



Utilizing the average node degree to assess the temporal growth rate of Twitter

Despoina Antonakaki¹ · Sotiris Ioannidis¹ · Paraskevi Fragopoulou¹

Received: 11 November 2016 / Revised: 4 February 2018 / Accepted: 7 February 2018 / Published online: 16 February 2018
© Springer-Verlag GmbH Austria, part of Springer Nature 2018

Abstract

Several models have been proposed that describe the evolution of the graph properties of many online social networks (OSNs) and explain the behavior of their users. These models are essential for understanding the growth dynamics of the underlying social graph. One of the most prominent OSNs is Twitter, since it covers a significant part of the online worldwide population. Nevertheless, investigating the validity of these models on Twitter entails many difficulties. The size of Twitter and the limitations of its access API make extremely difficult the estimation of many graph properties and therefore the evaluation of the proposed models. In this study, we present a simple and efficient method to fit an already existing model, which describes the *densification power law* property of modern OSNs. This model states that the average degree of an OSN increases over time. In a case study, we assess this model in two large samples of Twitter, and we demonstrate how it can portray the altering growth periods of Twitter. Finally, we make some remarks on several events during the early period of Twitter that may have affected its growth rates.

Keywords Twitter · Evolution · Average node degree · Temporal growth rate · Online social networks · Densification power law

1 Introduction

Today, a large proportion of online activity is happening through online social networks (OSNs). The study of this activity provides valuable insight regarding the growth and dynamics of an OSN. One of the major research objects in this area is the graph that represents the social network. In this graph, nodes represent individual users (or accounts), and edges are friendship (or following) relationships between them.

The study of the underlying graph of a social network can give valuable insight, regarding the dynamics that govern the creation, and evolution of the network (Strogatz 2001). One of the most common strategies in this area is to model the

node degree distribution of the network. This distribution is denoted as the function $P(d)$ of the percentage of nodes with degree equal to d . By studying this distribution, we can often make accurate assumptions regarding the rationale followed by users when making new connections. For example, a $P(d)$, that follows a Poisson distribution, indicates a random graph. The Gaussian distribution indicates a single-scale network, where various limiting factors (i.e., aging) constrain nodes from becoming “very rich”, and emerges, usually, in acquaintance networks (Amaral et al. 2000). Finally, a power-law distribution ($P(d) \sim d^{-\lambda}$) indicates a scale-free network. This last class of networks arise in various cases, for example, when the probability of a node receiving new connections depends on the number of connections, it already has. This also is known as the “preferential attachment” model (Barabási 1999).

The degree distribution of Twitter has been extensively studied (Sadikov and Martinez 2009; Kwak et al. 2010; Myers et al. 2014). A study of 2009, that examined the topology of 54.3 million users (Sadikov and Martinez 2009), found that both the outgoing and incoming degree follow a power law, with exponents 1.95 and 2.13, respectively. Nevertheless, a study of 2010 with 41.7 million users (Kwak

✉ Despoina Antonakaki
despoina@ics.forth.gr

Sotiris Ioannidis
sotiris@ics.forth.gr

Paraskevi Fragopoulou
fragopou@ics.forth.gr

¹ FORTH-ICS, Heraklion, Crete, Greece

et al. 2010) concluded that Twitter deviates from other social networks and that the outgoing degree distribution is not a power law. The most recent study (Myers et al. 2014) with the largest sample size (175 million users) demonstrates that the indegree is best fitted by a power law with $\lambda = 1.35$, whereas the outdegree is best fitted by a log-normal distribution with $\mu = 3.56$ and $\sigma^2 = 2.87$. Given the plethora of contradicting findings, we can conclude that the elucidation of Twitter's degree distribution is an active research question. However, all studies agree that the outgoing degree distribution follows *partly* a power law for users, with less than $\sim 10^5$ followers.

Instead of trying to fit the degree distribution to a known function, another line of work tries to include *time* as a factor and locates models that describe the evolution or else the temporal growth of the network. The main design principle of a mathematical model that captures the evolution of modern OSNs is to be able to formally describe the behavior of users in a way that the structure and properties of the network can be accurately predicted (Kumar et al. 2006). For example, studies have modeled the 'rich get richer' property (Barabási 1999), the 'small world phenomenon' (Kleinberg 2000) and the decreasing diameter observation (Leskovec et al. 2005). Areas like graph sampling (Leskovec and Faloutsos 2006), graph generation (Leskovec et al. 2008b) and spam detection (Benevenuto et al. 2010) rely heavily on models that describe accurately the growth of a network. A model can also be useful for predicting the actions of individual users or for identifying events in time that caused structural changes in the graph (Chan et al. 2012). For example, Barbieri et al. (2014) and Bliss et al. (2013) suggested following recommendation systems based on evolution models of social networks. Similar work has been done in the area of community detection (Barbieri et al. 2013), measurement of users' influence (Morales et al. 2014; Bray 2015), and the study of temporal variation of hashtag popularity (Yang and Leskovec 2011).

A well-studied family of models are the "preferential attachment", and the "copying model" (Kleinberg et al. 1999). These models can generate scale-free networks, where the average degree of the network remains constant, and its effective diameter slowly grows. Leskovec et al. (2007) noticed that these assumptions do not apply in many modern social networks. In contrast, they suggested that, as the number of nodes increases, the average degree also increases (the graph becomes more dense), whereas the diameter decreases (the graph shrinks). These significant differences may originate from the possibility that the outdegree distribution of these networks is not always a power law. Below on this paper, we expand on the specifics of this model, to which we refer as the "Leskovec model".

The "Leskovec model" has been extended in (Kleinberg and Boguñá 2014) to incorporate the layer of the existing yet

unobserved off-line social network. Since the latter model is more appropriate for local social networks (for example, nation-wide OSNs), we believe that the Leskovec model is more suitable for the study of Twitter. There is an extensive discussion on whether Facebook also follows this model, but without a decisive answer (Backstrom et al. 2012). Additionally, more elaborate models have been proposed that take into account better graph metrics, and suggest more thorough measurement techniques (Wei and Carley 2015). Yet, efforts to validate these models are limited to graphs, in the order of 10^5 nodes at best, which is incomparable to the size of modern social networks, like Twitter.

Although the "Leskovec model" has a sound mathematical base, and it has been applied for the study of the evolution of other social networks (Backstrom et al. 2012; Leskovec et al. 2008a), it has not been studied on Twitter. There are three reasons for this: The first is that acquiring a sufficient sample size for Twitter (estimated on this paper as ~ 90 million nodes) is extremely difficult, given the current limitations of Twitter's API. The second is that Twitter's API does not reveal the creation time of the links (following relationships); hence, moving back in time is not easy. Finally, if $|V|$ is the number of nodes, and $|E|$ the number of edges in the graph, the computational complexity of deriving the diameter is in the order of $O(|V||E|)$, which, in the case of Twitter, can reach the prohibitive amount of 10^{20} calculations for a sufficient sample size.

In this paper, we present a case study, where we apply the "Leskovec model" on the average outdegree of Twitter. We overcome the first two difficulties, by acquiring two large samples of Twitter, and by applying an approximation method to infer the link creation time. Here, we do not focus on the diameter, due to the third difficulty, although we agree that measurements of this property can further assess the validity of the "Leskovec model", and estimate the "shrinking" observation of the social graph. We add this task on our future work.

The estimation of the average degree has lower computational complexity ($O(|V|)$), whereas it can sufficiently portray the 'densification law' described in the "Leskovec model." We subsequently demonstrate how this modeling can delineate periods of diverse growth rates, that Twitter underwent, especially on its early days (before 2010). Also, the average degree has a significant meaning in the modeling of users' behavior, since it has been associated with the Dunbar's Number theory, which states that humans can have a finite number of stable social interactions in the range of 100–200 (Gonçalves et al. 2011). Finally, since OSNs are constantly growing, historic data are increasingly difficult and expensive to be collected (Batrincea and Treleaven 2015); the methods that contribute to the study of social networks' "archeology" can be of extreme importance.

The major contributions of this paper are:

- We present a fast, efficient and practical method to fit a widely accepted model that describes the evolution of the average node degree for large OSNs.
- We fit this model on one of the largest samples of Twitter's OSN and show how the growth of the average degree fluctuates over time.
- Based on the extracted growth pattern, we pinpoint several turning point events that took place on the early period of Twitter.

The rest of the paper is organized as follows: Sect. 2 presents our datasets and our collection methods. Section 3 applies and evaluates a known heuristic to build a snapshot of Twitter dataset, for a given period. Section 4 describes how Leskovec's model can accurately describe the growth of Twitter, with insufficient data. Section 5 presents the various growth periods of Twitter and some events that may have affected it. We conclude and discuss our findings, limitations, and future work in Sect. 6.

1.1 Terminology

In directed graphs, like social networks, the degree of a node is the sum of its outdegree and its indegree. The outdegree is the number of edges with direction outward to the node, whereas indegree is the number of inward directed edges. The average node outdegree is the average outdegree over all nodes. The density Q of a network is defined as the ratio of the number of edges E to the maximum possible number of edges and is defined as $(2E)/(N(N-1))$, where N is the number of nodes.

On Twitter, any user can “follow” any other user with a public profile. So given a specific user, A , we call the set of users that follow A as “followers”, and the set of users that are followed by A , as “friends”. We also use the term “following(s)”, to refer to any link on the social graph, regardless its direction.

2 Data collection

We used two independent datasets for our analysis. The first dataset consists of all the followers and friends of 92 million users. We used the random walk network sampling algorithm to obtain this dataset, which according to Leskovec and Faloutsos (2006) is the best method for capturing temporal graph patterns. Briefly, this algorithm simulates a random walk on the graph. Initially, we select a random user of Twitter, and we extract all her friends and followers. Then, we randomly select one of the newly added nodes and we repeat this procedure. At every step, we return at the starting point, with a probability of $p = 0.15$, and begin a new walk. For time-efficiency reasons, we initiated 11

random walks that were running concurrently. Each of these 11 “walks” had a different starting seed. We selected these seeds by randomly selecting 11 users, each one residing on a different geographic location (according to its latest tweet). These locations were: Canada, USA, Mexico, Argentina, UK, Greece, South Africa, Russia, Indonesia, Japan and Australia. Whenever a new sample (friend or follower of a node) was collected, we stored it in a Mongo database.

The same study that suggests random walk as an efficient sampling technique, (Leskovec and Faloutsos 2006), also addresses the issue of sufficient sample size for capturing graph metrics. A 15% sampling size is enough for measuring the graph properties of a graph as it grows and evolves. It is estimated that when the sampling happened (from September 2015 to April 2016), Twitter had 500–600 million users, who half of them were active users. Therefore, we reckon that 92 million users are a sufficient sampling size, since it constitutes 15% to 18% of the complete network. The average number of followers per user was 624, and the average number of friends was 763. For the remaining of this work, we will refer to this as the *BIG* dataset.

The second dataset contains the followers and friends of all users that are present on the study of Kwak et al. (2010). This dataset contains the entire graph of Twitter as of July 2009 and contains 40.8 million users. In the period from November 2014 to January 2015, we downloaded all the followers and friends of these users. Each user on this dataset has on average 210 friends and 214 followers. We will refer to this dataset as the *KWAK* dataset.

KWAK represents a sample of the early stage of Twitter, whereas *BIG* does not focus on a specific period. This will allow us to focus on some interesting events that took place during the early period of Twitter.

In Figs. 1 and 2, we show the evolution of the average outdegree and the density, respectively, for the two datasets. From these plots, we observe a first validation of the ‘densification law’ of the Leskovec model, which states that the average degree increases over time.

3 Generating time snapshots

Although Twitter does not reveal the creation time of followings, we can apply a heuristic that produces a lower bound estimation (Meeder et al. 2011). This heuristic is based on the fact that Twitter's API returns the lists of followers and friends of a user ordered according to the link creation time. This list contains the unique IDs of friends and followers. These IDs are increased monotonically. Consequently, the order of these IDs also reveals the subscription order of these users: Between any two users, the one that has the lower ID, subscribed earlier in Twitter. Therefore, the link creation time of the following relationship, between users

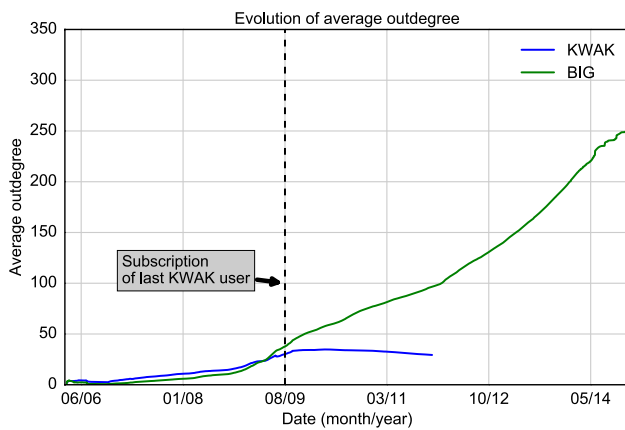


Fig. 1 The average outdegree of Twitter’s social network is increased over time. This is agrees with the “Leskovec model” and is evident in both datasets (*BIG* and *KWAK*). We also notice that the average outdegree of *KWAK* peaks and drops after August 2009. Since all users in *KWAK* have subscribed in Twitter before that date, a fair proportion of them were inactive when the sampling happened (5–6 years later). Inactive users do not add new followers; therefore, the measurement of the average outdegree past that date (August 2009) with the *KWAK* dataset is not representative of the real average outdegree value of Twitter. Nevertheless, this demonstrates that the influx of new users after 2009 compensated this effect and resulted in the increase of the average outdegree, as it is shown in the line representing the *BIG* dataset

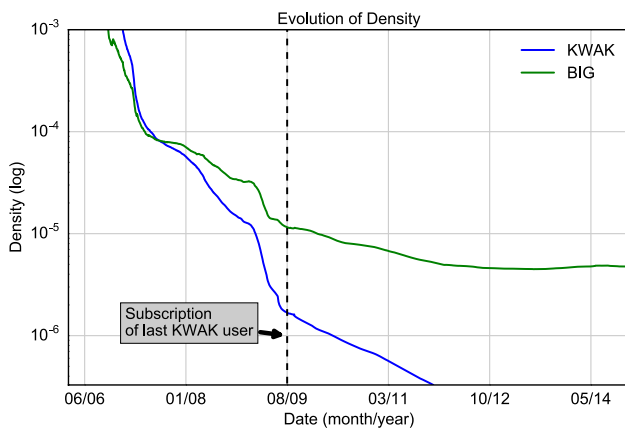


Fig. 2 The density of Twitter’s social network decreases over time. The density of a network at time t is defined as $Q_t = (2E_t)/(N_t(N_t - 1))$. N_t is the number of nodes, and E_t is the number of edges at time t . This plot demonstrates the ‘densification law’ of the Leskovec model

A and B , can be approximated by the most recent account creation time, among all users, that followed B prior to A . The computational complexity of this heuristic is $O(|E|)$.

The accuracy of the inferred link creation times of this heuristic depends on the number of friends and followers of a user. The higher the number of friends, or followers of a user, the more accurate this heuristic is. For celebrities

(users with more than 5000 followers), the link creation time is estimated with an accuracy level of several minutes. For users with lower number of followers or friends, the error can be higher reaching days or even weeks. Given the fact that the range of time this study covers spans over 9 years (2006–2015), we do not expect these inaccuracies to introduce significant errors in our analysis.

Another consideration is that the heuristic assumes that users’ IDs are ordered according to account creation time or else that users’ IDs are increased monotonically. In fact, that was actual the case when the heuristic was published. According to Twitter, this is not always the case. The strict monotonic order was guaranteed, until approximately 2011. After that, ID’s kept increasing, but the monotonic order is not guaranteed. To validate this, we plotted the Twitter IDs for 10 million random users ordered according to the creation time of their accounts. Figure 3 shows that although, the increase of Twitter IDs is not always monotonic after approximately 2013, it has a relatively canonical distribution. Moreover, since in the remaining of this paper, we focus on the early period of Twitter (before 2010), we do not expect this discrepancy to affect our findings.

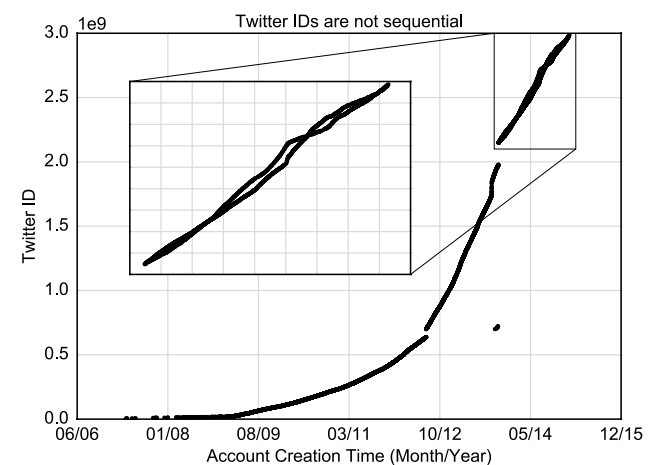


Fig. 3 Twitter provided unique user IDs, on 10 million random users. The x axis shows the subscription date of users, and y axis shows the IDs provided by Twitter. In general, we notice a monotonic increase of these IDs over time. Nevertheless, there are small deviations, where the increase is not monotonic. In the subplot, we notice that, after a period, there are two different ID sequences. Our model, that estimates the friendship creation date, assumes strong monotonic ID increases (user A with ID greater than user’s B ID is assumed to have subscribed later than B). Since we mainly focus on the early period of Twitter (before 2010), we do not expect these deviations to have any effect on our friendship date estimations

4 The average outdegree of Twitter

Two of the main properties that characterize the structure of OSNs are the average outdegree and the diameter of the graph. In Twitter, the outdegree of a node (or else, a user) is the number of other users (or else “friends”) that this user follows. The diameter is the longest shortest path in the graph, among all pairs of nodes. Usually, measuring the evolution of an OSN over time involves the study of the evolution of these parameters. It has been proposed (Broder et al. 2000; Albert et al. 1999) that the main “laws” that characterize the evolution of OSNs are: (1) constant average degree, and (2) slowly growing diameter.

In an influential study, Leskovec et al. (2007) suggested that both these laws are wrong and fail to describe the evolution of many modern social networks. In contrast, the authors proposed an alternative set of laws, based on empirical observations. These laws are: (1) Increasing average degree, and (2) decreasing diameter.

Leskovec et al. (2007) also suggested a model that describes the evolution of average degree. This model takes into account two parameters. The first is the community branching factor, b . If we model the graph as a tree of branching sub-communities, then b is the fanout of this tree. Fanout is the maximum number of children that a parent node might have in a tree. A large b is a characteristic of a dense network with tight communities. The second parameter is the Difficulty Constant, c and represents the difficulty to create a cross-community link in the graph.

The growth of the average outdegree of an OSN depends on the relation between these parameters. If the branching factor is higher than the difficulty factor, then the network’s outdegree increases superlinearly. If these parameters are equal, it increases logarithmically, and if the difficulty parameter is greater than the branching parameter, the network has a constant average outdegree through time. As defined by Leskovec et al. (2007), the expected average outdegree of a network (\bar{d}) is proportional to:

$$\begin{aligned} \bar{d} &= n^{1-\log_b(c)} && \text{if } 1 \leq c < b \\ &= \log_b(N_t) && \text{if } c = b \\ &= \text{constant} && \text{if } c > b \end{aligned}$$

while nodes (N_t) increase through time (t).

In this formula, b is the community branching factor and c is the difficulty constant. In case of superlinear growth (when $c < b$) the exponent $g = 1 - \log_b(c)$ is a quantification of the growth of the network. We will refer to value g as the “growth exponent.”

4.1 Fitting the model to Twitter

In both our datasets, we performed incremental measurements of the average outdegree, for every day of the dataset. Then, we fitted the Leskovec model, to a “sliding window” of the average outdegree. The size of the window was 200 days, and it moved from the first 200 days of our dataset to the last 200 days, with a timestep of 1 day. At each step, we fit all three functions of the Leskovec model to the current window, with the Levenberg–Marquardt algorithm (Marquardt 1963). Then, we assigned the midpoint of the window to the model with the lowest root mean square error. The Levenberg–Marquardt algorithm also produced estimations for the b and c parameters.

Figure 4 shows the average outdegree of the graph, while the nodes are increasing in the *BIG* dataset. Red parts are following a superlinear growth, whereas yellow parts follow a logarithmic scale and this pattern agrees with the Leskovec model. For parts with superlinear growth, we also have plotted the “growth exponent”, g . As predicted by Leskovec et al. (2007), Twitter’s social graph indeed does not have a constant outdegree distribution. In contrast, in most of the graph it exhibits a superlinear growth. The growth factor during the superlinear growth, oscillates drastically, reaches a maximum, and then, it is reduced. In the subsequent section, we will present some events that may have caused these variations.

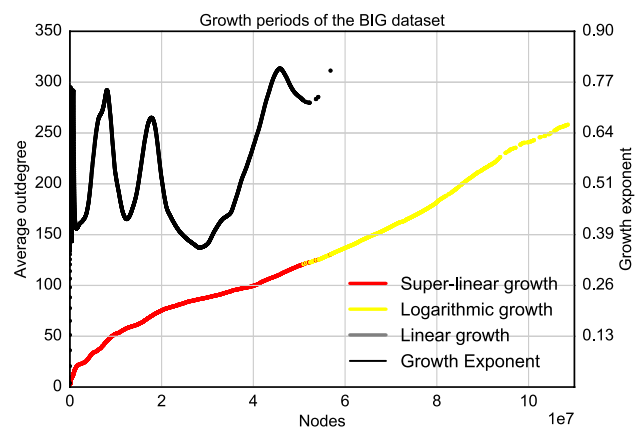


Fig. 4 Here we plot the evolution of the average outdegree for the *BIG* dataset, according to the nodes. X axis shows the number of nodes at a given time, and y axis shows the growth exponent of the graph (black line) and the average outdegree (colored line). The average outdegree is colored according to which function (superlinear, logarithmic or linear) estimates it better based on the Levenberg–Marquardt algorithm. Here, we notice that the average outdegree initially increased in a superlinear rate, and after 50 million nodes it slowed its rate into a logarithmic growth. Also, the growth exponent of the superlinear phase shows large variations that are indicative of various events that altered Twitter’s evolution in its early period (color figure online)

In Fig. 5 we present the same plot for the KWAK dataset, KWAK dataset, contains users that have registered to Twitter anytime before the end of 2009. Consequently, we cannot make any estimations of the overall outdegree distribution of the network, for any day past the end of 2009, based on these users. This is due to the fact that followings in the network are happening at a higher rate from recently subscribed users, compared to older ones. This is why we notice that the outdegree distribution drops after the end of 2009, which does not reflect a real tendency for the network. Similarly, we cannot produce reliable estimations of the growth exponent, when the window goes over the end of 2009. Nevertheless, we have a better resolution of the growth rate and the average outdegree of the graph, for the period marked from the start of Twitter until the end of 2009. In Figs. 6 and 7, we show the temporal evolution of the average outdegree for both datasets, respectively.

Finally, in Fig. 8 we plot only the growth exponent on the same time scale for both datasets. From this plot, we notice that the growth exponent is able to delineate various periods of increased or decreased superlinear growth. Moreover, it can help pinpoint various time points, where potential events might have taken place, that alter the growth rate of the network. In Fig. 9, we plot the first gradient of the growth exponent for both graphs. In this figure, it is clear that although in different scale, the two plots increase or decrease with the same gradient.

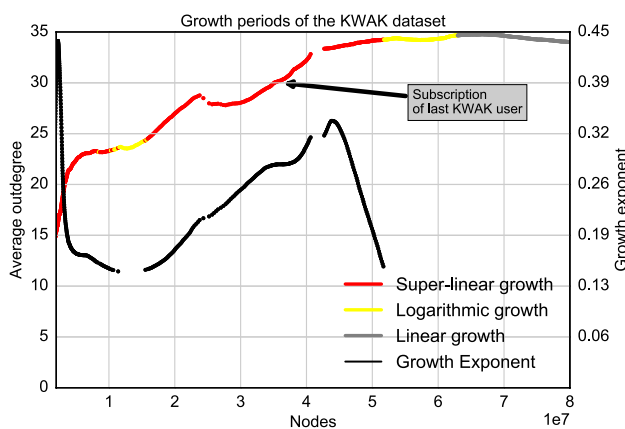


Fig. 5 Here we plot the evolution of the average outdegree for the KWAK dataset. The semantics of the lines in this plot are the same as in Fig. 4. Notice that the last user subscription of KWAK happened when this dataset had 37 million nodes. Therefore, the large drop of the growth exponent after that is an artifact and not a real event

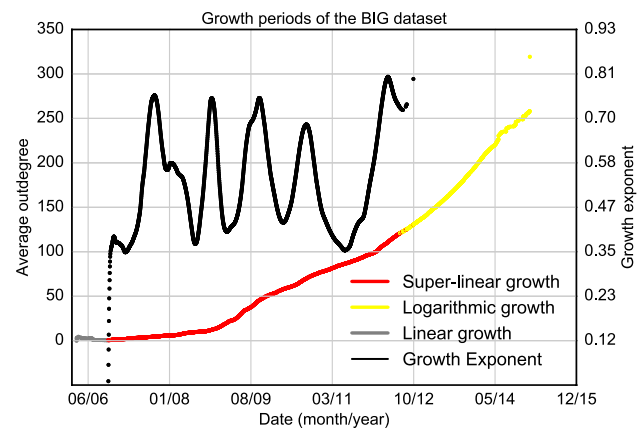


Fig. 6 Here, we plot the temporal evolution of the average outdegree and growth exponent for the BIG dataset. The semantics of the lines in this plot are the same as in Fig. 4. The only difference is that x axis shows the date, and y axis shows the growth exponent of the graph (black line) and the average outdegree (colored line) of the graph at that date (color figure online)

5 Events at the early stage of Twitter

As we have described, the growth exponent is able to capture various periods at the early stage of Twitter. These periods are marked either with increased or decreased superlinear growth. In Fig. 10, we plot the growth exponent annotated with events that have affected Twitter according to Wikipedia (2004). It is important to note that, in this study, we do not infer causal relationships between these events and the growth exponent, since a simple coincidence is not enough to justify a causal relation between an event and a growth

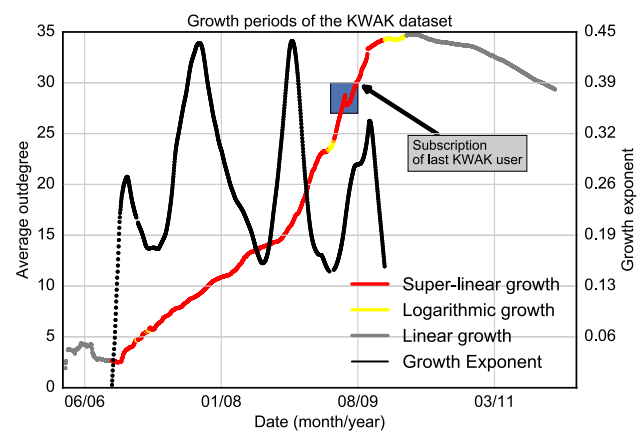


Fig. 7 Here, we plot the temporal evolution of the average outdegree for the KWAK dataset. The blue rectangle indicates a “bump” on the plot of average outdegree, which coincides with the disruption of the Twitter service from the death of Michael Jackson (discussed in Sect. 5.2) (color figure online)

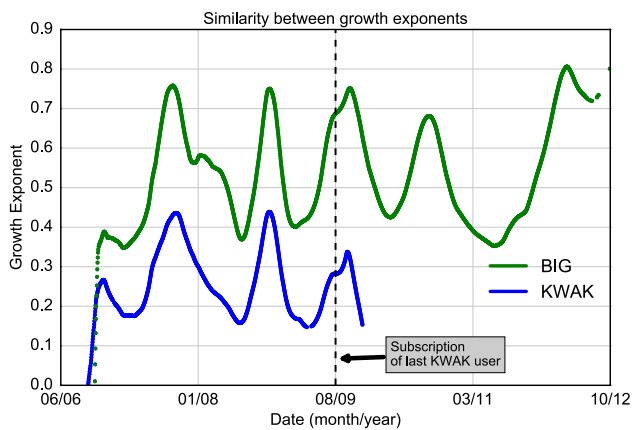


Fig. 8 Here, we plot the comparison of the growth exponent between the *BIG* and the *KWAK* datasets. The estimation of the growth exponent (g) depends on the number N of nodes in the graph ($\bar{d} = N^g$). The average outdegree (\bar{d}) between the two datasets, for the period before the subscription of the last *KWAK* user, is approximately the same (see Fig. 1). Nevertheless, since *KWAK* focuses exclusively on this time period, it contains more samples (nodes) than *BIG*. Therefore, our fitting model algorithm generated higher g values for the *BIG* dataset. However, we notice that, despite the inherent differences between the two datasets (size, users), the fluctuations of the growth exponent delineate almost the same time periods

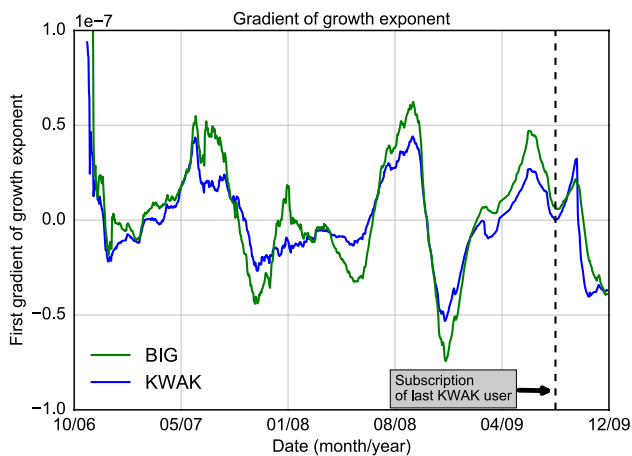


Fig. 9 The estimation of the growth exponent (g) is sensitive to various parameters of the sampling method. Yet, the rate of increase or decrease of g shows a relevant tolerance to these parameters. To demonstrate this, we plot the first gradient of g , between the *BIG*, and *KWAK* datasets

change. The purpose of this paragraph is to put these events into perspective, according to the changes of the growth exponent. Additional work is required, in order to quantify how these events might have actually affected (or not) the growth of Twitter.

Nevertheless, the importance of some of these events (like the SXSW conference) has been validated by Twitter’s

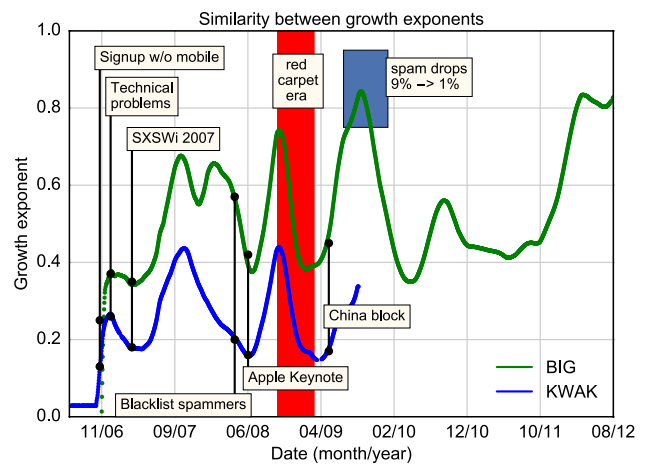


Fig. 10 Here, we plot the events that may have affected the growth exponent g of Twitter during its early period. When g decreases, then the average outdegree grows at a smaller rate. Periods of decreased g indicate higher rates of addition of new users, and periods of increased g indicate higher rates of new connections (followings)

officials. In the beginning (July 2006), Twitter was an experimental service developed exclusively for use with mobile phones. In October 2006, it was possible to sign up without the use of mobile phone (Widrich 2011). This change marked a transition to a regular OSN, and we also notice a first increase in growth exponent. In the end of 2006, several technical problems indicative of the service immaturity (Duncan 2007) may have slowed down Twitter’s growth. Moreover, several rival services (e.g., FriendFeed, Pownce, Jaiku, Brightkite) appeared to attract potential new users (Lardinois 2008). The decisive breakthrough of Twitter happened in March 2007, at the SXSW conference (Shah 2010), where Twitter won the top award and got a lot of attention. The user base of Twitter grew significantly, during this period.

In May 2008, Twitter applies its first action against spam, by massive deleting many spam accounts. Whether spam increases or decreases the superlinear growth is an open question that we discuss below. In June 2008, a lot of blogs and websites were expecting that Twitter will not withstand the extreme traffic from Apple’s keynote conference. Nevertheless, Twitter did not have any failures which was a sign of a transition to a more mature and stable service, and as a consequence an increase of the growth exponent. The period from November 2008 to April 2009 has been characterized as the “red carpet era” of Twitter, due to the attraction of many personalities from the show business industry. It is estimated (Judge 2010) that 54% of the most popular Twitter users started using Twitter during this period. The increase of growth rate is visible in the *BIG* dataset, but not in *KWAK*, and this is the only difference between the growth rate of the two datasets.

Another measurement that we performed, was the average outdegree per isolated day. For these measurements, we counted the average degree of the graph for each day, without taking into account any previously formed edges. This gives an evaluation of the density of the graph that was generated each day. We were surprised to find that this average degree was increasing each day until June of 2009, where it peaked and then started to decrease. We located two events that happened in this period. The first was the blocking of Twitter in China, and the second was the death of the famous pop artist Michael Jackson. In Fig. 12, we show these measurements annotated with these two events. On the following subsections, we discuss how these events might have affected Twitter.

5.1 Blocking from China

China blocked Twitter in early June 2009. Although we do not see any change in the growth exponent, we speculate that this might have reduced the average degree per isolated day, as we notice in Fig. 12. To test this hypothesis, we searched a dataset that contained the user objects of 250 million users.

A user object is a data structure that contains several meta-information about a user's profile, like language, location, creation time and other profile preferences. User objects can be requested from Twitter's API, and they include the last tweet of a user. In this dataset, we looked for users whose last tweet was tagged with geo-location information and we measured the percentage of those located in China. We also measured this percentage per year according to the account creation time, and according to the time this last tweet was sent. Unfortunately, Twitter, enabled geo-tagging of tweets in August of 2009 and it was very slowly adopted by users due to lack of support from Twitter clients (Bryant 2010). As an effect, statistics, prior to 2010, based on geo-tagged tweets are unavailable. To tackle this, we also looked into the "location" field of user objects. This field is a user-defined location, so there is no guarantee that the actual location of the tweet is in China. Nevertheless, since it was the only location information available for tweets prior to blockage, we also measured the percentage of users that specified their location as "China" (in English or in Chinese) in their profile.

In Fig. 11, we plot the percentages of Chinese users in Twitter, for each year, between 2006 and 2015. We used 4 different criteria to determine whether a user is Chinese or not. The first (black bars) is the account creation year of the users whose last tweet was geo-tagged in a location within China, and the second (gray bars) is the year of the last tweet for these users. The third (red bars) is the account creation years of the users who are self-described as Chinese in their profile, and the fourth (pink bars) is the year of the last tweet for these years.

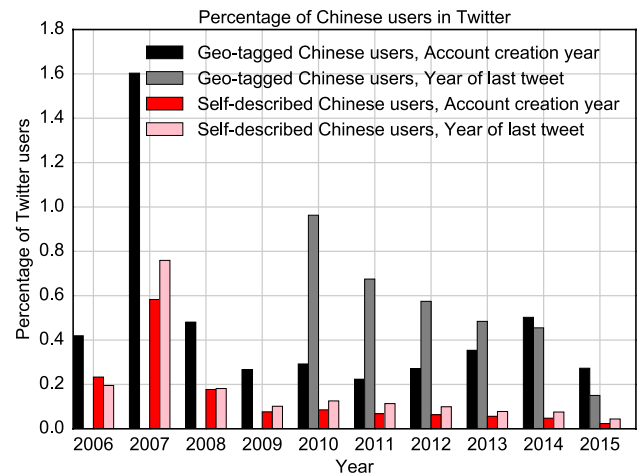


Fig. 11 Percentages of Chinese users in Twitter between 2006 and 2015. Black bars and gray bars show the percentages of users, who were identified as Chinese, according to geo-tagging. Red bars and pink bars show the percentages of self-identified Chinese users, according to the description of their profile. In black bars and red bars, we show the account creation year. In gray bars and pink bars, we show the year of their last tweet. Since geo-tagging was enabled in August of 2009, we do not have geo-tagged tweets before that period (color figure online)

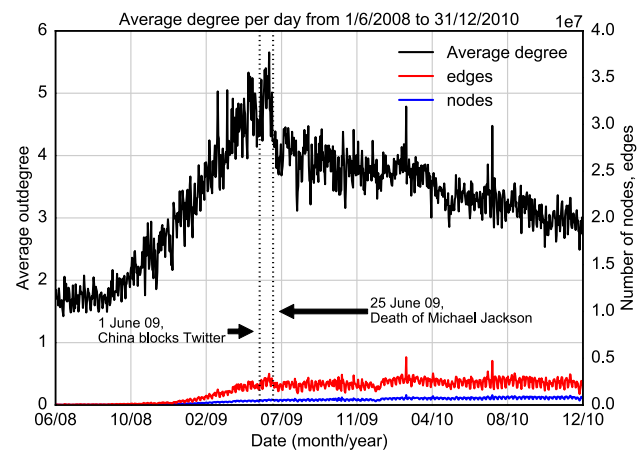


Fig. 12 For each day from 1/6/2008 until 31/12/2010, we extracted the following relationships that happened that day on the *BIG* dataset. Then, we constructed the social graph of each day, and we measured its average outdegree (black lines). We also show the number of nodes (blue) and edges (red) for each one of this daily graph. We notice that the daily average degree peaked between the dates when China blocked Twitter and the death of popular singer Michael Jackson. We also notice that the number of nodes and edges of the daily graph were stabilized on that period (color figure online)

In this figure, we notice that the percentage of Chinese users peaked at 2007. After 2007, the percentage drops even more than before the blockage takes place. From these measurements, the maximum decrease was 1.4% (from 1.6% at 2007 to 0.2% at 2009) for the account creation year of

users that posted geo-tagged tweets from China. Although in Fig. 12, we notice that in July 2009 the average degree per day starts to decrease, we believe that this large change cannot be attributed to the sudden blockage of only 1.4% of Twitter users.

5.2 Death of Michael Jackson

On 25 of June 2009, the excessive online traffic that sparked from the death of the famous pop artist Michael Jackson, created a disruption of many websites including Twitter. This disruption is noticeable, as a small “bump”, in the plot of average outdegree, in Fig. 7. From this technical problem, Twitter recovered quickly. Researchers have used this event, in order to study the propagation patterns of Twitter (Ye and Wu 2010), as well as the emotional content of related posts (Kim et al. 2009). However, this event might also have contributed to the increase of popularity of Twitter, in the long term, due to the publicity that this disruption reached. Figure 12 shows that before this disruption, the average outdegree of the daily graph was increasing linearly in time. On the day that coincides with the death of Michael Jackson, this increase stops abruptly. After that, the daily average outdegree decreases constantly and converges to a value close to 2. One hypothesis is that the death of Michael Jackson made Twitter suddenly increasingly popular, attracting users that enrolled in a high rate. Since new users have lower outdegree compared to older ones, this might have contributed to the decrease of the overall average.

5.3 Spam filtering

Spam in Twitter has been an important issue (Benevenuto et al. 2010). Studies have shown that the click-through rate of Twitter spam is significantly higher than mail spam (Grier et al. 2010). In August 2009, the percentage of spam tweets in Twitter had reached 9%, affecting its public image as a “clean” service that did not propagate spam or malicious sites. To mitigate this, in August 2009 Twitter embedded a spam filtering mechanism on its URL shortening service. According to a report from Twitter, this technique reduced the spam percentage to 1% in February 2010 (Chowdhury 2010). This period coincides with the beginning of a long decrease of the superlinear growth rate. This is shown as a blue shade in Fig. 10. The hypothesis in this case is that the superlinear growth rate was affected by spam accounts. In order to increase their target base, spam accounts were following as many users as possible with the hope that these users would follow back, thus making them potential targets for spam or malicious URLs. This might have contributed to the increase in the growth of average outdegree. The application of the spam filter stopped or slowed down this promotion technique and stabilized the average outdegree.

6 Discussion and conclusions

Existing models of the evolution of social networks are of extreme importance for elucidating their structure and explaining the behavior of their users. Unfortunately, the enormous size of modern OSNs and the prohibitive computational complexity of many essential graph properties make this modeling a cumbersome procedure. Here, we have demonstrated a computationally efficient method to model the growth of an OSN, based on the simple property of node average degree.

Our methodology consists of four distinct parts. The first is the application of the heuristic that approximates the friendship time-creation with a time complexity of $O(IEI)$. The second is the sorting of all edges according to this approximation. The third is the calculation of the average outdegree of the network, for every day, between June 2006 and January 2015 (in total 3100 days). The time complexity of this part is $O(TIV)$, where T is the total number of time periods ($T = 3100$). The final part is fitting the average outdegree for all days to the “Leskovec model” and requires minutes of computation in a commodity computer. The only computational challenging part is sorting all edges of the network, which can be easily parallelized. The rest computational part can take place in a single workstation. Overall, the complete computation required approximately one day, in a high end workstation (single 4-core Intel i7 processor, 3.4GHz, 16Gb RAM).

In our experiments, we demonstrated that, approximately, the same growth periods could be delineated from two fundamentally different samples of Twitter. The first (*BIG*) contains 92 million users and was created with the random walk sampling method. The second (*KWAK*) is approximately half of the size of the first and contains the friends and followers of a relatively old (2009) dataset.

We have also demonstrated how this method can portray fluctuations of growth, even years before the sampling of the OSN happened. In our case, we focus on events that happened more than 5 years before the OSN was sampled.

Although the outdegree distribution of Twitter is most likely not a power law, there is an open question, whether the moments of this distribution are well defined. The mean of a power-law distribution with exponent $\lambda < 2$ diverges, meaning that repetitive measurement in independent samples, will result in very large fluctuations (Newman 2005). The latest and largest study (Myers et al. 2014) concluded that this distribution is the best fit by a log-normal function, which has all its moments well defined. In order to fit the “Leskovec model,” we do not measure the average outdegree in independent samples, but instead we measure the average degree of the same sample of the social graph, on a day-by-day base,

therefore, we do not expect large fluctuations. This also is evident from the relative smooth form of the average outdegree plot, in Fig. 1. Nevertheless, the proper elucidation of the form of this distribution with adequately large sample sizes is a crucial open question that needs to be addressed.

Additionally, future work is required, in order to establish reliable causal relationships between the presented events and the alterations of the growth exponent. Additionally, we plan to apply heuristics for the estimation of the second half of the Leskovec model parameter, which is the diameter. This will allow us to estimate the effect and size of the “Shrinking parameter” of Twitter.

As a final comment, we believe that this approach will help researchers to model efficiently the evolution of large OSNs and delve into their past, in order to investigate “which” and more important “how” specific events alter their growth patterns.

Acknowledgements We would like to thank the anonymous reviewers that provided valuable comments and feedback. We are also grateful to prof. Marian Boguna and Kolja Kleineberg for the discussions and the contribution on the infrastructure at the University of Barcelona. Also we would like to thank Hariton Efstathiades and Demetris Antoniadis for their valuable comments as well as the University of Cyprus on the valuable contribution of their infrastructure in order to complete the experiments. This work was supported by the following research projects: FP7 Marie-Curie ITN iSocial funded by the EC under Grant Agreement No. 316808, UNICORN: Funded by the European Commission (H2020-ICT-2016-1/ICT-06-2016) and EUNITY: Funded by the European Commission (H2020-DS-2016-2017/DS-05-2016).

References

- Albert R, Jeong H, Barabási A-L (1999) Internet: diameter of the world-wide web, vol 401. Nature Publishing Group, London, pp 130–131
- Amaral LAN, Scala A, Barthelemy M, Stanley HE (2000) Classes of small-world networks. *Proc Natl Acad Sci* 97(21):11,149–11,152
- Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S (2012) Four degrees of separation. In: Proceedings of the 3rd annual ACM web science conference on WebSci '12, ACM Press, New York, NY, USA, pp 33–42. <http://dl.acm.org/citation.cfm?id=2380718.2380723>
- Barabási A (1999) Emergence of scaling in random networks. *Science* 286(5439):509–512. <https://doi.org/10.1126/science.286.5439.509>
- Barbieri N, Bonchi F, Manco G (2013) Cascade-based community detection. In: Proceedings of the sixth ACM international conference on web search and data mining, ACM, pp 33–42
- Barbieri N, Bonchi F, Manco G (2014) Who to follow and why: link prediction with explanations. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 1266–1275
- Batrinca B, Treleaven PC (2015) Social media analytics: a survey of techniques, tools and platforms. *AI Soc* 30(1):89–116
- Benevenuto F, Magno G, Rodrigues T, Almeida V (2010) Detecting spammers on twitter. In: Annual collaboration, electronic messaging, anti-abuse and spam conference (CEAS)
- Bliss CA, Frank MR, Danforth CM, Dodds PS (2013) An evolutionary algorithm approach to link prediction in dynamic social networks. *CoRR* abs/1304.6257. <http://dblp.uni-trier.de/db/journals/corr/corr1304.html#abs-1304-6257>
- Bray P (2015) Social authority: our measure of Twitter influence. <http://moz.com/blog/social-authority>. Accessed 20 Aug 2017
- Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A, Wiener J (2000) Graph structure in the web. *Comput Netw* 33(1):309–320
- Bryant M (2010) Twitter geo-fail? Only 0.23% of tweets geotagged. <https://thenextweb.com/2010/01/15/twitter-geofail-023-tweet-s-geotagged/>
- Chan J, Bailey J, Leckie C, Houle M (2012) ciForager: incrementally discovering regions of correlated change in evolving graphs. *ACM Trans Knowl Discov Data* 6(3):1–50. <https://doi.org/10.1145/2362383.2362385>
- Chowdhury A (2010) State of Twitter spam. <https://blog.twitter.com/2010/state-of-twitter-spam>. Accessed 20 Aug 2017
- Duncan R (2007) Making the switch from Twitter to Jaiku. <http://goo.gl/JMuhKA>. Accessed 20 Aug 2017
- Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on twitter networks: validation of Dunbar's number. *PLoS ONE* 6(8):e22,656. <https://doi.org/10.1371/journal.pone.0022656>
- Grier C, Thomas K, Paxson V, Zhang M (2010) @ spam: the underground on 140 characters or less. In: Proceedings of the 17th ACM conference on Computer and communications security—CCS '10, ACM Press, New York, NY, USA, p 27. <https://doi.org/10.1145/1866307.1866311>
- Judge P (2010) Barracuda Labs 2010, annual security report. Technical report. Barracuda Networks Inc
- Kim E, Gilbert S, Edwards M, Graeff E (2009) Detecting sadness in 140 characters. Webecology project
- Kleinberg J (2000) Navigation in a small world. *Nature* 406(6798):845. <https://doi.org/10.1038/35022643>
- Kleineberg K-K, Bogaña M (2014) Evolution of the digital society reveals balance between viral and mass media influence. *Phys Rev X* 4(031):046. <https://doi.org/10.1103/PhysRevX.4.031046>
- Kleinberg JM, Kumar R, Raghavan P, Rajagopalan S, Tomkins AS (1999) The web as a graph: measurements, models, and methods. In: Asano T, Imai H, Lee DT, Nakano S, Tokuyama T (eds) Computing and combinatorics. Springer, Berlin, Heidelberg, pp 1–17
- Kumar R, Novak J, Tomkins A (2006) Structure and evolution of online social networks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining—KDD '06, ACM Press, New York, NY, USA, p 611. <https://doi.org/10.1145/1150402.1150476>
- Kwak H, Lee C, Park H, 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 '10, ACM Press, New York, NY, USA, p 591. <http://dl.acm.org/citation.cfm?id=1772690.1772751>
- Lardinois F (2008) Twitter survives Stevenote—but FriendFeed was the place to be. <http://goo.gl/aGyGW0>. Accessed 20 Aug 2017
- Leskovec J, Faloutsos C (2006) Sampling from large graphs. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining—KDD '06, ACM Press, New York, NY, USA, p 631. <http://dl.acm.org/citation.cfm?id=1150402.1150479>
- Leskovec J, Kleinberg J, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceeding of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining—KDD '05, ACM Press, New York, NY, USA, p 177. <http://dl.acm.org/citation.cfm?id=1081870.1081893>

- Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. *ACM Trans Knowl Discov Data: TKDD* 1(1):2
- Leskovec J, Backstrom L, Kumar R, Tomkins A (2008a) Microscopic evolution of social networks. In: *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining—KDD 08*, ACM Press, New York, NY, USA, p 462. <https://doi.org/10.1145/1401890.1401948>
- Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2008b) Statistical properties of community structure in large social and information networks. In: *Proceeding of the 17th international conference on World Wide Web—WWW '08*, ACM Press, New York, NY, USA, p 695. <https://doi.org/10.1145/1367497.1367591>
- Marquardt DW (1963) An algorithm for least-squares estimation of nonlinear parameters. *J Soc Ind Appl Math* 11(2):431–441
- Meeder B, Karrer B, Sayedi A, Ravi R, Borgs C, Chayes J (2011) We know who you followed last summer: inferring social link creation times in twitter. In: *Proceedings of the 20th international conference on World Wide Web*, ACM, pp 517–526
- Morales A, Borondo J, Losada JC, Benito RM (2014) Efficiency of human activity on information spreading on twitter. *Soc Netw* 39:1–11
- Myers SA, Sharma A, Gupta P, 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, International World Wide Web Conferences Steering Committee*, pp 493–498
- Newman ME (2005) Power laws, pareto distributions and Zipf's law. *Contemp Phys* 46(5):323–351
- Sadikov E, Martinez MMM (2009) Information propagation on twitter. CS322 project report
- Shah D (2010) The March of Twitter: analysis of how and where Twitter spread. <https://goo.gl/RiWs4n>. Accessed 20 Aug 2017
- Strogatz SH (2001) Exploring complex networks. *Nature* 410(6825):268
- Wei W, Carley KM (2015) Measuring temporal patterns in dynamic social networks. *ACM Trans Knowl Discov Data* 10(1):1–27. <https://doi.org/10.1145/2749465>
- Widrich L (2011) How twitter evolved from 2006 to 2011. <https://blog.bufferapp.com/how-twitter-evolved-from-2006-to-2011>. Accessed 20 Aug 2017
- Wikipedia (2004) Timeline of twitter. https://en.wikipedia.org/wiki/Timeline_of_Twitter. Accessed 20 Aug 2017
- Yang J, Leskovec J (2011) Patterns of temporal variation in online media. In: *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, pp 177–186
- Ye S, Wu SF (2010) Measuring message propagation and social influence on twitter. *com. SocInfo* 10:216–231