On the Retweet Decay of the Evolutionary Retweet Graph

Giambattista Amati[1], Simone Angelini[1], Francesca Capri[2], Giorgio Gambosi[2],
Gianluca Rossi[2], and Paola Vocca[3](✉)

[1] Fondazione Ugo Bordoni, Rome, Italy
{gba,sangelini}@fub.it
[2] University of Rome Tor Vergata, Rome, Italy
francesca.capri.ext@mise.gov.it,
{giorgio.gambosi,gianluca.rossi}@uniroma2.it
[3] University of Tuscia, Viterbo, Italy
vocca@unitus.it

Abstract. Topological and structural properties of social networks, like Twitter, is of a major importance in order to understand the nature of user activities, for example how information propagates or how to identify influencing accounts. A deeper analysis of these properties may have a crucial impact on the design of new applications and of existing ones.

In a social network there are different relations among nodes that can be defined and analyzed by keeping track of how the generated links evolve over time. So far, all evolutionary studies analyze the graph in a *cumulative way*, that is, once a link is inserted in a graph it is never eliminated [9,12]. However, in social networks like Twitter interactions are more volatile, and after a period of life they should die.

In this paper, we consider the *Retweet Graph*, where links are generated by the retweet action made by an user. The life of a tweet is limited in time, and it spans from the time it is generated, to the last time it is retweeted. To take into account the dynamics of Twitter users, we consider a model in which, when a tweet expires, we delete all the edges representing the retweet action relative to this tweet and all users corresponding to involved nodes become inactive, unless they are alive with respect to a different retweeting activity. In particular, we define a new version of the usual Retweet Graph, the *Dynamic Retweet Graph (DRG)*: when a tweet has been retweeted for the last time all the edges related to this tweet are deleted. This allows to model the decay of tweet relavance in Twitter. To evaluate the structural properties of a DRG, we consider three different Twitter streams, derived by monitoring the Twitter flow on three different contexts: two of them are based on a specific event (the 2015 Black Friday and the 2015 World Series) while the third is the Firehose of the whole Twitter stream, filtered by the Italian language.

We study the differences between the DRG graphs and the corresponding cumulative ones by comparing standard metrics for social networks, such as average distance, clustering coefficient, in-degree and out-degree distributions. The analysis shows an important difference

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2017
O. Gaggi et al. (Eds.): GOODTECHS 2016, LNICST 195, pp. 243–253, 2017.
DOI: 10.1007/978-3-319-61949-1_26

between the cumulative graphs and the corresponding DRGs, both on the way they grow, and on the way the observed measures evolve.

Keywords: Graph analysis · Social media · Twitter graph · Retweet graph · Graph dynamics

1 Introduction

The analysis of the topological characteristics of graphs derived from social community systems has a remarkable significance either to derive social properties (the degree of separation, the community detection, the degree distribution, etc.) or to analyze the information flow (how data spread over the network, which are the authoritative users, etc.). Twitter is largely dissimilar to other kind of social networks mainly because all information is open to everyone. On the contrary, on Facebook users adopt very restricted privacy policies and accessibility is only once a friendship relation is established. Additionally, Twitter lets users act freely, e.g. by retweeting posts, following or mentioning any account, and such interactions among accounts result in different kinds of networks [1].

The most studied relationship is the follower/following one, also known as the *Follow Graph*, and is obtained by yielding a directed edge from a vertex a to a vertex b when a follows b. However, it is an almost static form of relation characterized by a small effective diameter (the 90% of reachable pairs have a minimum distance less than this value), a non-power-law in-degree distribution (the followers distribution), and low reciprocity, which overall marks a deviation from known features of other social networks [8]. More generally, the follow graph shows structural properties of both a social and an information network [13].

The follow graph has also been studied in [7] in order to identify authoritative accounts.

The Twitter dataset is prohibitive to crawl on a huge scale because of the very restrictive policy of Twitter's API, therefore the analysis of the follow graph is impractical, and, moreover, it cannot be used to fully describe user interests, because the follow graph does not contain temporal or text information [13].

A different kind of network induced by Twitterverse is the *Retweet Graph*. A Retweet Graph is described as a directed graph, where nodes are accounts and edges between nodes are when one retweets the second. Studies on the retweet graph can be found in [2,4,6,15,16].

The graph representation of a social network is frequently used to assess the temporal evolution of the network, and there are several mathematical models that predict the growth and the trends evolution [4,15,18]. In these models the graph representing the social network is considered in a cumulative way, that is once an edge or a vertex is inserted it is never eliminated, even when the relation or the account they represent does not exist anymore (e.g. the two accounts terminate the following/follower relationship, the retweeted tweet is obsolete, etc.). Whilst this assumption may be still plausible in the case of the following/follower relationship, which is inherently static because, the deletion of a following/follower link takes place less often, and thus a cumulative evolutionary graph model may be reasonable, it is largely unrealistic when considering more dynamic relations as the retweeting one.

In this paper, in order to take into consideration the dynamics of the Twitterverse, we introduce a variant of the retweet graph, that is, the *Dynamic Retweet Graph (DRG*, for short*)*. The graph is built as follows: when a tweet is retweeted for the last time, all edges associated to this tweet are removed. This assumption would model the expiration of a tweet in Twitter.

The DRG behavior is tested against three distinct Twitter collections, derived by monitoring the activities in three distinctive contexts: two collections refer to a specific event (*event driven*), that is, they are derived by filtering the Twitter stream using a set of words related to these specific events (the first one refers to the 2015 Black Friday whilst the second the 2015 World Series); the third one is the whole Italian Twitter stream, filtered by the Italian language, denoted *Italian Firehose*. To derive the Italian Twitter Firehose we use a list of the most frequently used Italian stop words and the Twitter native selection function for languages.

We observed the evolution of the DRGs for two months and compute the basic structural measures that are normally used to evaluate social networks. Then we compare results on DRGs with their corresponding cumulative graphs [2], in terms of clustering coefficient, in-degree and out-degree distributions, average distance.

Results point out a substantial difference between the DRGs and the relative cumulative graphs both on how they grow and the way the structural measures evolve. Only the Italian Firehose shows a similar behavior of the cumulative and the dynamic graph, whilst for the event driven graphs some properties show different trends. In particular, in the case of the Italian Firehose retweet graph the sizes of the edges and vertices sets and the considered measures do not show discontinuity in their growth, while the event driven graphs show a skewed distribution and reach a predictably saturation point at end of the event. The measures for the firehose retweet graph are more similar to a social network, such as Facebook, than to an information network, whilst the event driven retweet graphs is the opposite.

The study of the temporal evolution of the retweet graphs is a preliminary work in order to better understand the nature of Twitter, how trends evolve over time, to detect both authoritative and spamming accounts, and to derive a suitable mathematical evolutionary model of Twitter communities.

1.1 Related Work

A large literature on the evolution of Twitter [8,13] compare Twitter to other social networks, or analyze the evolution over time of the Twitter social graph in order to model topic trends [5,17] or model just the twitter network [5,14,15,17,18]. One of the open problem is to assess the social nature of Twitter, whether it is more a social network than a social media, or is both showing a double nature of the Twitterverse [8,13]. On the other hand, the evolutionary behavior of Twitter is mostly studied for trends analysis. News streams in Twitter, as well as in many other news media, show a star-like shape [5]. This shape is shown by means of a dataset gathered on Twitter during the Iranian election on the 2009 [17], reflecting the tendency that flows spread widely and not deeply. Starting from the Superstar model, that represents the condensation phenomenon occurring in the largest component of a retweet graph [4], a evolutionary mathematical model for the retweet graph based on the density distribution of edges, and the density of the largest connected component [14,15]. Finally, a classification method allows to rapidly identify categories of trends by means of different triggers that starts trends, through [18].

2 Graph Construction and Evolution

A DRG $G = (V, E, \ell)$ is a graph where the vertex set V is made of Twitter accounts and a direct edge $(a, b) \in E$ exists if and only if a has retweeted at least one tweet of b, that can be itself already a retweet. Since an user a may retweet many tweets of b we maintain them distinct labeling the edge (a, b) with the id of the original tweet and the timestamp on which this retweet occurs. Labels are represented with a list $\ell(e)$ associated to each edge $e = (a, b)$, consisting of pairs (i, t) where i is the id of a tweet and t is the timestamp in which a retweets i from b. The pairs of $\ell(e)$ are sorted for timestamps in non-decreasing order.

Starting from the information represented, for each tweet i we define the *date of birth* denoted by $\texttt{dob}(i)$, as the timestamp of the first retweet of i. Dually, the *date of death* of i, denoted $\texttt{dod}(i)$, is the timestamp of the last retweet of i. In a formal way,

$$\texttt{dob}(i) = \min_{e \in E} \{t : (i, t) \in \ell(e)\}$$

and

$$\texttt{dod}(i) = \max_{e \in E} \{t : (i, t) \in \ell(e)\}.$$

Therefore, a tweet i is *alive at time* t if and only if $\texttt{dob}(i) \leq t \leq \texttt{dod}(i)$.

A node $v \in V$ is *alive at time* t if and only if there is a tweet connecting the node v that is alive.

With the above definitions we can derive a sequence of DRG graphs G_t based on time variation. Let $V_t \subseteq V$ be the set of alive nodes of V at time t. The graph $G_t = (V_t, E_t)$ at time t is a subgraph of G, when E_t all edges e are relative to alive tweets. For simplicity, we give an example. If G is the retweet graph (see the left side of Fig. 1) then G_{20} coincides with G since the tweets with labels 1 or 2 are alive before the timestamp 20, according to periods in the left part of Fig. 1. Differently, E_{35} contains only edges (c, a) and (a, b).

We set a 4 hours interval in our experiments, that is the series of DRGs are $(G_{t_i})_{i \geq 0}$ and $t_{i+1} - t_i = 4$.

In order to compare results with the corresponding cumulative graphs we use the same data sets as in paper [2]. Among the three different collections, two are *event driven*, that is filtered by using a set of words specifying two events,

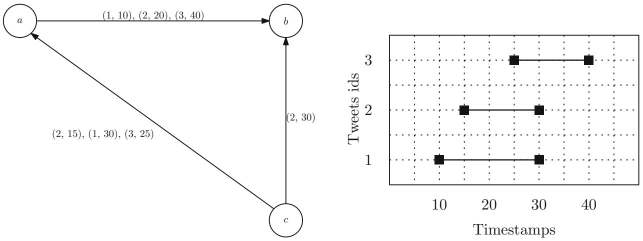


Fig. 1. On the left, a retweet graph with edges labeled with $\langle i, t\rangle$. On the right, a the life-time interval are represented.

the 2015 Black Friday and the 2015 World Series; the last is the whole Italian Twitter stream, also called the *Italian Firehose*. We filtered by the most frequent Italian stop-words and by the Twitter native selection function for languages. Table 1 shows the sizes of the cumulative retwet graphs of the three collections.

Table 1. Dimensions of the three cumulative graphs

	Black friday	World series	Italian firehose
Vertices	2.7e+06	4.74e+05	2.541739e+06
Edges	3.9e+06	8.4e+05	1.3708317e+07
Tweets/edges	2.603	2.3	5.45
Tweets/vertices	3.66	4	29.4

Figure 2 describes the dimensional evolution of the three graphs over the observation.

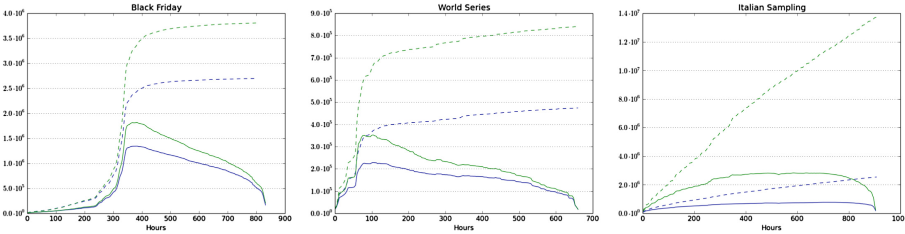


Fig. 2. Vertex number (blue) and edge number (green) of the Italian Firehose, World Series, and Black Friday as a function of hours. Solid lines represent the trend of the DGR, dashed lines represent the trend of the cumulative graphs

Figure 3 shows that the graphs densify over time, having the size edge set growing more than the node set. A densification law for edges with respect the number of nodes, known as *Densification Power Law (DPL)* also holds for all DRG graphs and follows a power-law [10,11].

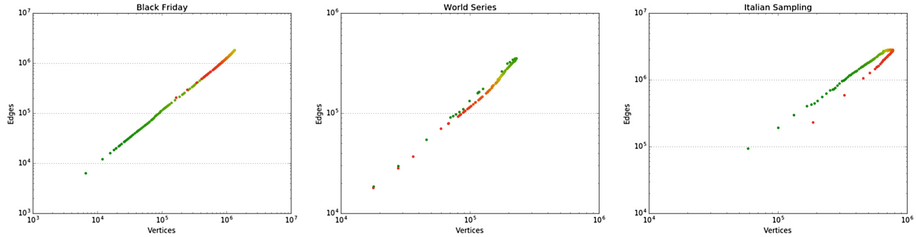


Fig. 3. Densification Power Law holds for all three DRG graphs (Black Friday coefficient = 1.06, World Series coefficient = 1.18, Italian Firehose coefficient = 1.32). The color change from green to red in accordance with the timestamp growth. When approaching the end of the stream the edge number decay is more rapid than node number decay in the case of the Italian Firehose.

Note that the event-driven graphs and the Firehose graph evolve differently: the event-driven graphs rapidly grow near the event time and afterward they have a gradual decline. On the contrary, we note that the Twitter Italian Firehose DRG graph has a slower growth and a more rapid decline, maybe due to the border effect. The DPL shows that the Firehose retweet graph behaves two ways (green line and red line): with the green line it increases, and, afterwards, with the red line it declines abruptly reaching the very last timestamp.

For what concern the event-driven graphs, the rapid growth near the event can be easily established with the growing interest for the event itself. Moreover, the slow decline suggests the diminished interest.

Concerning the Firehose graph, the increase and the decline is because of "border effects". Beginning from empty graph G_0, we have a sequence of increasing sizes before reaching a stable configuration. In a similar way, while reaching the final state G_{final} in which the graph is again empty, the stable configuration starts to decay. At this point the curve does not decrease vertically due to the fact the date of death is the final time the tweet is retweeted.

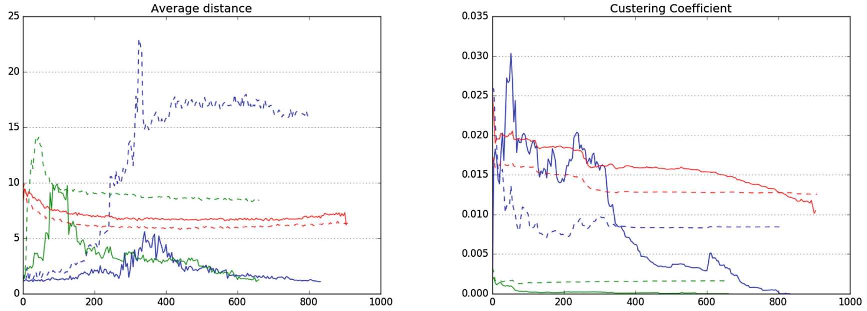
3 Measures Description and Trends

3.1 Average Distance

The average distance $Avg(G)$ is defined as follows:

$$Avg(G) = \frac{\sum_{d \geq 1} d \cdot \text{number of pair of vertices at distance} \quad d}{\text{number of connected pairs of vertices}}$$

Figure 4a shows the evolution of the average distances. Again, the event-driven graphs are different to the Firehose graph. In this last graph the average distance



(a) (b)

Fig. 4. Trend of the average distance (a) and cluster coefficient (b) in the three observed retweet graphs: Black Friday in blue; World Series in green and Italian Firehose in red. The DRGs are represented with solid lines and the corresponding cumulative graphs with dashed lines.

is almost constant while in the event driven case we have a climax in proximity of the event.

The Italian Firehose has the same distance tendency and average distance as the corresponding cumulative graph. On the contrary, event-driven DRG graphs are unstable and both growth and decay are very steep having reached a peak and that they do no longer converge. Moreover, the average distance value is much smaller than the corresponding cumulative graphs. On the contrary, event-driven DRG graphs are very unstable and the growth and the decay are very rapid reaching a peek and they do not converge. In addition, the average distance magnitude of event-driven DRG graphs is much smaller than the corresponding cumulative graphs.

3.2 Clustering Coefficient

Figure 4b shows the global clustering coefficient trend. Barrat and Weigt [3] introduced the global clustering coefficient in the physical and mathematical context as largely used in social sciences. Clustering coefficient correlates to the probability of obtaining a triangle of connections from two consecutive connecting edges, in other words, the clustering coefficient is the probability that a friend of your friend is likely to be your friend. Mathematically:

$$C = \frac{3 \times \text{number of triangles in the network}}{\text{number of paths of length 2}}$$

The Italian Firehose DRG's clustering coefficient is similar to that of the cumulative graph. Black Friday's DRG is different to the cumulative retweet graph, but the World Series has the DRG and the cumulative graphs following a similar trend. This is mainly because the observation of the World series event begun when it was already started.

3.3 Out-Degree and In-Degree Distributions

Figure 5 represents the distribution of the in-degrees with respect to a specific G_t. The y-axis represents the number of nodes having the in-degree in the x-axis. For event driven graphs, the timestamp t is close to the event, while for the Italian Firehose the timestamp is in the middle of the interval. For both axis we use the logarithmic scale.

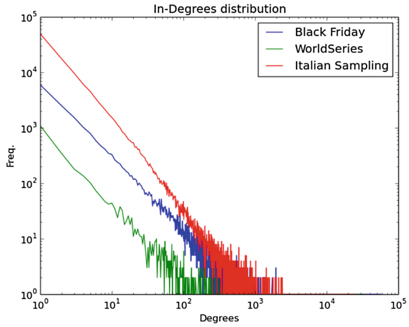


Fig. 5. The in-degrees distribution of a DRG at a specific timestamp.

From Fig. 5 it follows that the in-degrees distributions of all the three graphs attain a power-law distribution. The same trend can be observed both with different time-stamps and for the out-degree distribution, whose plots are not reported here for lack of space. Figure 6 reports the trend of the power-law exponents of the in- and out- degrees distribution.

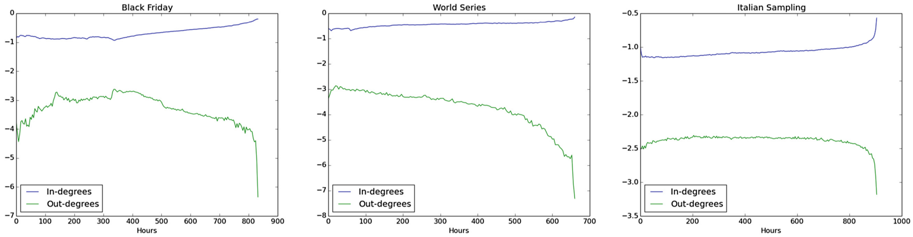


Fig. 6. Power-law exponents of the out-degrees (green) and of the in-degrees (blue) distributions of the World Series, Black Friday and Italian Firehose as functions of hours. (Color figure online)

4 Discussion and Conclusions

There are well known properties that actual graphs fulfil, which include heavy tails for the out-degree and in-degree distribution, shrinking average distance and diameters, and the Densification Power Law (DPL) [10,11]. The retweet graphs of all collections fulfill all these properties [2]. In this paper, we have analyzed the tweet lifetime and its natural temporal decay, by considering only the subgraphs (DRG) that remain active during the stream, and, thus, by cutting off those elements (accounts and interactions) in the graph that become obsolete in the whole network. The analysis performed shows a substantial difference in the evolution for the two classes of graphs: the event driven (both the Black Friday and the World series) and the Italian Firehose. It is interesting to note that for the Italian Firehose, which can be seen as the superposition of a sequence of event-based subgraphs, the DRGs evolution has the same structural properties of the cumulative graph. Moreover, in the case of the average distance and of the clustering coefficient no change occurs, which substantiates that the Firehose is indeed the result of the union of different communities induced by different event-driven streams. On the contrary, for the event-based graphs we have a border effect on all structural metrics, all growing and decaying in a similar way, and reaching a climax in the middle of their lifetime interval with a value much smaller than the values of their corresponding cumulative graphs. Additionally, all the structural properties valid for the cumulative graphs, with the exception of the Densification Power Law and Degree Power Laws, do not hold for the DRGs. In fact, the average distance and the clustering coefficient tend to converge superlinearly. Note that, from the analysis performed, the model proposed (DRG) better captures the different evolutionary behavior of the two kinds of retweet graphs (event-driven and the firehose). In fact, in [2], the topological measures on the cumulative retweet graphs without taking into account the edges decay, do not show a substantial difference. Additionally, the generative models proposed in literature [4,14,15] fail in representing the edges decay in the event-driven retweet graphs. Hence, a promising future research direction is to define a tighter mathematical model.

Acknowledgments. This work was conducted in the Laboratory of Big Data of ISCOM-MISE (Institute of communication of the Italian Ministry for Economic Development). Francesca Capri was supported by a grant of ISCOM-MISE.

References

1. Amati, G., Angelini, S., Bianchi, M., Fusco, G., Gambosi, G., Gaudino, G., Marcone, G., Rossi, G., Vocca, P.: Moving beyond the Twitter follow graph. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, KDIR 2015, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), Lisbon, Portugal, 12–14 November 2015, vol. 1, pp. 612–619 (2015)

2. Amati, G., Angelini, S., Capri, F., Gambosi, G., Rossi, G., Vocca, P.: Twitter temporal evolution analysis: comparing event and topic driven retweet graphs. In: Proceedings of the International Conference on Big Data Analytics, Data Mining and Computational Intelligence, BIGDACI 2016, Funchal, Madeira, Portugal, 2–4 July 2016, vol. 1 (2016)
3. Barrat, A., Weigt, M.: On the properties of small-world network models. Eur. Phys. J. B-Condens. Matter Complex Syst. **13**(3), 547–560 (2000)
4. Shankar Bhamidi, J., Steele, M., Zaman, T., et al.: Twitter event networks and the superstar model. Ann. Appl. Probab. **25**(5), 2462–2502 (2015)
5. 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)
6. Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M.: Aggregate characterization of user behavior in Twitter and analysis of the retweet graph. ACM Trans. Internet Technol. **15**(1), 4 (2015)
7. Java, A., Song, X., Finin, T., Tseng, B.: Why we Twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, WebKDD/SNA-KDD 2007, pp. 56–65. ACM, New York (2007)
8. 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, WWW 2010, pp. 591–600. ACM, New York (2010)
9. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., Ghahramani, Z.: Kronecker graphs: an approach to modeling networks. J. Mach. Learn. Res. **11**, 985–1042 (2010)
10. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD 2005, pp. 177–187. ACM, New York (2005)
11. Leskovec, J., Kleinberg, J., Faloutsos, C.: Graph evolution: densification and shrinking diameters. ACM Trans. Knowl. Discov. Data **1**(1), 2 (2007)
12. Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C.: Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In: Jorge, A.M., Torgo, L., Brazdil, P., Camacho, R., Gama, J. (eds.) PKDD 2005. LNCS, vol. 3721, pp. 133–145. Springer, Heidelberg (2005). doi:10.1007/11564126_17
13. Myers, S.A., Sharma, A., Gupta, P., Lin, J.: Information network or social network?: the structure of the twitter follow graph. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion, WWW Companion 2014, Republic and Canton of Geneva, Switzerland, pp. 493–498. International World Wide Web Conferences Steering Committee (2014)
14. ten Thij, M., Bhulai, S., Kampstra, P.: Circadian patterns in Twitter. In: Data Analytics, pp. 12–17 (2014)
15. ten Thij, M., Ouboter, T., Worm, D., Litvak, N., Berg, H., Bhulai, S.: Modelling of trends in Twitter using retweet graph dynamics. In: Bonato, A., Graham, F.C., Prałat, P. (eds.) WAW 2014. LNCS, vol. 8882, pp. 132–147. Springer, Cham (2014). doi:10.1007/978-3-319-13123-8_11
16. Yang, M.-C., Lee, J.-T., Lee, S.-W., Rim, H.-C.: Finding interesting posts in Twitter based on retweet graph analysis. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012, pp. 1073–1074. ACM, New York (2012)

17. 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)
18. Zubiaga, A., Spina, D., Martinez, R., Fresno, V.: Real-time classification of Twitter trends. J. Assoc. Inf. Sci. Technol. **66**(3), 462–473 (2015)