








The Analysis of Influential Users Evolution in Microblogging Social Networks (Extended Abstract)

Giambattista Amati¹ , Simone Angelini¹, Giorgio Gambosi² ,
Gianluca Rossi² , and Paola Vocca³  

¹ Fondazione Ugo Bordoni, Rome, Italy
{gba,sangelini}@fub.it

² University of Rome “Tor Vergata”, Rome, Italy
{giorgio.gambosi,gianluca.rossi}@uniroma2.it

³ University of Tuscia, Viterbo, Italy
vocca@unitus.it

Abstract. In this paper, we study the evolution of the most influential users in the microblogging social network platform Twitter. To this aim, we consider the *Dynamic Retweet Graph (DRG)* proposed in [3] and partially analyzed in [2, 4]. The model of the evolution of the Twitter social network is based on the retweet relationship. In a DRGs, 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 consider the following measures of centrality: *degree*, *closeness*, and *pagerank-centrality* which have been widely studied in the static case. Here we analyze them on the sequence of DRG temporal graphs with special regard to the distribution of the 75% most central nodes.

We derive the following results: (a) in all cases the closeness measure produces many nodes with high centrality, so it is useless to detect influential users; (b) for the other measures almost all nodes have null or very low centrality and (c) the number of vertices with significant centrality are often the same; (d) the above observations hold also for the whole DRG and, (e) central nodes in the sequence of DRG temporal graphs have high centrality in static graphs.

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

1 Introduction

One of the fundamental and most studied features in a social network is the detection of central nodes, which can usually be considered as the *most important* nodes [6, 7, 12]. Centrality is widely-used for measuring the relative importance

This work was conducted in the Laboratory of Big Data of ISCOM-MISE (Institute of communication of the Italian Ministry for Economic Development).

of nodes within a graph and has many applications: in social networks to determine the most influential or well-connected people; in the Web graph to rank pages in a search; in a terrorist network, to detect agents that are critical for facilitating the transmission of information; for the dissemination of information in P2P Networks, Decentralized Online Social Networks and Friend-to-Friend Network [10].

There is a plethora of centrality definitions: degree centrality [16], closeness centrality [5], graph centrality [13], stress centrality [17], betweenness centrality [11], each one of them useful to detect specific properties and with significantly different computational costs. Here we consider four of them: the *degree*, *closeness*, *betweenness*, and *PageRank*-centrality.

Degree centrality, i.e. the degree d_v of a vertex v , is the simplest measure of centrality: it just takes into account how many direct, “one hop” connections each node has to other nodes of the network, hence it can be applied to detect popular individuals, agents who are likely to hold most information or individuals who can quickly connect with the wider network. The degree centrality is very cheap to compute but, being a purely local notion, it is often unable to recognize the relevance of certain nodes.

One of the most popular measures, but computational expensive for large graphs, is betweenness-centrality. It detects nodes which act as “bridges” between other nodes in a network. It does this by identifying all the shortest paths and then counting how many times each node falls on one. Betweenness centrality is suitable for finding vertices who influence flows (such as information flow) in the network.

A third measure considered below is closeness-centrality, which, after computing the set of all-pairs shortest paths, assigns each node a score based on the number of shortest paths to which it belongs. This definition of centrality is useful for quickly finding the agents who are in good position to influence the entire network but in a highly connected network often most nodes have a similar score.

Finally, Pagerank-centrality was introduced in [8] and it recursively quantifies a “value” or the PageRank of a node based on: (i) the number of links it receives, (ii) the link propensity of the linkers (that is, the number of outgoing links of each in-going node), and (iii) the centrality of the linkers, that is their PageRank.

In order to analyze the evolution the influential users we study the distribution of the centrality measures on a model of the Twitter network, the *Dynamic Retweet Graph (DRG)* proposed in [3] and partially analyzed in [2, 4].

This model has two major features: (i) we consider the retweet graph since it allows to better represent relationships among users related to information flow in Twitter [14, 15] and (ii) once a tweet has been retweeted for the last time all the edges representing that tweet are deleted, to model the decay of relevance of the tweet content.

The temporal model we consider coincides with the other ones in the growing phase. That is a new vertex is added whenever a new user starts or retweets a tweet, and a new directed edge (a, b) is inserted when a user a retweets for

the first time a tweet of b , if an edge already exists then a timestamp is added to it. Conversely, the decreasing stage happens when a tweet is never retweeted again. Then, all vertices and edges not involved in other retweeting processes are deleted at once. As shown in previous experimentations [2,4], this evolutionary model better captures the information flow in Twitter.

DRGs seem to better represent the double nature of the Twitter platform: social network and news media [14,15].

For what concerns the use of the centrality measure to assess influential or authoritative users Kwak et al. [14] compared three measures of influence: in-degree centrality, PageRank centrality in the following/follower network and the number of retweets on Twitter. Cha et al. [9] compared three different measures of influence: in-degree centrality, the number of retweets and mentions on Twitter. The results indicate that users with high in-degree were not necessarily influential.

In this paper we study the evolution of the most influential users in the microblogging social network platform Twitter with respect to the above four centrality measures (betweenness, degree, closeness, and PageRank) and we analyze their behavior on the DRG evolutionary model of the retweet social networks proposed in [3].

We consider two different kind of data sets, first introduced in [1] and updated and refined in [3]: the *event driven* retweet graphs based on the events *Black Friday 2015* and the *World Series 2015* and the *Italian Sampling*, that is the *firehose* retweet graph, filtered by language (i.e. Italian) from the whole Twitter stream.

The four centrality measures are analyzed on three levels: (i) with respect to the sequence of DRG temporal graphs; (ii) with respect to the static cumulative graph, that is the graph that contains all nodes and edges and (iii) with respect to the kind of networks considered, that is *event driven* or the firehose.

We derive that the model proposed allows to detect the most authoritative users, since:

1. in all cases the closeness centrality provides too many central nodes, hence it is useless to detect influential users;
2. with regard the other measures, almost all nodes have null or very low centrality;
3. vertices with centrality values above 75% of the maximum is a small set and they are often repeated in the three centrality measures;
4. the above observations hold also for the static graphs (the whole DRG);
5. central nodes in the sequence of DRG temporal graphs have high centrality in static graphs.

2 DRG Temporal Graphs

In this paper we will use a definition of Dynamic Retweet Graph (DRG) slightly different from the one in [4].

A DRG graph $G = (V, E, \ell)$ is defined as follows: the set V of nodes are Twitter accounts and a directed edge $e \in E$ represents an interaction (a retweet) between two accounts. In particular, there is a directed edge from an account a to 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 . This edge information is implemented with a list $\ell(e)$ 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 $\ell(e)$ are sorted by non-decreasing order of their timestamp.

From the data that we have collected in G we define, for all tweets i , the *date of death* of i (in short, $\text{dod}(i)$) as the timestamp of the latest retweet of i . Formally,

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

Consequently we define the *expiration date* of an edge e (in short, $\text{ed}(e)$) as the time after which all tweets associated to e will be dead. Formally,

$$\text{ed}(e) = \max\{\text{dod}(i) : (i, t) \in \ell(e)\}.$$

On the contrary, the *creation date* of an edge $e = (a, b)$ (in short, $\text{cd}(e)$) is the time when b retweets a for the first time, formally:

$$\text{cd}(e) = \min\{t : (i, t) \in \ell(e)\}.$$

Let t be a timestamp, we define a *DRG temporal graph* at time t as the subgraph $G_t = (V_t, E_t)$ of the DRG G at time t defined as follows: E_t contains any edge e such that $\text{cd}(e) \leq t \leq \text{ed}(e)$; V_t is the set of nodes induced by E_t .

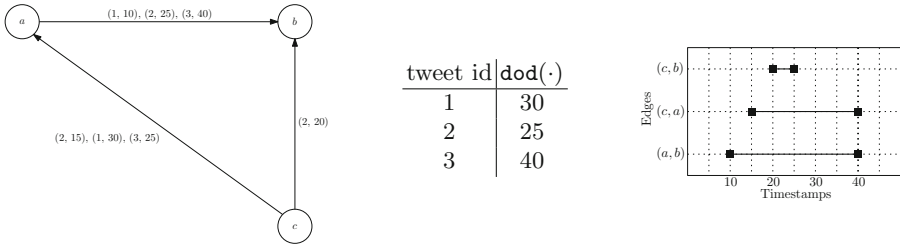


Fig. 1. On the left side, an example of a DRG retweet graph. Edges are labelled by pairs with the id of the tweet and the timestamp of the retweet. The center table shows the date of death of all tweets in the graph. On the right side, for each edge of G is represented its creation and expiration date.

For example if G is the retweet graph represented in the left part of Fig. 1, G_{30} contains edges (a, b) and (c, a) and the induced vertices since (c, b) expires at timestamp 25. For all $20 \leq t \leq 25$, G_t contains all edges of G .

3 Data Sets

For the experiments we use the same dataset as [3] 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 sampling retweet graph, filtered by the Italian language from the whole Twitter stream. To obtain the Italian Twitter sample we use a list of the most used Italian stop words and the Twitter native selection function for languages. In Table 1 the size of the three graphs are shown. In Fig. 2 we show the evolution of the size of the three datasets over the period of observation. Note that the event-driven datasets (World Series and Black Friday) show a rapid growth close to the events, and then a slow decline. Differently, the Italian Sampling show a smooth and stable behavior, ignoring the border effects.

Table 1. Size of the dataset

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

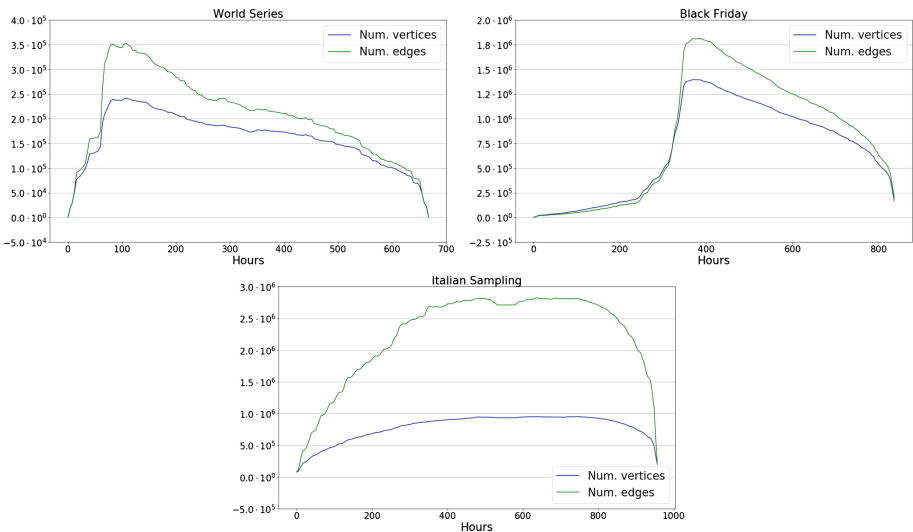


Fig. 2. Number of vertices (blue) and number of edges (green) of: World Series, Black Friday, and Italian Sampling, as functions of hours. (Color figure online)

4 Experimentation

For each graph G in our dataset, we consider the sequence of DRG temporal graphs $(G_{t_i})_{i \geq 0}$ where $t_{i+1} - t_i$ is 4 h. For each G_t we compute the four centrality values (betweenness, closeness, degree, and PageRank centrality) of each vertex of the graph.

Given the centrality measure c , the *relative centrality value* with respect to c of a vertex u is the ratio $c(u)$ and the maximum value of $c(\cdot)$.

Preliminary considerations. First of all, for each centrality measure $c(\cdot)$ and for each G_t , we consider the number of nodes with centrality values above 90% of the maximum. Figure 3(a) shows the behavior of the closeness centrality: observe that this value is almost always greater than 30%. This means that

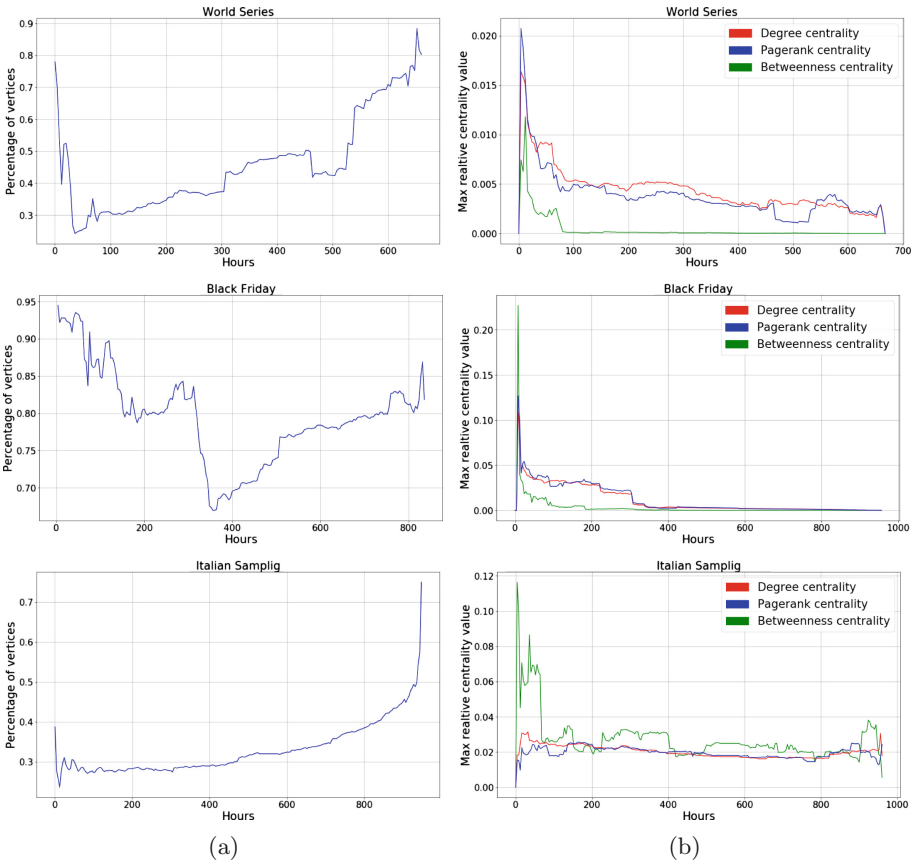


Fig. 3. (a) Trend over time of the ratio of nodes whose closeness centrality is above the 90% of the maximum. (b) The 99.9-th percentile evolution over time of the three relative centrality measures.

closeness centrality is not very suitable to determine the more influential nodes in the graph. Conversely, the other centrality measures (degree, betweenness, and PageRank) show an opposite behavior: excluding the first and last timestamp, 99.9% of vertices always have centrality values below 20% of the maximum. This is shown in Fig. 3(b) which shows the evolution over time of the three centrality values below which the 99.9% of all values fall (99.9-th percentile). Observe that, from Fig. 3(b) it results that the highest values are at the very beginning of time sequences, when there is still much instability. After that, values fall below 0.05.

Analysis of temporal graphs. From the previous observations it follows that if we restrict ourselves to the betweenness, degree and PageRank measures, the number of vertices for which the centrality value is meaningful is so small that we can study them one by one.

We say that a node is *central* (with respect to a centrality measure) if its centrality value is at least 75% of the maximum. Let G be a DRG, c be a centrality and t be a timestamp, we define $A_{G,c,t}$ as the set of central node of G_t with respect to c .

In Fig. 4 are shown the sets $A_{G,c,t}$ for the World Series (Italian Sampling and Black Friday are similar and are omitted for lack of space). In the y -axis are reported the vertex ids. Let us consider one of the diagrams in the figure relative

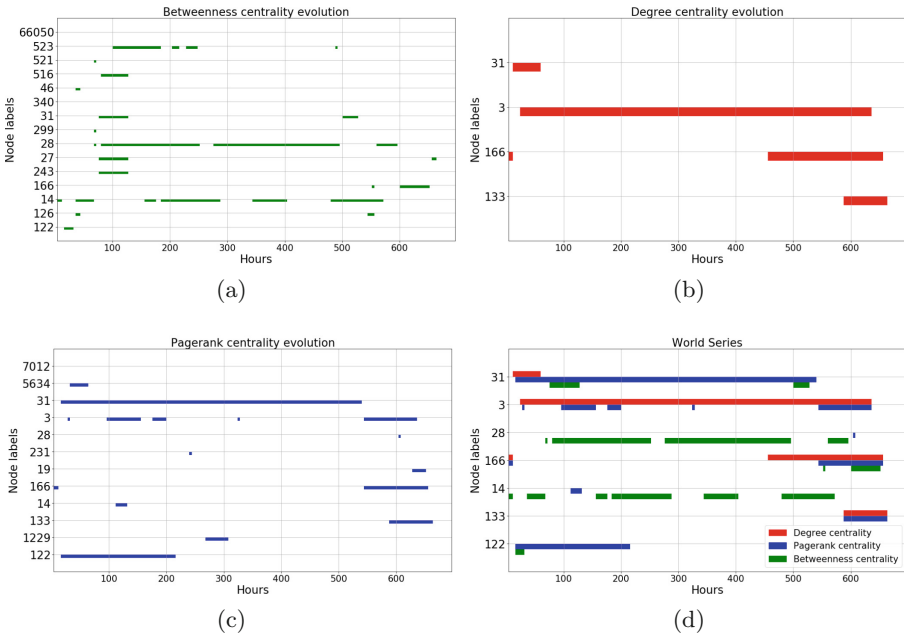


Fig. 4. Temporal evolution of $A_{G,c,t}$ for the world Series relative to centrality measures: betweenness (a), degree (b), and PageRank (c). Diagram (d) shows the overlapping of (a), (b), and (c) with respect to at least two measures.

to measure c : a segment in correspondence to node u that intersects timestamp t means that $u \in A_{G,c,t}$. From the above analysis we get the following observations:

- For all datasets, the degree centrality always produces a total number of central nodes lower than the other measures. Conversely, betweenness centrality is the one that produces more.
- For all datasets and all the centrality measures, there are nodes that are central for a long time: this trend is more prominent for degree and pagerank centrality.
- Another important result that turns out is a significant overlap between the central vertices with respect to the three measures. In Fig. 4(d) the diagrams in Fig. 4(a)–(c) are merged by taking into account only vertices in common with at least two measures. For example vertex 31 is central for most of the time over the three measures.

Comparison with the static cumulative DRGs. The latest analysis involves the centrality measures of the static cumulative DRGs G representing the three datasets. Like DRGs temporal graphs, a large portion of vertices, varying from 28% (for World Series) to 50% (for Black Friday), have closeness centrality above 90% of the maximum, hence, we discard it.

On the contrary for the betweenness, degree, and PageRank centrality, almost all the nodes have centrality below 1% of the maximum. Table 2 shows, for each dataset and for each measure the percentage of vertices whose relative centrality value is at most 0.01.

Table 2. Percentage of vertices whose relative centrality value is at most 0.01.

World Series			Black Friday			Italian Sampling		
Between.	Degree	Pagerank	Between.	Degree	Pagerank	Between.	Degree	Pagerank
99.934%	99.948%	99.932%	99.97%	99.96%	99.97%	99.93%	99.78%	99.84%

Now we will focus on vertices with high relative centrality. Table 3 lists the id of vertices of the World Series dataset whose relative betweenness centrality is at least 0.5. Some of these nodes compare also in Fig. 4(a). That is, there are nodes that are central in both the static cumulative DRG and in the temporal graphs. Table 4 lists, for all central nodes with respect the betweenness centrality in the World Series temporal graphs (see Fig. 4(a)), the relative centrality in the static cumulative DRG. It is interesting to note that all the listed nodes but one (node 166) belong to the 0.066% (=100 – 99.934, see Table 2) of vertices whose relative centrality is at least 0.01. That is almost all nodes that are central in temporal graphs are also central in the whole graph.

Table 3. Vertices of the whole World Series dataset whose relative betweenness centrality is at least 0.5.

Vertex id	Relative centrality	Vertex id	Relative centrality
299	1.00	122	0.62
31	0.69	11374	0.52
27	0.67		

Table 4. Relative centrality in the whole World Series dataset of nodes that are central in the temporal graphs.

Vertex id	Relative centrality	Vertex id	Relative centrality
299	1.00	243	0.19
31	0.69	340	0.18
27	0.67	126	0.10
122	0.62	516	0.07
14	0.49	521	0.05
46	0.25	66050	0.03
28	0.23	166	< 0.01
523	0.20		

For what concerns the other centrality measures and the Italian Sampling and Black Friday datasets we have similar results that are not reported in this extended abstract for lack of space.

5 Discussion and Conclusions

In this paper we have studied the evolution of four centrality measures (betweenness, degree, closeness, and PageRank) on the DRG temporal retweet graphs based on three datasets: Black Friday, World Series, and Italian Sampling. Our main results can be summarized as follows: (i) too many nodes are central with respect closeness centrality, hence this measure is useless to detect influential users; (ii) for the other measures, the number of nodes with very low centrality is very high and the sets of central nodes (with centrality values above 75% of the maximum) are very small and quite similar in the three measures; (iii) similar results hold also for the static cumulative graphs where the sets of nodes with relevant centrality contain central nodes in the sequence of DRG temporal graphs.

As pointed out in [4], the DRG temporal graphs derived from our datasets are quite sparse: this could explain the small number of central nodes respect to the three centrality measures.

According to the above analysis the approach based on the DRG temporal graph and the centrality measures represent a promising approach for detecting influencer in the microblogging Twitter platform.

References

1. Amati, G., Angelini, S., Bianchi, M., Costantini, L., Marcone, G.: A scalable approach to near real-time sentiment analysis on social networks. In: DART 2014 Information Filtering and Retrieval. Proceedings of the 8th International Workshop on Information Filtering and Retrieval co-located with XIII AI*IA Symposium on Artificial Intelligence (AI*IA 2014), vol. 1314, pp. 12–23. CEUR-WS.org, December 2014
2. Amati, G., Angelini, S., Capri, F., Gambosi, G., Rossi, G., Vocca, P.: Modelling the temporal evolution of the retweet graph. *IADIS Int. J. Comput. Sci. Inf. Syst.* **11**(2), 19–30 (2016)
3. Amati, G., Angelini, S., Capri, F., Gambosi, G., Rossi, G., Vocca, P.: Twitter temporal evolution analysis: comparing event and topic driven retweet graphs. In: BIGDACI 2016 - Proceedings of the International Conference on Big Data Analytics, Data Mining and Computational Intelligence, Funchal, Madeira, Portugal, July 2–4, 2016, vol. 1 (2016)
4. Amati, G., Angelini, S., Capri, F., Gambosi, G., Rossi, G., Vocca, P.: On the retweet decay of the evolutionary retweet graph. In: Gaggi, O., Manzoni, P., Palazzi, C., Bujari, A., Marquez-Barja, J.M. (eds.) GOODTECHS 2016. LNICSITE, vol. 195, pp. 243–253. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61949-1_26
5. Bavelas, A.: Communication patterns in task-oriented groups. *J. Acoust. Soc. Am.* **22**(6), 725–730 (1950)
6. Bonacich, P.: Power and centrality: a family of measures. *Am. J. Sociol.* **92**(5), 1170–1182 (1987)
7. Borgatti, S.P.: Centrality and network flow. *Soc. Netw.* **27**(1), 55–71 (2005)
8. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **30**(1–7), 107–117 (1998)
9. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P.K.: Measuring user influence in twitter: the million follower fallacy. *ICWSM* **10**(10–17), 30 (2010)
10. Conti, M., De Salve, A., Guidi, B., Ricci, L.: Epidemic diffusion of social updates in Dunbar-based DOSN. In: Lopes, L., Žilinskas, J., Costan, A., Cascella, R.G., Kecskemeti, G., Jeannot, E., Cannataro, M., Ricci, L., Benkner, S., Petit, S., Scarano, V., Gracia, J., Hunold, S., Scott, S.L., Lankes, S., Lengauer, C., Carretero, J., Breitbart, J., Alexander, M. (eds.) Euro-Par 2014. LNCS, vol. 8805, pp. 311–322. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-14325-5_27
11. Freeman, L.C.: A set of measures of centrality based on betweenness. *Sociometry* **40**, 35–41 (1977)
12. Freeman, L.C.: Centrality in social networks conceptual clarification. *Soc. Netw.* **1**(3), 215–239 (1978)
13. Hage, P., Harary, F.: Eccentricity and centrality in networks. *Soc. Netw.* **17**(1), 57–63 (1995)
14. 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, New York, NY, USA, pp. 591–600. ACM (2010)

15. 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 23rd International Conference on World Wide Web, WWW 2014 Companion, New York, NY, USA, pp. 493–498. ACM (2014)
16. Nieminen, J.: On centrality in a graph. *Scand. J. Psychol.* **15**, 322–336 (1974)
17. Shimbel, A.: Structural parameters of communication networks. *Bull. Math. Biophys.* **15**(4), 501–507 (1953)