IADIS International Journal on Computer Science and Information Systems
Vol. 11, No. 2, pp. 19-30
ISSN: 1646-3692

MODELLING THE TEMPORAL EVOLUTION OF THE 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.
[2]Univ. of Rome Tor Vergata, Rome, Italy
[3]Univ. of Tuscia, Viterbo, Italy.

ABSTRACT

Topological properties of graphs derived from social network platforms, like Twitter, give important insights on the nature of the social activities or on the way information spreads over the network. It may have also a relevant impact on designing new applications and improving already existing services. Different types of relations among the nodes define different graphs that can be analyzed, by tracking how relations evolve over time. Usually, this is performed in a cumulative way: once an edge is inserted, it is never deleted, see Leskovec et al. (2005) and Leskovec et al. (2010). However, the tweet life is limited, spanning from its birth to the very last retweet it receives. Therefore, we want to analyze the dynamics of evolutionary graphs, that is deleting tweets and thus edges among the nodes when they naturally expire as well as accounts that become therefore inactive. We introduce a variant of the retweet graph which takes into account the dynamics of Twitter users: *Dynamic Retweet Graph (DRG)*. In a DRG, once a tweet has been retweeted the last time all the edges representing this tweet are deleted, to model the decay of tweet life in the social platform. We analyze the characteristics of this graph using three different Twitter streams, built on three different contexts: two are event based (the 2015 Black Friday and the 2015 World Series), the third is the firehose of the whole Twitter stream, filtered by the Italian language. We use some standard social network analysis metrics to compare the structural properties of the DRG graph with cumulative evolving graphs.

KEYWORDS

Graph analysis, social media, Twitter graph, retweet graph, graph dynamics

1. INTRODUCTION

The study of the topological properties of graphs derived from social network platforms has a great importance from both social and information point of views. Twitter has specific characteristics that makes it substantially different from other social networks such as, for example Facebook, because of its openness to account interactions. Moreover, Twitter allows different users activities such as following other users, retweetting posts, mention or hashtagging other users, and such interactions between users induce a new kind of network, Amati et al. (2015). The following/follower graph (from now on will be denoted as *follow graph*) is the most studied: it represent a relatively static type of relation, and is obtained by associating nodes to users and assuming a directed edge from a node *a* to a node *b* if *a* follows *b*. The first quantitative study of the follow graph, Kwak et al. (2010), have found a non-power-law for the follow distribution, a short effective diameter, and low reciprocity, which overall marks a deviation from known characteristics of human social networks. These outcomes were successively strengthen, Myers et al. (2014), by the observation that the Twitter follow graph exhibits structural characteristics of both an information network and a social network. Other work, Java et al. (2007), studied the follow graphs to identify authoritative accounts.

Unfortunately, the follow graph datasets are prohibitive to crawl on a massive scale due to the very restrictive policy of Twitter, and, additionally, it could be not so meaningful for describing the Twitter behavior since the follow graph may not completely explain how information spreads over the network, Myers et al. (2014).

Another kind of network derivable from the Twitterverse is the *Retweet graph*. A Retweet graph is defined as a directed graph where nodes are accounts and edges between accounts *a* and *b* is set when *a* retweets a tweet of *b*. Also the retweet graph has been widely studied see Yang et al. (2012), Build et al. (2015), Ten-Thij et al. (2014), Amati et al. (2016) to cite a few.

Graph representation of a network are often used to evaluate the temporal evolution of the network, and mathematical models are derived to predict the network growth and the trends evolution, see for example Ten-Thij et al. (2014), Bhamidi et al. (2015), Zubiaga et al. (2015). All these works consider the graph growth of the Twitterverse in a cumulative way: once an edge is inserted, it is never deleted. While this approach could be reasonable when considering a more static relationship such as the follow one (the deletion of a follow link occurs rarely and hence, a cumulative follow network is a good evolutionary model), it is particularly unrealistic when more dynamic relations, such as retweets, are also considered. In this paper, we introduce a variant of the retweet graph which takes into account the dynamics of Twitter users: *Dynamic Retweet Graph* (DRG, for short). In a DRG, once a tweet has been retweeted for the last time all the edges representing this tweet are deleted, to model the expiration of a tweet in a stream of the social medium. In contrast to the DRG, in the Cumulative Retweet Graph (in short CRG), vertices and edges once inserted will no longer be removed.

We analyze the characteristics of these graphs using three different Twitter collections, built by monitoring the activities in three different contexts: two such collections are event driven (related to the 2015 Black Friday and the 2015 World Series), while the third one is obtained from the overall Twitter stream, filtered by language (Italian), denoted *Italian Firehose*. To obtain the Italian Twitter Firehose we use a list of the most used Italian stop-words and the Twitter native selection function for languages.

We analyze the evolution of the DRG over a period of two months, and compare the main structural measures that are generally used to characterize the nature of graphs with the ones derived from the same datasets considering CRG: average distance, clustering coefficient, in- and out-degree distribution;, number of strongly connected components, size of biggest strongly connected component. See also Amati et al. (2016) for a preliminary analysis on CRG.

Results show a significant difference between CRG and the corresponding DRG, both in the way they grow, and in the way the above measures evolve. We have seen that only the DRG for the Italian Firehose dynamically maintains the same structural properties of the CRG, whilst the event based do not preserve some structural properties.

In this paper, after a fast survey on related bibliography (Section 2), in Section 3 we formally define the DRG and the CRG together with other notions that are used in the paper. In Section 4, we analyze the evolution of the graphs by using the measures that we have previously describes. We close the paper with some final considerations in Section 5, here we also describe some interesting problems still left open.

2. RELATED WORK

There is a large literature on Twitter social network evolution: Kwak et al. (2010) and Myers et al. (2014) compare Twitter with other social networks; Bhattacharya and Ram (2012) and Zhou et al. (2010) study the temporal evolution to model topic trends; Bhattacharya and Ram (2012), Zhou et al. (2010), Ten-Thij et al. (2014), Ten-Thij et al. (2015) and Zubiaga et al. (2015) that assess authoritative users. The analysis to assess the social nature of Twitter, whether it is a social network or a social media, is not conclusive since both can be explained, see Kwak et al. (2010) and Myers et al. (2014). On the other side, the temporal evolution of Twitter is mainly studied for trends analysis. The diffusion of news in Twitter and in several popular news media show a star-like phenomenon of the information flows, Bhattacharya and Ram (2012). Similar results are derived for the diffusion of information on Twitter during the Iranian election on 2009, Zhou et al. (2010). The results showed that the flows tend to be wide, not too deep and their size follow a power law-distribution. In Bhamidi et al. (2015), the authors proposed and validated the superstar random graph model to represent the condensation phenomenon represented by the largest component of the retweet graph. Based on this approach, Ten-Thij et al. (2014) and successively Ten-Thij et al. (2015) define a mathematical model that describes the evolution of a retweet graph on some basic characteristics, such as the density of edges and the size and density of the largest connected component. In Zubiaga et al. (2015), the authors explore the types of triggers that spark trends on Twitter, through a categorization that allows to quickly identify types of trend.

3. GRAPH CONSTRUCTION

The DRG (Dynamic Retweet Graph) $G = (V, E, L)$ is defined as follows: the set V of nodes are Twitter accounts and an edge e in E represents an interaction (a retweet) between two accounts. In particular, there is a directed edge from an account a towards an account b, if a has retweeted at least one tweet of b, that can be itself already a retweet. Observe that user a may retweet more tweets of b. For this reason we keep distinct all edges (a, b) by attaching the id of the original tweet and the timestamp on which this retweet occurs. This edge information is implemented with a list $L(e)$ associated to every edge $e = (a, b)$ that contains 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 $L(e)$ are sorted for timestamps in non-decreasing order.

From the data that we have collected in *G* we define, for all tweets *i*, the *date of birth* of *i* (in short, *dob(i)* is the timestamp of the first retweet of *i*, similarly, the *date of death* of *i* (in short, *dod(i)*) is the timestamp of the last retweet of *i*. Formally,

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

and

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

A tweet with id *i* is *alive at time t* if and only if $dob(i) \leq t \leq dod(i)$.

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.

By using these definitions we construct a time series of DRG $\{G_t\}$ Let *t* be a timestamp, we define the subgraph $G_t = (V_t, E_t)$ at time *t* where E_t contains any edge *e* in *E* of alive tweets, that is, *e* is in E_t if and only if at least one of the tweets in the list *L(e)* is alive at time *t*; V_t is the set of alive nodes of E_t.

For example if *G* is the DRG represented in the left part of Figure 1, V_{20} contains all nodes of *G* since all nodes are alive, because of the tweets with ids 1 or 2 that are the only tweets alive before the timestamp 20, and similarly E_{20} contains all edges of *G* (see the left part of Figure 1. On the contrary E_{35} contains only edges *(a, b)* and *(c, a)*

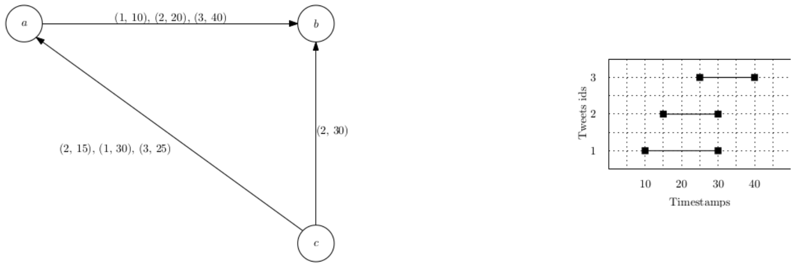

Figure 1. On the left side, an example of a DRG. Edges are labeled by pairs with the id of the tweet and the timestamp of the retweet. On the right side, there is a graphical representation of the life-time of the tweets

In a similar way we denote by CRG (which stands for Cumulative Retweet Graph) the graph of retweets that grows in cumulative way (once an edge is inserted, it is never deleted). For this graphs a tweet *i* is alive at time *t* if and only if $t \geq dob(i)$.

In our experiment settings we study the properties of the sequence of graphs $\{G_{t(i)}\}_{i \geq 0}$ where $t(i+1) - t(i)$ is 4 hours.

For our experiments we use a dataset that consists in two different classes of retweet graphs: the event driven retweet graph, filtered by topics about specific events (i.e. the Black Friday 2015 and the World Series 2015) and the firehose retweet graph, filtered by the Italian language from the whole Twitter stream. To obtain the Italian Twitter Firehose we use a list of the most used Italian stop-words and the Twitter native selection function for languages. In Table 1 it is shown the dimensions of the three graphs.

Table 1. Dimensions of the Final graphs

	Italian Firehose	Black Friday	World series
Vertices	2.541739e+06	2.7e+06	4.74e+05
Edges	1.3708317e+07	3.8e+06	8.40e+05
Tweets/edges	5.45	2.603	2.3
Tweets/vertices	29.4	3.66	4

In Figure 2 we show the evolution of the dimensions of the three datasets over the period of observation. The figure shows both the trends of the CRG and the DRG..

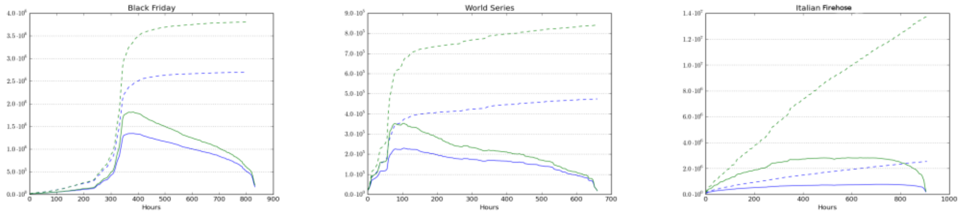

Figure 2. Number of vertices (blue) and number of edges (green) of: Black Friday, World Series and Italian Firehose as functions of hours. With dashed lines are represented the CRG. and with the solid lines are represented the DRG

Most of the graphs densify over time, with the number of edges growing more than the number of nodes, and this densification follows a power-law pattern (the Densification Power Law, DPL), see Leskovec et al. (2005) and Leskovec et al. (2007). This behavior can be found in the growth of the CRG of our three datasets, see Figure 3.

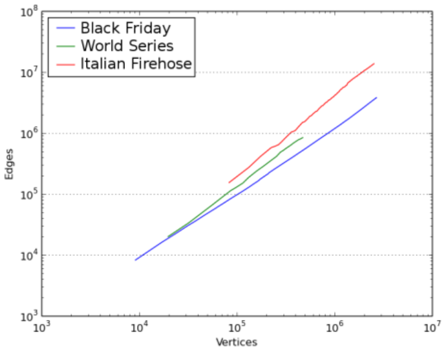

Figure 3. The densification of the three CRG in log-log scale

Figure 4 shows the densification of the three DRG, here we can observe that DPL also holds for all DRG graphs.

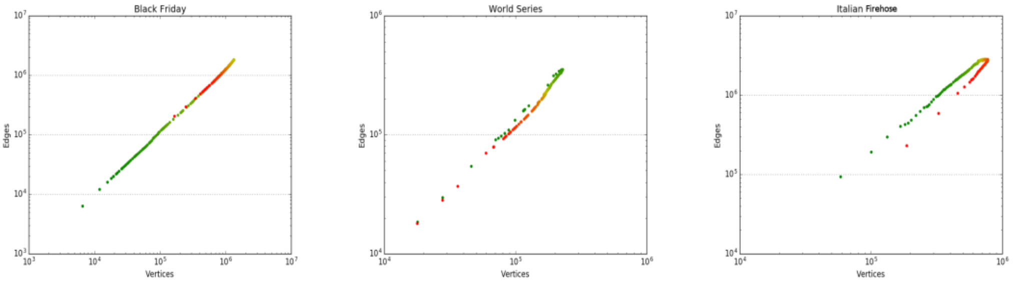

Figure 4. Densification Power Law holds for all three DRG graphs. The gradient color goes from green to red along with the timestamp growth. In the case of the Italian Firehose, when approaching the end of the stream the number of edges decays more quickly than the number of vertices

In Table 2 are reported the coefficients of the power laws that explain the densification of our datasets.

Table 2. Power law coefficients

	Black Friday	World Series	Italian Firose
CRG	1.10	1.19	1.34
DRG	1.06	1.18	1.32

Referring to Figure 4, it is interesting to note that the event-driven graphs and the firehose graph evolve in two different way: the event-driven ones show a rapid growth close to the event, and then a slow decline. On the contrary, the Twitter firehose graph have a slower growth and a rapid decline. The DPL trends indeed shows that the Twitter firehose graph follows two lines: initially it follows the green line, going up, and then it turns downwards, as soon as it approaches the final timestamp, with a steeper red line.

About the event-driven graphs, the rapid growth in proximity of the event is justified by the interest for that event. And also the gradual loss of interest explains the slow decline.

Regarding the Italian Firehose DRG, the growth and decline is due to "border effects". Starting from G_0, the empty graph, we observe a number of intermediate sizes before reaching a stationary configuration. Similarly, approaching G_{final}, the final empty graph, the stationary configuration starts to decay. The tail effect is due to the death of tweets born near to the end of time listening window. The final part of the curve is not vertical because the date of death is the last time the tweet is retweeted.

4. GRAPH EVOLUTION

In our analysis we have considered a number of measures both for the CRG than for the DRG and for the three datasets. These measures include, in addition to the number of vertices and edges (NumVertives and NumEdges), the following ones: the maximum in- and out-degree of vertices (MaxInDegree, MaxOutDegree); number of strongly connected components (NumCCs); the size of the biggest strongly connected component (MaxCCsize); the average distance between vertices (AverageDist) and the clustering coefficient (ClusterCoeff). These measures may or may not be statistically correlated. In Figure 5 are shown the Pearson correlation coefficients between the observed measures. We note that NumVertices, NumEdges, MaxInDegree, MaxOutDegree, NumCCs and MaxCCsize are strongly related in the case of CRG. This behavior does not occur in DRG..

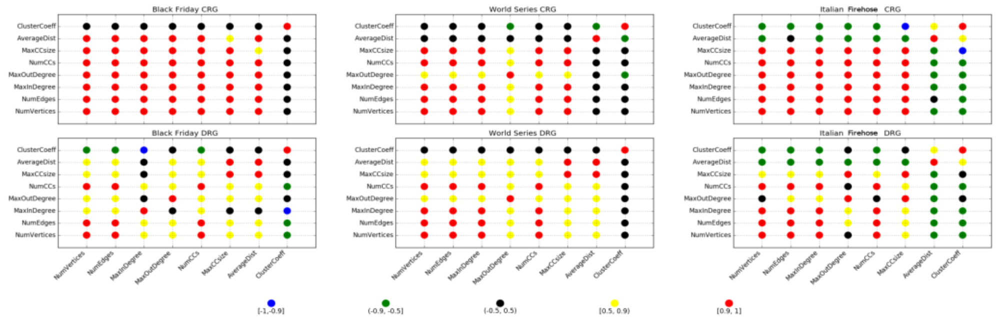

Figure 5. Pearson correlation coefficients between the observed measures. The coefficients are represented by colors

4.1 Average Distance

The average distance is obtained by considering the distances between all the connected pairs of vertices. Let *d* be an integer, *N(d)* is defined as the number of pairs of vertices of *G* at distance exactly *d*. Then, the average distance, *Avg(G)*, of *G* is

$$Avg(G) = \sum_{d \geq 1} d \cdot N(d)/S.$$

Where *S* denotes the number of connected pairs of vertices, that is $S = \sum_{d \geq 1} N(d)$.

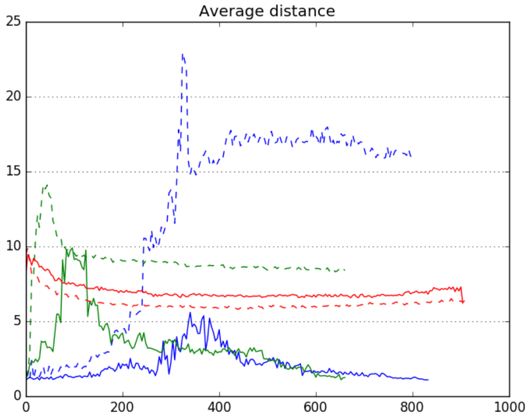

Figure 6. Trends of the average distance for the CRG (dashed lines) DRG (solid lines) for the three datasets: Black Friday in blue; World Series in green and Italian Firehose in red

The trend of the average distances over the time is shown in Figure 6. In the Italian Firehose the average distance is almost constant and shows the same trend and the same magnitude in the CRG and DRG: in particular the CRG have an average distance slightly smaller than the DRG. On the contrary, event-driven graphs are very unstable and growth and 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 conresponding CRG..

4.2 Clustering Coefficient

As a second feature we considered the evolution of the global clustering coefficient as widely used in social science and introduced by Barrat and Weigt (2000) in the mathematical and physical literature. The global clustering coefficient quantifies the probability that if a vertex *a* is connected to vertex *b* and vertex *b* is connected to vertex *c* then the vertex *a* will also be connected to vertex *c*. In other words, the probability that the friend of your friend is likely also to be your friend. Thus, let T be the number of triangles in the network and let P be the number of path of length 2, the clustering coefficient be quantified as follows: $C = 3{\cdot}T/P$. *M*ost social networks are characterized by relatively hight clustering coefficients: in particular on those social networks the global clustering coefficient is higher than in random networks Watts and Strogatz (1998).

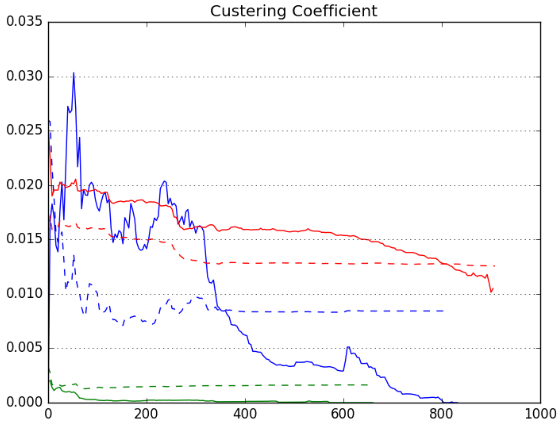

Figure 7. Trends of the clustering coefficient for the CRG (dashed lines) DRG (solid lines) for the three datasets: Black Friday in blue; World Series in green and Italian Firehose in Red

In Figure 7, we show the clustering coefficient evolution in the three datasets in the case of CRG and DRG. For World Series we have very low clustering coefficient for both the CRG that for the DRG. In the other two cases we get values slightly higher in the case of CRG which tend to decrease considerably as time goes. However these values are an order of magnitude lower than the ones observed in social networks, see Myers et al. (2014).

4.3 In-degree and out-degree Distributions

The plot in Figure 8 shows the distribution of the in-degrees for a particular DRG G_t. In the *y*-axis are represented the number of vertices with in-degree that corresponds to the value in the *x*-axis. The time-stamp *t* is chosen so that it is close to the event (in the case of event-driven graph) or in the middle of temporal observation window (in the case of Italian Firehose). The plot is in logarithmic scale in both the axes.

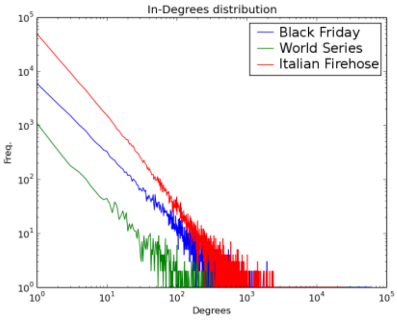

Figure 8. The in-degrees distributions of three particular DRG, one for dataset

By observing Figure 8 we deduce that the in-degrees distributions of the three graphs follow a power-law distribution. We observe the same behavior also with other timestamps or considering the out-degrees distribution.

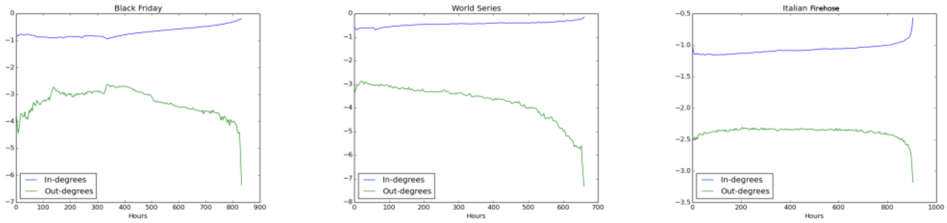

Figure 9. Power-law exponents of the in-degrees distribution (blue) and out-degrees distribution (green) of Italian Firehose, World Series, and Black Friday DRG

Starting from these consideration we have derived the trend of the power-law exponents of the in- and out- degrees distribution. The results are shown in Figure 9. It is important to note that, except in the beginning an the end of the observation periods that suffer for the border effect, the power-low exponents are substantially constant over time.

4.4 Other Properties

In the three datasets and for both CRG and DRG, the number of connected components evolves in the same way as the number of vertices, so much that from Figure 5 these two measures appear to be strongly correlated. On the other hand, the trends of the sizes of the biggest strongly connected components are shown in Figure 10.

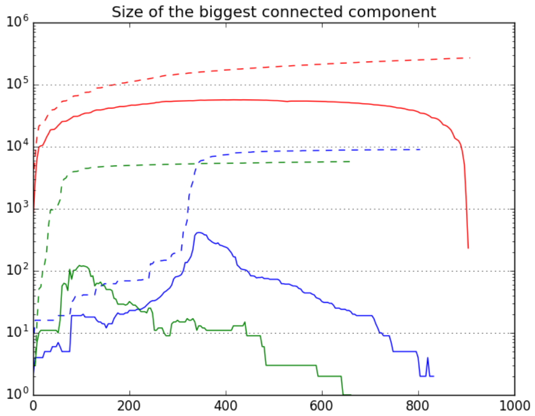

Figure 10.Trends of the size of the biggest strongly connected component for the CRG (dashed lines) DRG (solid lines) for the three datasets: Black Friday in blue; World Series in green and Italian Firehose in Red

Although there is not a strong correlation between the number of vertices and the size of strongly connected components (see Figure 5), the two event-driven DRG show a peak in correspondence of the event.

5. DISCUSSION AND CONCLUSIONS

Thanks to the Big Data technology, we have performed an extensive analysis of the evolution of retweet graphs relative to three Twitter streams for different periods of time. This is one of the first papers that systematically studies the temporal growth of graphs generated by a social network. We conducted the analysis on two types of graphs: the event-driven graphs and the graph constructed by an Italian stream of tweets (Italian Firehose). We considered two opposing models of evolution graph: the Cumulative Retweet Graphs (CRG) in which the vertices and edges corresponding to users and tweets once added, will never be deleted; the Dynamic Retweet Graphs (DRG) in which vertices and edges that correspond to inactive users and obsolete tweets are cutting off from the graph.

There are well known properties that real graphs derived from social networks satisfy, such as heavy tails for the in-degree and out-degree distribution, shrinking average distance and diameters, and the Densification Power Law (DPL), see Leskovec et al. (2005) and Leskovec et al. (2007). From our analysis it follows that also, the CRG, both of the whole Italian Firehose and the two event-based streams, satisfy such properties - see also Amati et al. (2016). Moreover, we have compared the behavior of the DRG with the CRG and we have seen that for the Italian Firehose, that contains the involution of many event-based subgraphs, the DRG dynamically maintains the same structural properties of the cumulative graph. Interestingly, the clustering coefficient and the average distance are very close, witnessing thus that the Italian Firehose is indeed the outcome of the union of different communities. Conversely, single event-based graphs show the border effects on all structural measures, generally growing and decaying in a similar manner, and reaching a peek activity around the middle of their lifetime much below the values of their corresponding cumulative graphs. Moreover, all real properties shown for other cumulative graphs, with the exception of the DPL and Degree Power Laws, do not hold: average distance and the clustering coefficient converge super-linearly.

One important problem still remains open that is to rigorously define a mathematical model that describes the evolution of Twitter graphs according to topic, communities and type of events.

ACKNOWLEDGMENTS

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

REFERENCES

Amati G., Angelini S, Capri F., Gambosi G., Rossi G., and Vocca P., 2016. 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.

Amati G., Angelini S., Bianchi M., Fusco G., Gambosi G., Gaudino G., Marcone G., Rossi G., and Vocca P., 2015. Moving beyond the twitter follow graph. In Proceedings of the International Conference on Knowledge Discovery and Information Retrieval, part of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015).

Barrat, A. and Weigt, M., 2000. On the properties of smallworld networks. Eur. Phys. J. B, 13, 547–560.

Bhamidi S., Steele J.M., and Zaman T., 2015. Twitter event networks and the superstar model. The Annals of Applied Probability, 25(5), 2462-2502.

Bhattacharya D. and Ram S., 2012. Sharing News Articles Using 140 Characters: A Diffusion Analysis on Twitter. Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM).

Bild D.R., Liu Y., Dick R. P., Morley Mao Z., and Wallach D.S., 2015. Aggregate characterization of account behavior in Twitter and analysis of the retweet graph. ACM Trans. Internet Technol., 15(1), 4:1–4:24.

Java A., Song X., Finin T., and Tseng B., 2007. Why we twitter: Understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNAKDD 2007 Workshop on Web Mining and Social Network Analysis, WebKDD/SNA-KDD 2007.

Kwak H., Lee C., Park H., and Moon S., 2010. What is twitter, a social network or a news media?. In Proceedings of the 19th International Conference on World Wide Web, (WWW 2010).

Leskovec J., Chakrabarti D., Kleinberg J., and Faloutsos C., 2015. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases.

Leskovec J., Chakrabarti D., Kleinberg J., Faloutsos C., and Ghahramani J., 2010. Kronecker graphs: An approach to modeling networks. . J. Mach. Learn. Res., 11, 985-1042.

Leskovec J., Kleinberg J., and Faloutsos C., 2007. Graph evolution: Densification and shrinking diameters. ACM Trans. Knowl. Discov. Data, 1(1), .

Myers S.A., Sharma A., Gupta P., and Lin J., 2014. 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).

Ten-Thij M, Ouboter T, Worm D, Litvak N., van den Berg H., and Bhulai S., 2015, Modelling of Trends in Twitter Using Retweet Graph Dynamics, , Algorithms and Models for the Web Graph, Lecture Notes in Computer Science.

Ten-Thij M., Bhulai S., and Kampstra P. , 2014. Circadian patterns in twitter. Data Analytics, , 12-17.

Watts D.J. and Strogatz S. , 1998. Collective dynamics of 'small-world' networks. Nature, 393 (6684), 440-442.

Yang M., Lee J., Lee S., and Rim H., 2012. 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).

Zhou Z., Bandari R., Kong J., Qian H., and Roychowdhury V., 2010. Information resonance on Twitter: watching Iran. . In Proceedings of the First Workshop on Social Media Analytics.

Zubiaga A., Spina D., Martínez R., Fresno V., 2015. Real-time classification of Twitter trends. Journal of the Association for Information Science and Technology, 66, 426.