Modelling of Trends in Twitter Using Retweet Graph Dynamics

Marijn ten Thij[1]($\boxtimes$), Tanneke Ouboter[2], Daniël Worm[2], Nelly Litvak[3],
Hans van den Berg[2,3], and Sandjai Bhulai[1]

[1] Faculty of Sciences, VU University Amsterdam, Amsterdam, The Netherlands
`{m.c.ten.thij,s.bhulai}@vu.nl`
[2] TNO, Delft, The Netherlands
`{tanneke.ouboter,daniel.worm,j.l.vandenberg}@tno.nl`
[3] Faculty of Electrical Engineering, Mathematics and Computer Science,
University of Twente, Enschede, The Netherlands
`n.litvak@utwente.nl`

Abstract. In this paper we model user behaviour in *Twitter* to capture
the emergence of trending topics. For this purpose, we first extensively
analyse tweet datasets of several different events. In particular, for these
datasets, we construct and investigate the retweet graphs. We find that
the retweet graph for a trending topic has a relatively dense largest con-
nected component (LCC). Next, based on the insights obtained from the
analyses of the datasets, we design a mathematical model that describes
the evolution of a retweet graph by three main parameters. We then quan-
tify, analytically and by simulation, the influence of the model parameters
on the basic characteristics of the retweet graph, such as the density of
edges and the size and density of the LCC. Finally, we put the model in
practice, estimate its parameters and compare the resulting behavior of
the model to our datasets.

Keywords: Retweet graph · Twitter · Graph dynamics · Random graph
model

1 Introduction

Nowadays, social media play an important role in our society. The topics people
discuss on-line are an image of what interests the community. Such trends may
have various origins and consequences: from reaction to real-world events and
naturally arising discussions to the trends manipulated e.g. by companies and
organisations [14]. Trending topics on *Twitter* are 'ongoing' topics that become
suddenly extremely popular[1]. In our study, we want to reveal differences in the
retweet graph structure for different trends and model how these differences arise.

---

[1] https://support.twitter.com/articles/101125-about-trending-topics

In *Twitter*[2] users can post messages that consist of a maximum of 140 characters. These messages are called tweets. One can "follow" a user in *Twitter*, which places their messages in the message display, called the timeline. Social ties are directed in *Twitter*, thus if user A follows user B, it does not imply that B follows A. People that "follow" a user are called "friends" of this user. We refer to the network of social ties in *Twitter* as the friend-follower network. Further, one can forward a tweet of a user, which is called a retweet.

There have been many studies on detecting different types of trends, for instance detecting emergencies [9], earthquakes [18], diseases [13] or important events in sports [11]. In many current studies into trend behaviour, the focus is mainly on content of the messages that are part of the trend, see e.g. [12]. Our work focuses instead on the underlying networks describing the social ties between users of *Twitter*. Specifically, we consider a graph of users, where an edge means that one of the users has retweeted a message of a different user.

In this study we use several datasets of tweets on multiple topics. First we analyse the datasets, described in Section 3, by constructing the retweet graphs and obtaining their properties as discussed in Section 4. Next, we design a mathematical model, presented in Section 5, that describes the growth of the retweet graph. The model involves two attachment mechanisms. The first mechanism is the preferential attachment mechanism that causes more popular messages to be retweeted with a higher probability. The second mechanism is the superstar mechanism which ensures that a user that starts a new discussion receives a finite fraction of all retweets in that discussion [2]. We quantify, analytically and with simulations, the influence of the model parameters on its basic characteristics, such as the density of edges, the size and the density of the largest connected component. In Section 6 we put the model in practice, estimate its parameters and compare it to our datasets. We find that what our model captures, is promising for describing the retweet graphs of trending topics. We close with conclusions and discussion in Section 7.

2   Related Work

The amount of literature regarding trend detection in *Twitter* is vast. The overview we provide here is by no means complete. Many studies have been performed to determine basic properties of the so-called "Twitterverse". Kwak et al. [10] analysed the follower distribution and found a non-power-law distribution with a short effective diameter and a low reciprocity. Furthermore they found that ranking by the number of followers and PageRank both induce similar rankings. They also report that *Twitter* is mainly used for News (85% of the content). Huberman et al. [8] found that the network of interactions within *Twitter* is not equal to the follower network, it is a lot smaller.

An important part of trending behaviour in social media is the way these trends progress through the network. Many studies have been performed on *Twitter* data. For instance, [3] studies the diffusion of news items in *Twitter* for

---

[2] www.twitter.com

several well-known news media and finds that these cascades follow a star-like structure. Also, [20] investigates the diffusion of information on *Twitter* using tweets on the Iranian election in 2009, and finds that cascades tend to be wide, not too deep and follow a power law-distribution in their size.

Bhamidi et al. [2] proposed and validated on the data a so-called superstar random graph model for a giant component of a retweet graph. Their model is based on the well-known preferential attachment idea, where users with many retweets have a higher chance to be retweeted [1], however, there is also a superstar node that receives a new retweet at each step with a positive probability. We build on this idea to develop our model for the progression of a trend through the *Twitter* network.

Another perspective on the diffusion of information in social media is obtained through analysing content of messages. For example, [17] finds that on *Twitter*, tags tend to travel to more distant parts of the network and URLs travel shorter distances. Romero et al. [16] analyse the spread mechanics of content through hashtag use and derive probabilities that users adopt a hashtag.

Classification of trends on *Twitter* has attracted considerable attention in the literature. Zubiaga et al. [21] derive four different types of trends, using 15 features to make their distinction. They distinguish trends triggered by news, current events, memes or commemorative tweets. Lehmann et al. [12] study different patterns of hashtag trends in *Twitter*. They also observe four different classes of hashtag trends. Rattanaritnont et al. [15] propose to distinguish topics based on four factors, which are cascade ratio, tweet ratio, time of tweet and patterns in topic-sensitive hashtags.

We extend the model of [2] by mathematically describing the growth of a complete retweet graph. Our proposed model has two more parameters that define the shape of the resulting graph, in particular, the size and the density of its largest connected component. To the best of our knowledge, this is the first attempt to classify trends using a random graph model rather than algorithmic techniques or machine learning. The advantage of this approach is that it gives insight in emergence of the trend, which, in turn, is important for understanding and predicting the potential impact of social media on real world events.

3   Datasets

We use datasets containing tweets that have been acquired either using the *Twitter* Streaming API[3] or the *Twitter* REST API[4]. Using the REST API one can obtain tweets or users from *Twitter*'s databases. The Streaming API filters tweets that *Twitter* parses during a day, for example, based on users, locations, hashtags, or keywords.

Most of the datasets used in this study were scraped by RTreporter, a company that uses an incoming stream of Dutch tweets to detect news for news agencies in the Netherlands. These tweets are scraped based on keywords, using

---

[3] https://dev.twitter.com/docs/streaming-apis
[4] https://dev.twitter.com/docs/api/1.1

the Streaming API. For this research, we selected several events that happened in the period of data collection, based on the wikipedia overviews of 2013 and 2014[5]. We have also used two datasets scraped by TNO - Netherlands Organisation for Applied Scientific Research. The *Project X* dataset contains tweets related to large riots in Haren, the Netherlands. This dataset is acquired by *Twitcident*[6]. For this study, we have filtered this dataset on two most important hashtags: *#projectx* and *#projectxharen*. The *Turkish-Kurdish* dataset is described in more detail in Bouma et al. [4]. A complete overview of the datasets, including the events and the keywords, is given in Table 1. The size and the timespans for each dataset are given in Table 2.

Table 1. Datasets: events and keywords (some keywords are in Dutch)

	dataset	keywords
PX	Project X Haren	projectx, projectxharen
TK	Demonstrations in Amsterdam related to the Turkish-Kurdish conflict	koerden, turken, rellen, museumplein, amsterdam
WCS	World cup speedskating single distanced 2013	wkafstanden, sochi, sotsji
W-A	Crowning of His Majesty King Willem-Alexander in the Netherlands	troonswisseling, troon, Willem-Alexander, Wim-Lex, Beatrix, koning, koningin
ESF	Eurovision Song Festival	esf, Eurovisie Songfestival, ESF, songfestival, eurovisie
CL	Champions Leage final 2013	Bayern Munchen, Borussia Dortmund, dorbay, borussia, bayern, borbay, CL
Morsi	Morsi deposited as Egyption president	Morsi, afgezet, Egypte
Train	Train crash in Santiago, Spain	Treincrash, treincrash, Santiago, Spanje, Santiago de Compostella, trein
Heat	Heat wave in the Netherlands	hittegolf, Nederland
Damascus	Sarin attack in Damascus	Sarin, Damascus, Syrië, syrië
Peshawar	Bombing in Peshawar	Peshawar, kerk, zelfmoordaanslag, Pakistan
Hawk	Hawk spotted in the Netherlands	sperweruil, Zwolle
Pile-up	Multiple pile-ups in Belgium on the A19	A19, Ieper, Kortrijk, kettingbotsing
Schumi	Michael Schumachar has a skiing accident	Michael Schumacher, ski-ongeval
UKR	Rebellion in Ukrain	Azarov, Euromaidan, Euromajdan, Oekraïne, opstand
NAM	Treaty between NAM and Dutch government	Loppersum, gasakkoord, NAM, Groningen
WCD	Michael van Gerwen wins PDC WC Darts	van Gerwen, PDC, WK Darts
NSS	Nuclear Security Summit 2014	NSS2014, NSS, Nuclear Security Summit 2014, Den Haag
MH730	Flight MH730 disappears	MH730, Malaysia Airlines
Crimea	Crimea referendum for independance	Krim, referendum, onafhankelijkheid
Kingsday	First Kingsday in the Netherlands	koningsdag, kingsday, koningsdag
Volkert	Volkert van der Graaf released from prison	Volkert, volkertvandergraaf, Volkert van der Graaf

For each dataset we have observed there is at least one large peak in the progression of the number of tweets. For example, Figure 1 shows such peak in *Twitter* activity for the *Project X* dataset.

When a retweet is placed on *Twitter*, the Streaming API returns the retweet together with the message that has been retweeted. We use this information to construct the retweet trees of every message and the user IDs for each posted message. The tweet and graph analysis is done using *Python* and its modules

---

[5] http://nl.wikipedia.org/wiki/2014 & http://nl.wikipedia.org/wiki/2013

[6] www.twitcident.com

Table 2. Characteristics of the datasets

| dataset	year	first tweet	last tweet	# tweets	# retweets
PX	2012	Sep 17 09:37:18	Sep 26 02:31:15	31,144	15,357
TK	2011	Oct 19 14:03:23	Oct 27 08:42:18	6,099	999
WCS	2013	Mar 21 09:19:06	Mar 25 08:45:50	2,182	311
W-A	2013	Apr 27 22:59:59	May 02 22:59:25	352,157	88,594
ESF	2013	May 13 23:00:08	May 18 22:59:59	318,652	82,968
CL	2013	May 22 23:00:04	May 26 22:59:54	163,612	54,471
Morsi	2013	Jun 30 23:00:00	Jul 04 22:59:23	40,737	13,098
Train	2013	Jul 23 23:00:02	Jul 30 22:59:41	113,375	26,534
Heat	2013	Jul 10 19:44:35	Jul 29 22:59:58	173,286	42,835
Damascus	2013	Aug 20 23:01:57	Aug 31 22:59:54	39,377	11,492
Peshawar	2013	Sep 21 23:00:00	Sep 24 22:59:59	18,242	5,323
Hawk	2013	Nov 11 23:00:07	Nov 30 22:58:59	54,970	19,817
Pile-up	2013	Dec 02 23:00:15	Dec 04 22:59:57	6,157	2,254
Schumi	2013-14	Dec 29 02:43:16	Jan 01 22:54:50	13,011	5,661
UKR	2014	Jan 26 23:00:36	Jan 31 22:57:12	4,249	1,724
NAM	2014	Jan 16 23:00:22	Jan 20 22:59:49	41,486	14,699
WCD	2013-14	Dec 31 23:03:48	Jan 02 22:59:05	15,268	5,900
NSS	2014	Mar 23 23:00:06	Mar 24 22:59:56	29,175	13,042
MH730	2014	Mar 08 00:18:32	Mar 28 22:40:44	36,765	17,940
Crimea	2014	Mar 13 23:02:22	Mar 17 22:59:57	18,750	5,881
Kingsday	2014	Apr 26 23:00:00	Apr 29 22:53:00	7,576	2,144
Volkert	2014	Apr 30 23:08:14	May 04 22:57:06	9,659	4,214
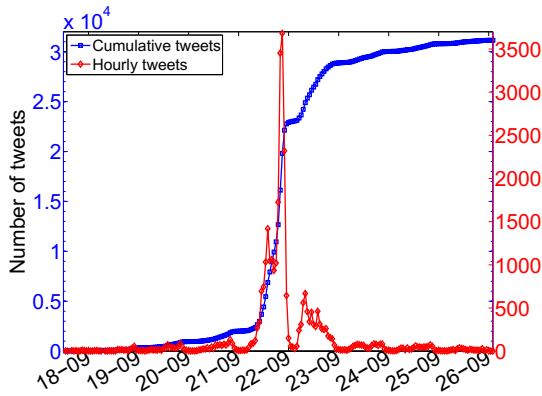
Fig. 1. *Project X* Number of tweets and cumulative number of tweets per hour

Tweepy[7] and *NetworkX*[8]. In this paper, we investigate the dynamics of retweet graphs with the goal to predict peaks in *Twitter* activity and classify the nature of trends.

4   Retweet Graphs

Our main object of study is the retweet graph $G = (V, E)$, which is a graph of users that have participated in the discussion on a specific topic. A directed

---

edge $e = (u, v)$ indicates that user v has retweeted a tweet of u. We observe the retweet graph at the time instances $t = 0, 1, 2, \ldots$, where either a new node or a new edge was added to the graph, and we denote by $G_t = (V_t, E_t)$ the retweet graph at time t. As usual, the out- (in-) degree of node u is the number of directed edges with source (destination) in u. In what follows, we model and analyse the properties of G_t. For every new message initiated by a new user u a tree T_u is formed. Then, $\mathcal{T}_t$ denotes the forest of message trees. Note that in our model a new message from an already existing user u (that is, $u \in \mathcal{T}_t$) does not initiate a new message tree. We define $|\mathcal{T}_t|$ as the number of new users that have started a message tree up to time t.

After analyzing multiple characteristics of the retweet graphs for every hour of their progression, we found that the size of the largest (weakly) connected component (LCC) and its density are the most informative characteristics for predicting the peak in *Twitter*. In Figure 2 we show the development of these characteristics in the *Project X* dataset. One day before the actual event, we observe a very interesting phenomenon in the development of the edge density of the LCC in Figure 2a. Namely, at some point the edge density of the LCC exceeds 1 (indicated by the dash-dotted gray lines), i.e. there is more than one retweet per user on average. We shall refer to this as the *densification* (or dens.) of the LCC. Furthermore, the relative size of the LCC increases from 18% to 25% as well, see Figure 2b.
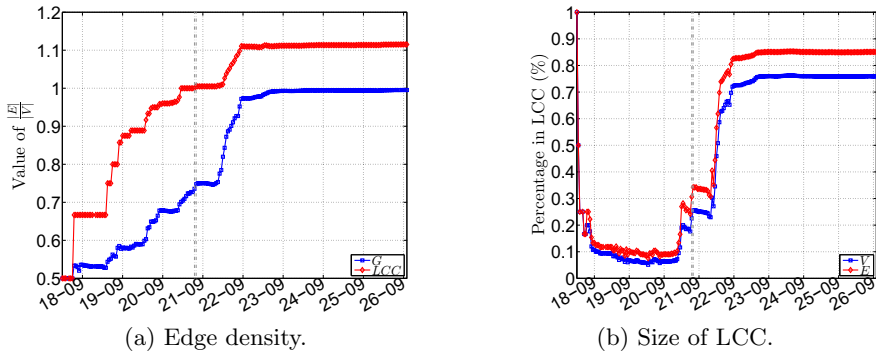

(a) Edge density.   (b) Size of LCC.

Fig. 2. Progression for the edge density (a) and the size of the LCC (b) in the *Project X* dataset

We have observed a densification of the LCC in each dataset that we have studied. Indeed, when the LCC grows its density must become at least one (each node is added to the LCC together with at least one edge). However, we have also observed that in each dataset the densification occurs before the main peak, but the scale of densification is different. For example, in the *Project X* dataset the densification already occurs one day before the peak activity. Plausibly, in this discussion, that ended up in riots, a group of people was actively participating before the event. On the other hand, in the *WCS* dataset, which tweets about an ongoing sport event, the densication of the LCC occurs during the largest

peak. This is the third peak in the progression. Hence, our experiments suggest that the time of densification has predictive value for trend progression and classification. See Table 3 for the density of the LCC in each dataset at the end of the progression.

5   Model

Our goal is to design a model that captures the development of trending behaviour. In particular, we need to capture the phenomenon that disjoint components of the retweet graph join together forming the largest component, of which the density of edges may become larger than one. To this end, we employ the superstar model of Bhamidi et al. [2] for modelling distinct components of the retweet graph, and add the mechanism for new components to arrive and the existing components to merge. For the sake of simplicity of the model we neglect the friend-follower network of *Twitter*. Note that in *Twitter* every user can retweet any message sent by any public user, which supports our simplification.

At the start of the progression, we have the graph G_0. In the analysis of this section, we assume that G_0 consists of a single node. Note that in reality, this does not need to be the case: any directed graph can be used as an input graph G_0. In fact, in Section 6 we start with the actual retweet graph at a given point in time, and then use the model to build the graph further to its final size.

We consider the evolution of the retweet graph in time $(G_t)_{t\geq 0}$. We use a subscript t to indicate G_t and related notions at time t. We omit the index t when referring to the graph at the end of the progression.

Recall that G_t is a graph of *users*, and an edge (u, v) means that v has retweeted a tweet of u. We consider time instances $t = 1, 2, \ldots$ when either a new node or a new edge is added to the graph G_{t-1}. We distinguish three types of changes in the retweet graph:

- $T1$: a new user u has posted a new message on the topic, node u is added to G_{t-1};
- $T2$: a new user v has retweeted an existing user u, node v and edge (u, v) are added to G_{t-1};
- $T3$: an existing user v has retweeted another existing user u, edge (u, v) is added to G_{t-1}.

The initial node is equivalent to a $T1$ arrival at time $t = 0$. Assume that each change in G_t at $t = 1, 2, \ldots$ is $T1$ with probability $\lambda/(1 + \lambda)$, independently of the past. Also, assume that a new edge (retweet) is coming from a new user with probability p. Then the probabilities of $T1$, $T2$ and $T3$ arrivals are, respectively $\frac{\lambda}{\lambda+1}, \frac{p}{\lambda+1}, \frac{1-p}{\lambda+1}$. The parameter p is governing the process of components merging together, while λ is governing the arrival of new components in the graph.

For both $T2$ and $T3$ arrivals we define the same mechanism for choosing the source of the new edge (u, v) as follows.

Let $u_0, u_1, \ldots$ be the users that have been added to the graph as $T1$ arrivals, where u_0 is the initial node. Denote by $T_{i,t}$ the subgraph of G_t that includes u_i

and all users that have retweeted the message of u_i in the interval $(0, t]$. We call such a subgraph a message tree with root u_i. We assume that the probability that a $T2$ or $T3$ arrival at time t will attach an edge to one of the nodes in $T_{i,t-1}$ with probability $p_{T_{i,t-1}}$, proportional to the size of the message tree:

$$p_{T_{i,t-1}} = \frac{|T_{i,t-1}|}{\sum_{T_{j,t-1} \subset \mathcal{T}_{t-1}} |T_{j,t-1}|}.$$

This creates a preferential attachment mechanism in the formation of the message trees. Next, a node in the selected message tree $T_{i,t-1}$ is chosen as the source node following the superstar attachment scheme [2]: with probability q, the new retweet is attached to u_i, and with probability $1 - q$, the new retweet is attached to any other vertex, proportional to the preferential attachment function of the node, that we choose to be the number of children of the node plus one.

Thus we employ the superstar-model, which was suggested in [2] for modelling the largest connected component of the retweet graph on a given topic, in order to describe a progression mechanism for a single retweet tree. Our extensions compared to [2] are that we allow new message trees to appear ($T1$ arrivals), and that different message trees may either remain disconnected or get connected by a $T3$ arrival.

For a $T3$ arrival, the target of the new edge (u, v) is chosen uniformly at random from V_{t-1}, with the exception of the earlier chosen source node u, to prevent self-loops. That is, any user is equally likely to retweet a message from another existing user.

Note that, in our setting, it is easy to introduce a different superstar parameter q_{T_i} for every message tree T_i. This way one could easily implement specific properties of the user that starts the message tree, e.g. his/her number of followers. For the sake of simplicity, we choose the same value of q for all message trees. Also note that we do not include tweets and retweets that do not result in new nodes or edges in a retweet graph. This could be done, for example, by introducing dynamic weights of vertices and edges, that increase with new tweets and retweets. Here we consider only an unweighted model.

5.1 Growth of the Graph

The average degree, or edge density, is one of the aspects through which we give insight to the growth of the graph. The essential properties of this characteristic are presented in Theorem 1. The proof is given in the Appendix.

Theorem 1. *Let τ_n be the time when node n is added to the graph. Then*

$$\mathbb{E}\left[\frac{|E_{\tau_n}|}{|V_{\tau_n}|}\right] = \frac{1}{\lambda + p} - \frac{1}{n(\lambda + p)}, \tag{1}$$

$$var\left(\frac{|E_{\tau_n}|}{|V_{\tau_n}|}\right) = \frac{(n-1)(\lambda + 1 - p)}{n^2(\lambda + p)^2}. \tag{2}$$

Note that the variance of the average degree in (2) converges to zero as $n \to \infty$ at rate $\frac{1}{n}$.

The next theorem studies the observed ratio between $T2$ and $T3$ arrivals (new edges) and $T1$ arrivals (new nodes with a new message). As we see from the theorem, this ratio can be used for estimating the parameter λ. The proof is given in the Appendix.

Theorem 2. *Let $G_t = (V_t, E_t)$ be the retweet graph at time t, let $\mathcal{T}_t$ be the set of all message trees in G_t. Then*

$$\mathbb{E}\left[\frac{|E_t|}{|\mathcal{T}_t|}\right] = \lambda^{-1} \cdot \left(1 - \left(\frac{1}{\lambda+1}\right)^t\right), \tag{3}$$

$$\lim_{t \to \infty} \frac{\lambda^3 t}{(1+\lambda)^2} var\left(\frac{|E_t|}{|\mathcal{T}_t|}\right) = 1, \tag{4}$$

Furthermore,

$$\frac{\lambda^{3/2}\sqrt{t}}{\lambda+1}\left(\frac{|E_t|}{|\mathcal{T}_t|} - \frac{1}{\lambda}\right) \xrightarrow{D} Z, \tag{5}$$

where Z is a standard normal $N(0,1)$ random variable, and $\xrightarrow{D}$ denotes convergence in distribution.

Note that, as expected from the definition of λ,

$$\lim_{t \to \infty} \mathbb{E}\left[\frac{|E_t|}{|\mathcal{T}_t|}\right] = \lambda^{-1}. \tag{6}$$

This will be used in Section 6 for estimating λ.

5.2   Component Size Distribution

In the following, we assume that G_t consists of m connected components $(C_1, C_2, \ldots, C_m)$ with known respective sizes $(|C_1|, \ldots, |C_m|)$. We aim to derive expressions for the distribution of the component sizes in G_{t+1}.

Lemma 3. *The distribution of the sizes of the components of G_{t+1}, given G_t is as follows,*

$$\begin{cases} |C_1|, \ldots, |C_i|, |C_j|, \ldots, |C_m|, 1 & w.p. \ \frac{\lambda}{\lambda+1} \\ |C_1|, \ldots, |C_i| + 1, |C_j|, \ldots, |C_m| & w.p. \ \frac{p}{\lambda+1} \cdot \frac{|C_i|}{|V|} \\ |C_1|, \ldots, |C_i| + |C_j|, \ldots, |C_m| & w.p. \ \frac{1-p}{\lambda+1} \cdot \frac{2 \cdot |C_i| \cdot |C_j|}{|V|^2 - |V|} \\ |C_1|, \ldots, |C_i|, |C_j|, \ldots, |C_m| & w.p. \ \frac{1-p}{\lambda+1} \cdot \frac{\sum_{k=1}^m |C_k|^2 - |C_k|}{|V|^2 - |V|} \end{cases} \tag{7}$$

The proof of Lemma 3 is given in the Appendix. Lemma 3 provides a recursion for computing the distribution of component sizes. However, the computations are highly demanding if not infeasible. Also, deriving an exact expression of the distribution of the component sizes at time t is very cumbersome because they are hard and they strongly depend on the events that occurred at $t = 0, \ldots, t - 1$. Note that if $p = 1$, there is a direct correspondence between our model and the infinite generalized Pólya process [5]. However, this case is uninformative as there are no $T3$ arrivals. Therefore, in the next section we resort to simulations to investigate the sensitivity of the graph characteristics to the model parameters.

5.3   Influence of q, p and λ

We analyze the influence of the model parameters λ, p and q on the characteristics of the resulting graph numerically using simulations. To this end, we fix two out of three parameters and execute multiple simulation runs of the model, varying the values for the third parameter. We start simulations with graph G_0, consisting of one node. We perform 50 simulation runs for every parameter setting and obtain the average values over the individual runs for given parameters.

Parameter q affects the degree distribution [2] and the overall structure of the graph. If $q = 0$, then the graph contains less nodes that have many retweets. If $q = 1$ each edge is connected to a superstar, and the graph consists of star-like sub graphs, some of which are connected to each other. In the *Project X* dataset, which is our main case study, $q \approx 0.9$ results in a degree distribution that closely approximates the data. Since degree distributions are not in the scope of this paper, we omit these results for brevity.

We compare the results for two measures that produced especially important characteristics of the *Project X* dataset: $\frac{|E_{\text{LCC}}|}{|V_{\text{LCC}}|}$ and $\frac{|V_{\text{LCC}}|}{|V|}$. These characteristics do not depend on q. In simulations, we set $t = 1,000$, $q = 0.9$ and vary the values for p and λ. the results are give in Figure 3.

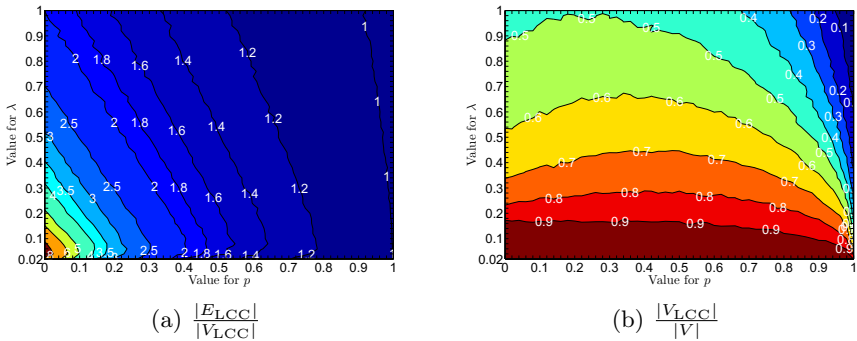(a) $\frac{|E_{\text{LCC}}|}{|V_{\text{LCC}}|}$ (b) $\frac{|V_{\text{LCC}}|}{|V|}$

Fig. 3. Numerical results for the model using $q = 0.9$ and $t = 1,000$

We see that the edge density in the LCC in Figure 3a decreases with λ and p. Note that according to (1), $|E|/|V|$ is well approximated by $1/(\lambda + p)$ when λ or p are large enough. The edge density in LCC shows a similar pattern, but it is slightly higher than in the whole graph. When λ and p are small, there are many $T3$ arrivals, and new nodes are not added frequently enough. This results in an unexpected non-monotonic behaviour of the edge density near the origin. For the fraction of nodes in the LCC, depicted in Figure 3b, we see that the parameter λ is most influential. The parameter p is of considerable influence only when it is large.

6   The Model in Practice

In this section we obtain parameter estimators for our model and compare the model to the datasets discussed in Section 3.

Using Theorem 2, we know that $\frac{|E_t|}{|T_t|}$ converges to λ^{-1} as $t \to \infty$. Thus, we suggest the following estimator for λ at time $t > 0$:

$$\hat{\lambda}_t = \frac{|T_t|}{|E_t|}. \tag{8}$$

Second, we derive an expression for $\hat{p}_t$ using (1) and substituting (8) for λ:

$$\hat{p}_t = \frac{|V_t| - |T_t| - 1}{|E_t|}. \tag{9}$$

Since the *Twitter* API only gives back the original message of a retweet and not the level in the progression tree of that retweet, we can not determine q easily from the data. Since this parameter does not have a large influence on the outcomes of the simulations, we choose this parameter to be 0.9 for all datasets.

Notice that we can obtain the numbers ($|E_t|$, $|T_t|$ and $|V_t|$) directly from a given retweet graph for each $t = 1, 2, \ldots$. The computed estimators for our datasets are displayed in Table 3.

Next, we compare 50 simulations of the datasets from the point of densification of the LCC until the graph has reached the same size as the actual dataset. We display the average outcomes of these simulations and compare them to the actual properties of the retweet graphs of each dataset in Table 3.

Here we see diverse results per dataset in the simulations. For the *CL, Morsi* and *WCD* datasets, the simulations are very similar to the actual progressions. However, for some datasets, for instance the *ESF* dataset, simulations are far off. In general, the model predicts the density of the LCC quite well for many datasets, but tends to overestimate the size of the LCC. We notice that current random graph models for networks usually capture one or two essential features, such as degree distribution, self-similarity, clustering coefficient or diameter. Our model captures both degree distribution and, in many cases, the density of the LCC. It seems that our model performs better on the datasets that have a

Table 3. Estimated parameter values using complete dataset, simulation and progression properties

dataset	$\hat{\lambda}$	$\hat{p}$	actual progression $\frac{|V_{LCC}|}{|V|}$	$\frac{|E|}{|V|}$	$\frac{|E_{LCC}|}{|V_{LCC}|}$	simulations (starting at dens.) $\frac{|V_{LCC}|}{|V|}$	$\frac{|E|}{|V|}$	$\frac{|E_{LCC}|}{|V_{LCC}|}$
PX	.23	.78	.76	1.00	1.12	.54	.75	1.08
TK	.42	.85	.25	.79	1.00	.54	.74	1.08
WCS	.49	.73	.20	.81	.99	.49	.95	1.90
W-A	.41	.52	.67	1.07	1.30	.40	.62	1.41
ESF	.38	.43	.73	1.24	1.48	.45	.69	1.42
CL	.40	.72	.44	.90	1.22	.46	.66	1.16
Morsi	.60	.55	.39	.87	1.20	.47	.67	1.17
Train	.54	.78	.28	.76	1.04	.50	.70	1.17
Heat	.42	.59	.60	.99	1.23	.41	.72	1.68
Damascus	.58	.51	.46	.92	1.24	.44	.65	1.30
Peshawar	.54	.68	.31	.82	1.18	.53	.75	1.25
Hawk	.38	.38	.82	1.31	1.45	.49	.76	1.43
Pile-up	.33	.64	.65	1.03	1.24	.58	.93	1.54
Schumi	.38	.83	.33	.82	1.08	.56	.77	1.07
UKR	.72	.37	.53	.91	1.12	.50	.75	1.38
NAM	.44	.48	.50	1.09	1.51	.45	.72	1.51
WCD	.26	.81	.66	.94	1.10	.64	.83	1.07
NSS	.26	.62	.79	1.13	1.26	.23	.35	1.21
MH730	.33	.52	.15	1.18	1.00	.56	.76	1.09
Crimea	.44	.63	.51	.93	1.19	.52	.72	1.12
Kingsday	.47	.92	.07	.72	1.11	.47	.67	1.15
Volkert	.29	.55	.79	1.18	1.31	.64	.87	1.22

singular peak rather than a series of peaks. We have observed on the data that each peak activity has a large impact on the parameters estimation. We will strive to adopt the model for incorporating different rules for activity during peaks, and improving results on the size of the LCC.

7   Conclusion and Discussion

We have found that our model performs well in modelling the retweet graph for tweets regarding a singular topic. However, there is a room for improvement when the dataset covers a prolonged discussion with users activity fluctuating over time.

A possible extension of the present work is incorporating more explicitly the time aspect into our model. We could for example add the notion of 'novelty', like Gómez et al. in [6], taking into account that e.g. the retweet probability for a user may decrease the longer he/she remains silent after having received a tweet. But also other model parameters may be assumed to vary over time. In addition, we propose to analyse the clustering coefficient of a node in the network model and, in particular, to investigate how it evolves over time. This measure (see [19]) provides more detailed insight in how the graph becomes denser, making it possible to distinguish between local and global density.

Appendix

A1. Proof of Theorem 1

Proof. The proof is based on the fact that the total number of edges $|E_{\tau_n}|$ equals a total number of the $T2$ and $T3$ arrivals on $(0, \tau_n]$. By definition, $(0, \tau_n]$ contains exactly $(n-1)$ of $T1$ or $T2$ arrivals, hence, the number of $T2$ arrivals has a Binomial distribution with number of trials equal to $(n-1)$, and success probability $P(T2 \,|\, T1 \text{ or } T2) = \frac{p}{\lambda + p}$. Next, the number of $T3$ arrivals on $[\tau_i, \tau_{i+1})$, where $i = 1, \ldots, n-1$, has a shifted geometric distribution, namely, the probability of k $T3$ arrivals on $[\tau_i, \tau_{i+1})$ is

$$\left(1 - \frac{1-p}{\lambda + 1}\right) \left(\frac{1-p}{\lambda + 1}\right)^k, \ k = 0, 1, \ldots.$$

Observe that there have been $n-1$ of these transitions from 1 node to n. Hence, the number of $T3$ arrivals on $(0, \tau_n]$ is the sum of $(n-1)$ i.i.d. Geometric random variables with mean $\frac{1-p}{\lambda + p}$. Summarizing the above, we obtain (1). For (2) we also need to observe that the number of $T2$ and $T3$ arrivals on $[0, \tau_n]$ are independent.

A2. Proof of Theorem 2

Proof. Let X_t be the number of $T2$ and $T3$ arrivals by time t. Note that $|E_t| = X_t$, and $|\mathcal{T}_t| = t - X_t + 1$, which is the number of T_1 arrivals on $[0, t]$, since the first node at time $t = 0$ is by definition a $T1$ arrival. Note that X_t has a binomial distribution with parameters t and $\mathbb{P}(T2 \text{ arrival}) + \mathbb{P}(T3 \text{ arrival}) = \frac{1}{\lambda + 1}$. Furthermore, the number of $T1$ arrivals is $t - X_t + 1$ since the first node at time $t = 0$ is by definition a $T1$ arrival. Hence,

$$\mathbb{E}\left[\frac{|E_t|}{|\mathcal{T}_t|}\right] = \sum_{i=1}^{t} \frac{i}{t - i + 1} \binom{t}{i} \left(\frac{1}{\lambda + 1}\right)^i \left(\frac{\lambda}{\lambda + 1}\right)^{t-i}$$

$$= \frac{1}{\lambda} \cdot \sum_{i=1}^{t} \binom{t}{i-1} \left(\frac{1}{\lambda + 1}\right)^{i-1} \left(\frac{\lambda}{\lambda + 1}\right)^{t-i+1},$$

which proves (3). Next, we write

$$\mathbb{E}\left[\left(\frac{|E_t|}{|\mathcal{T}_t|}\right)^2\right] = \sum_{i=0}^{t} \left(\frac{i}{t - i + 1}\right)^2 \binom{t}{i} \left(\frac{1}{\lambda + 1}\right)^i \left(\frac{\lambda}{\lambda + 1}\right)^{t-i}$$

$$= \frac{1}{\lambda} \cdot \sum_{i=1}^{t} \frac{i}{t - i + 1} \binom{t}{i-1} \left(\frac{1}{\lambda + 1}\right)^{i-1} \left(\frac{\lambda}{\lambda + 1}\right)^{t-i+1}$$

$$= \frac{1}{\lambda} \mathbb{E}\left[\frac{t+1}{t - X_t} \mathbb{1}_{\{X_t \leq t-1\}}\right] - \frac{1}{\lambda}\left(1 - \left(\frac{1}{1 + \lambda}\right)^t\right), \tag{10}$$

where $\mathbb{1}_{\{A\}}$ is an indicator of event A. Denoting

$$Z_t = \frac{X_t - \mathbb{E}\left[X_t\right]}{\sqrt{var\left(X_t\right)}} = \frac{(\lambda+1)X_t - t}{\sqrt{\lambda t}}, \tag{11}$$

we further write

$$\mathbb{E}\left[\frac{t+1}{t-X_t}\mathbb{1}_{\{X_t \le t-1\}}\right] = \mathbb{E}\left[\frac{(t+1)(\lambda+1)}{\lambda t(1-\frac{Z_t}{\sqrt{\lambda t}})}\mathbb{1}_{\{Z_t \le \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}}\right]. \tag{12}$$

We now split the indicator above as follows:

$$\mathbb{1}_{\{Z_t \le -\sqrt{\lambda t}\}} + \mathbb{1}_{\{-\sqrt{\lambda t} < Z_t < \sqrt{\lambda t}/2\}} + \mathbb{1}_{\{\sqrt{\lambda t}/2 \le Z_t \le \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}}. \tag{13}$$

For the first and the third term we use the Chernoff bound:

$$\mathbb{E}\left[\frac{1}{1-\frac{Z_t}{\sqrt{\lambda t}}}\mathbb{1}_{\{Z_t \le -\sqrt{\lambda t}\}}\right] \le 2e^{-\lambda t/4}, \tag{14}$$

$$\mathbb{E}\left[\frac{1}{1-\frac{Z_t}{\sqrt{\lambda t}}}\mathbb{1}_{\{\sqrt{\lambda t}/2 \le Z_t \le \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}}\right] \le \frac{\sqrt{\lambda t}}{\lambda+1}2e^{-\lambda t/16}, \tag{15}$$

and notice that both expressions above converge to zero faster than $1/t$. For the second case, note first that $\mathbb{E}\left[Z_t\right] = 0$ and hence it follows from (11) and (13)–(15) that, as $t \to \infty$,

$$\mathbb{E}\left[Z_t \mathbb{1}_{\{-\sqrt{\lambda t} < Z_t < \sqrt{\lambda t}/2\}}\right] = o\left(\frac{1}{t}\right).$$

Then we use the Taylor expansion to obtain:

$$\left|\mathbb{E}\left[\frac{1}{1-\frac{Z_t}{\sqrt{\lambda t}}}\mathbb{1}_{\{-\sqrt{\lambda t} < Z_t < \sqrt{\lambda t}/2\}}\right] - 1\right|$$
$$\le \mathbb{E}\left[\frac{Z_t^2}{\lambda t}\right] + 2\mathbb{E}\left[\frac{|Z_t|^3}{(\lambda t)^{3/2}}\right] + o\left(\frac{1}{t}\right), \tag{16}$$

as $t \to \infty$. By the central limit theorem, $Z_t \xrightarrow{D} Z$ as $t \to \infty$. Furthermore, for $r > 0$, the convergence of moments holds [7]: $\lim_{t \to \infty} \mathbb{E}\left[|Z_t|^r\right] = \mathbb{E}\left[|Z|^r\right]$. In particular, in (16), $\mathbb{E}\left[|Z_t|^3\right]$ converges to a constant, and $\mathbb{E}\left[Z_t^2\right]$ converges to 1 as $t \to \infty$. Thus, using (10)–(12) and (3) we write

$$var\left(\frac{|E_t|}{|\mathcal{T}_t|}\right) = \mathbb{E}\left[\left(\frac{|E_t|}{|\mathcal{T}_t|}\right)^2\right] - \left(\mathbb{E}\left[\frac{|E_t|}{|\mathcal{T}_t|}\right]\right)^2$$
$$= \mathbb{E}\left[\frac{(t+1)(\lambda+1)}{\lambda t(1-\frac{Z_t}{\sqrt{\lambda t}})}\mathbb{1}_{\{Z_t \le \sqrt{\lambda t} - \frac{\lambda+1}{\sqrt{\lambda t}}\}}\right] - \frac{1}{\lambda} - \frac{1}{\lambda^2} + o\left(\frac{1}{t}\right).$$

Now, subsequently using (13) – (16), we get

$$
var\left(\frac{|E_t|}{|\mathcal{T}_t|}\right) = \frac{1}{\lambda}\frac{(t+1)(\lambda+1)}{\lambda t}\left(1+\frac{1}{\lambda t}+o\left(\frac{1}{t}\right)\right)
$$
$$
-\frac{1}{\lambda}-\frac{1}{\lambda^2}+o\left(\frac{1}{t}\right),
$$

which results in (4). Statement (5) is proved along similar lines: we apply the expansion directly to the random variable

$$
\frac{X_t}{t-X_t+1} = \frac{(t+1)(\lambda+1)}{(\lambda t+\lambda+1)(1-Z_t\frac{\sqrt{\lambda t}}{\lambda t+\lambda+1})}\mathbb{1}_{\{Z_t\le\sqrt{\lambda t}\}}-1,
$$

and then use the Chernoff bounds and the CLT to obtain the result.

A3. Proof of Lemma 3

Proof. Assume the arrival at time $t+1$ is of type $T1$. This occurs w.p. $\frac{\lambda}{\lambda+1}$, and then a new component consisting of size one is created in G_{t+1}, corresponding to the first case in (3).

Next, consider a $T2$ arrival, which occurs w.p. $\frac{p}{\lambda+1}$. We now add a node to an existing component C_i w.p. $\frac{|C_i|}{|V|}$. Thus the probability that we add the new node to C_i is $\frac{p}{\lambda+1}\cdot\frac{|C_i|}{|V|}$.

Last, we consider a $T3$ arrival. In this case we have two options. The new edge can either join two components, or join two nodes that are already in one component. For the first case, we derive the probability that C_i and C_j join as

$$
\mathbb{P}\left(C_i \text{ and } C_j \text{ merge}\right) = \frac{1-p}{\lambda+1}\cdot\frac{2\cdot|C_i|\cdot|C_j|}{|V|^2-|V|}.
$$

Then for the second case, the number of ways a T3 arrival links two nodes that are already connected in a component, say C_i, is $|C_i|(|C_i|-1)$. Therefore with probability $\frac{\sum_{k=1}^{m}|C_k|^2-|C_k|}{|V|^2-|V|}$ the component size does not change.

References

1. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)
2. Bhamidi, S., Steele, J.M., Zaman, T.: Twitter event networks and the superstar model, arXiv preprint ar Xiv:1211.3090 (2012)
3. Bhattacharya, D., Ram, S.: Sharing news articles using 140 characters: A diffusion analysis on Twitter. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM, pp. 966–971. IEEE (2012)
4. Bouma, H., Rajadell, O., Worm, D., Versloot, C., Wedemeijer, H.: On the early detection of threats in the real world based on open-source information on the Internet. In: International Conference on Information Technologies and Security, ITSEC (2012)

5. Chung, F., Handjani, S., Jungreis, D.: Generalizations of Polya's urn problem. Annals of Combinatorics **7**(2), 141–153 (2003)
6. Gómez, V., Kappen, H.J., Litvak, N., Kaltenbrunner, A.: A likelihood-based framework for the analysis of discussion threads. World Wide Web, 1–31 (2012)
7. Hall, P.: On the rate of convergence of moments in the central limit theorem for lattice distributions. Transactions of the American Mathematical Society **278**(1), 169–181 (1983)
8. Huberman, B., Romero, D., Wu, F.: Social networks that matter: Twitter under the microscope. Available at SSRN 1313405 (2008)
9. Klein, B., Laiseca, X., Casado-Mansilla, D., López-de-Ipiña, D., Nespral, A.P.: Detection and extracting of emergency knowledge from twitter streams. In: Bravo, J., López-de-Ipiña, D., Moya, F. (eds.) UCAmI 2012. LNCS, vol. 7656, pp. 462–469. Springer, Heidelberg (2012)
10. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the 19th International Conference on World Wide Web, pp. 591–600. ACM (2010)
11. Lanagan, J., Smeaton, A.F.: Using Twitter to detect and tag important events in live sports, pp. 542–545. AAAI (2011)
12. Lehmann, J., Gonçalves, B., Ramasco, J.J., Cattuto, C.: Dynamical classes of collective attention in Twitter. In: Proceedings of the 21st International Conference on World Wide Web, pp. 251–260. ACM (2012)
13. Paul, M.J., Dredze, M.: You are what you tweet: Analyzing Twitter for public health. In: Fifth International AAAI Conference on Weblogs and Social Media, ICWSM 2011 (2011)
14. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., Menczer, F.: Truthy: mapping the spread of astroturf in microblog streams. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 249–252. ACM (2011)
15. Rattanaritnont, G., Toyoda, M., Kitsuregawa, M.: A study on characteristics of topicspecific information cascade in Twitter. In: Forum on Data Engineering, DE 2011, pp. 65–70 (2011)
16. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 695–704. ACM (2011)
17. Sadikov, E., Martinez, M.M.M.: Information propagation on Twitter. CS322 Project Report (2009)
18. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
19. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (1998)
20. Zhou, Z., Bandari, R., Kong, J., Qian, H., Roychowdhury, V.: Information resonance on Twitter: watching iran. In: Proceedings of the First Workshop on Social Media Analytics, pp. 123–131. ACM (2010)
21. Zubiaga, A., Spina, D., Fresno, V., Martínez, R.: Classifying trending topics: a typology of conversation triggers on Twitter. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 2461–2464. ACM (2011)