a shape that varies according to two parameters: α and β. Hence, the model states that the number of retweeters among a user list of followers is a Beta-binomial random variable of parameters n, α and β, where n is the number of followers. Since α and β are unknown, we use MLE to gain an estimation of such parameters. Applying MLE to the model, we obtain a non-linear system that has to be solved to find α and β. Theoretical analytical methods are too difficult or fail in solving exactly that system, so the well-known Newton-Raphson method was used.

Results are shown in the last section. We found that the problem is ill-conditioned and the experiments gain very different values of α and β. But, as a notable result, the obtained Beta distributions candidate to be that of the p probability of retweet have all the same shape, since $\alpha < 1$ and $\beta \geq 1$.

In conclusion, the proposed model shows some limitations (it was not possible to obtain unique α and β), but give some information about the shape of the target probability. Further developments have to be made in order to improve the precision of the model, mainly by considering other features of the tweets, or by relaxing the hypothesis that a follower is more likely to retweet his/her friends tweet than others.

References

1. Angiani, G., Fornacciari, P., Iotti, E., Mordonini, M., Tomaiuolo, M.: Models of participation in social networks. Social Media Performance Evaluation and Success Measurements p. 196 (2016)
2. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science 286(5439), 509–512 (1999)
3. Barabâsi, A.L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical mechanics and its applications 311(3), 590–614 (2002)
4. Baumann, H., Crescenzi, P., Fraigniaud, P.: Parsimonious flooding in dynamic graphs. In: Proceedings of the 28th ACM symposium on Principles of distributed computing. pp. 260–269. ACM (2009)
5. Clementi, A., Crescenzi, P., Doerr, C., Fraigniaud, P., Pasquale, F., Silvestri, R.: Rumor spreading in random evolving graphs. Random Structures & Algorithms 48(2), 290–312 (2016)
6. Clementi, A.E.F., Macci, C., Monti, A., Pasquale, F., Silvestri, R.: Flooding time of edge-markovian evolving graphs. SIAM journal on discrete mathematics 24(4), 1694–1712 (2010)
7. Cristani, M., Fogoroasi, D., Tomazzoli, C.: Measuring homophily. vol. 1748 (2016), https://www.scopus.com/inward/record.uri?eid=2-s2.0-85012298603&partnerID=40&md5=81df100456c2118853ca823496097c79
8. Cristani, M., Tomazzoli, C., Olivieri, F.: Semantic social network analysis foresees message flows. vol. 1, pp. 296–303 (2016), https://www.scopus.com/inward/record.uri?eid=2-s2.0-84969287486&partnerID=40&md5=6d7a0bb42fd4f45cdb48b8dc1193907a

9. Doerr, B., Fouz, M., Friedrich, T.: Why rumors spread so quickly in social networks. Communications of the ACM 55(6), 70–75 (2012)
10. Fan, W., Yeung, K.H.: Virus propagation modeling in facebook. In: Procs. of the 2010 Int'l Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 331–335. IEEE (2010)
11. Feige, U., Peleg, D., Raghavan, P., Upfal, E.: Randomized broadcast in networks. Random Structures & Algorithms 1(4), 447–460 (1990)
12. Feller, W.: An introduction to probability theory and its applications, vol. 2. John Wiley & Sons (2008)
13. Fortunato, S.: Community detection in graphs. Physics reports 486(3), 75–174 (2010)
14. Fountoulakis, N., Panagiotou, K., Sauerwald, T.: Ultra-fast rumor spreading in social networks. In: Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms. pp. 1642–1660. Society for Industrial and Applied Mathematics (2012)
15. Giakkoupis, G.: Tight bounds for rumor spreading in graphs of a given conductance. In: Symposium on Theoretical Aspects of Computer Science (STACS2011). vol. 9, pp. 57–68 (2011)
16. Ibragimov, I.A., Has' Minskii, R.Z.: Statistical estimation: asymptotic theory, vol. 16. Springer Science & Business Media (2013)
17. Karp, R., Schindelhauer, C., Shenker, S., Vocking, B.: Randomized rumor spreading. In: Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on. pp. 565–574. IEEE (2000)
18. Kee, K.F., Sparks, L., Struppa, D.C., Mannucci, M.A., Damiano, A.: Information diffusion, facebook clusters, and the simplicial model of social aggregation: a computational simulation of simplicial diffusers for community health interventions. Health communication 31(4), 385–399 (2016)
19. Klein, A., Ahlf, H., Sharma, V.: Social activity and structural centrality in online social networks. Telematics and Informatics 32(2), 321–332 (2015)
20. Kuhn, F., Lynch, N., Oshman, R.: Distributed computation in dynamic networks. In: Proceedings of the forty-second ACM symposium on Theory of computing. pp. 513–522. ACM (2010)
21. Kuhn, F., Oshman, R.: Dynamic networks: models and algorithms. ACM SIGACT News 42(1), 82–96 (2011)
22. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Procs. of the 19^{th} Int'l Conference on World Wide Web. pp. 591–600. ACM (2010)
23. Moreno, Y., Nekovee, M., Pacheco, A.F.: Dynamics of rumor spreading in complex networks. Physical Review E 69(6), 066130 (2004)
24. Pittel, B.: On spreading a rumor. SIAM Journal on Applied Mathematics 47(1), 213–223 (1987)
25. Rényi, A., Erdős, P.: On random graphs. Publicationes Mathematicae 6(290-297), 5 (1959)
26. Shah, D., Zaman, T.: Rumors in a network: Who's the culprit? IEEE Transactions on Information Theory 57(8), 5163–5181 (2011)
27. Wang, D., Wen, Z., Tong, H., Lin, C.Y., Song, C., Barabási, A.L.: Information spreading in context. In: Procs. of the 20^{th} Int'l Conference on World Wide Web. pp. 735–744. ACM (2011)
28. Ye, S., Wu, S.F.: Measuring message propagation and social influence on twitter. com. SocInfo 10, 216–231 (2010)

29. Zaman, T., Fox, E.B., Bradlow, E.T., et al.: A bayesian approach for predicting the popularity of tweets. The Annals of Applied Statistics 8(3), 1583–1611 (2014)
30. Zaman, T.R., Herbrich, R., Van Gael, J., Stern, D.: Predicting information spreading in twitter. In: Workshop on Computational Social Science and the Wisdom of Crowds. vol. 104, pp. 599–601. Citeseer (2010)
31. Zhou, Y., Guan, X., Zhang, Z., Zhang, B.: Predicting the tendency of topic discussion on the online social networks using a dynamic probability model. In: Procs. of the 2008 Workshop on Collaboration and Collective Intelligence. pp. 7–11. ACM (2008)