Online Social Network Evolution: Revisiting the Twitter Graph

Hariton Efstathiades, Demetris Antoniades,
George Pallis, Marios D. Dikaiakos
Dept. of Computer Science
University of Cyprus
{h.efstathiades, danton,
gpallis, mdd}@cs.ucy.ac.cy

Zoltán Szlávik, Robert-Jan Sips
Center for Advanced Studies
IBM Benelux
Amsterdam, The Netherlands
{zoltan.szlavik, rjsips}@nl.ibm.com

***Abstract*—In 2010 the popular paper by Kwak et al. [11] presented the first comprehensive study of Twitter as it appeared in 2009, using most of the Twitter network at the time. Since then, Twitter's popularity and usage has exploded, experiencing a 10-fold increase. As of 2015, it has more than 500 million users, out of which 316 million are active, i.e. logging into the service at least once a month.[1] In this study we revisit the network observed by Kwak et al. to examine the changes exhibited in both the graph and the behavior of the users in it. Our results conclude to a denser network, showing an increase in the number of reciprocal edges, despite the fact that around 12.5% of the 2009 users have now left Twitter. However, the network's largest strongly connected component seems to be significantly decreasing, suggesting a movement of edges towards popular users. Furthermore, we observe numerous changes in the lists of influential Twitter users, having several accounts that where not popular in the past securing a position in the top-20 list as new entries.**

***Keywords*-Online Social Networks Evolution; Social Media; Twitter Graph Analysis; Removed Users Analysis;**

I. Introduction

Twitter is one of the most popular Online Social Network (OSN) to time. It first appeared in 2006 and has been receiving growing attention ever since. Nowadays, the platform has more than 500 million users, out of which 316 million are considered active, i.e. users who log into the service at least once a month. Twitter allows its users to publish short messages, 140 characters long (including videos, pictures and URLs), in order to communicate their ideas, products, emotional state with their followers. Over the years Twitter has been used in a variety of different situations, e.g. allowing protesters to communicate over the Arab Spring [17, 30]. The extensive usage of Twitter enables researchers to analyze the generated information for several applications such as event detection [1, 3], user location analysis [6, 18, 31], health care [26], recommendation and early warning systems [27], temporal trends and information diffusion [15, 22].

In their popular study of the Twitter network, Kwak et al., examined the full Twitter graph as it appeared in 2009 [11]. The dataset that they have collected and studied is the largest publicly available Twitter dataset according to the number of nodes and edges. With their analysis they provided insights about the overall network topology, online activity of the users and influential users that existed at that time. Their results summarize the characteristics of Twitter in 2009 and its power as a new medium of information sharing.

In this study we revisit the same sample of users and collect the full information that is available from the Twitter API. We collect a total of 34.6 million user profiles, connected through 2.05 billion relationships. Based on the provided insights and data, we aim in analyzing the Twitter network as is today, and provide a comparison with the snapshot of 2009.

We address the different characteristics of the 2009 Twitter network, as it appears to be connected today, and examine the changes in connectivity of the network in general and the users in particular. To the best of our knowledge this work is the first quantitative study on the entire Twittersphere, that examines the long term evolution of the Twitter network.

Our contributions can be summarized as follows:

1) We observe a network that gets denser through the years, with the number of edges between the users in 2015 being almost double than 2009.
2) We clearly observe a "rich-get-richer" phenomenon, since the increased number of edges is mainly directed towards the most popular users.
3) Despite the increased number of edges, network connectivity seems to be decreasing. The Largest Strongly Connected component of the network decreases by 20%, in number of nodes, showing that the connections not only increase in total but are also redirected.
4) In the 2009 most of the popular users where popular in both followers and PageRank classification. Our study reveals a decoupling of the two methods, where most popular users through PageRank are not necessarily the ones with the highest in-degree.
5) We identify the reasoning behind users who left the Twittersphere and correlate it with their position in the graph. Our analysis suggests that users who have been banned from Twitter have different degree distributions than others, while the participation in the

[1] https://about.twitter.com/company (Last accessed: Jun. 2016)

Snapshot	*Vertices*	*Edges*	*Density*
TW2009	40,103,281	1,468,365,182	1.83×10^{-6}
TW2015	34,664,106	2,056,655,361	3.42×10^{-6}
TW2009C	34,664,106	933,256,652	1.55×10^{-6}

Table I
DESCRIPTION OF THE 3 DIFFERENT TWITTER GRAPH SNAPSHOTS.

largest Strongly Connected Component of users who intentionally left the network is by 10% higher than the rest. Furthermore, PageRank classification suggests that several users maintained highly ranked positions before their disappearance.

The remainder of this paper is structured as follows: We describe the experimental setting and the datasets used in the paper in Section II; Section III presents the topological analysis of the Twitter network and the comparison between the different snapshots. In Section IV we rank users based on the number of followers and PageRank, and compare the results with the ones of Kwak et al. study [11]. Finally, Section V describes the study performed on users who have been disappeared from the network and present the derived insights. Section VI presents the related work and Section VII summarizes our findings and concludes the study.

II. COLLECTED DATA

Our analysis is based on two different snapshots of the same Twitter network: (i) the complete Twitter 2009 graph, as collected and shared by [11], and (ii) the collection of the same list of Twitter users and their social graph as it appeared in late 2015. The 2009 graph was made available by Kwak et al.[2] According to the authors, the dataset represents the complete social graph of Twitter in 2009. Using the list of Twitter users that appeared in *TW2009* we perform a large-scale collection, through the current version of the Twitter API[3], with respect to platform's terms of use and users' privacy.

In order to collect this large scale Twitter dataset in a short-period of time we perform a distributed data collection campaign. Since Twitter API policy has been updated from IP-based to Application-based [10], we follow a crowd-crawling approach asking Twitter users to authorize our multiple applications to make request for public information on their behalf. We manage to configure a large number of Twitter applications instances in order to reduce the waiting time between the requests[4]. We implement this approach on 3 different machines; an action that enables us to collect the ego-networks of 1.2M users per day.

[2]http://an.kaist.ac.kr/traces/WWW2010.html (Last accessed: Jun. 2016)
[3]https://dev.twitter.com/rest/public (Last accessed: Jun. 2016)
[4]https://dev.twitter.com/rest/public/rate-limiting (Last accessed: Jun. 2016)

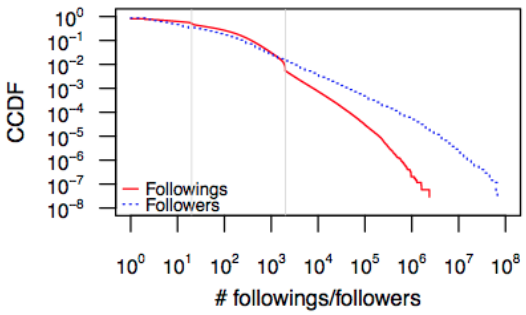


Figure 1. Complementary Cumulative Distribution Function (CCDF) of *followings* and *followers*.

Through this collection we retrieve the same set of Twitter users and their ego-network state (followers and followings) in November 2015. From this network we remove any connections (edges) that are directed towards or coming from users who do not belong in 2009 set. Thus, our *TW2015* snapshot contains only the connections that existed and have arise between the users that consisted the Twitter social network in 2009.

Table I presents the details of the two snapshots. As a first general observation we can see that more than 5 million users from *TW2009* have disappeared in the *TW2015* snapshot. The reason for a user not to appear in the snapshot can be explained through three different scenarios: (i) the user has been banned from the network due to violations of the terms of use (ii) the user intentionally removed her account deleting herself from the Twitter Online Social Network platform (iii) the user updated her privacy settings and made her information (profile and ego-network) private (not publicly accessible through the Twitter API). We further examine the properties of these three user categories in Section V.

In addition to the two full graphs of the 2009 Twitter users we also examine and compare, where relevant, the 2009 graph as it would appear if the users that belong to the above three categories where not existent in 2009. The *TW2009C* presents the snapshot of this case.

As we can see from Table I, the social network has become denser through the years. The connections between the same network users in 2015 are more than double the ones that existed in 2009 (*TW2015* Vs. *TW2009C*). This observation shows that the same set of users are constantly identifying each other creating new connections between them and becoming interested in the content they share. In the next section we examine in more detail the difference between the snapshot connections, trying to identify whether these new relationships are additional to the ones existed in 2009 or whether there is a general move of connections, with some users losing their followers while others gain more attention.

III. The Twitter Graph Evolution

In this section we present a study on the different snapshots of the 40.1M users of Twitter, regarding their graph metrics. For each analysis step we describe the procedure followed, results and derived insights. Furthermore, we present a comparison between the networks and discuss their topological differences.

A. Basic Analysis

Similarly to Kwak et al. [11], we analyze the characteristics of the followers and followings of the TW2015 users. The following relationship is directly related with a user's action: the individual chooses to follow another profile due to her own reasoning. On the other hand, the follower relationship is influenced by an indirect action; an individual maintaining an active profile, posting interesting content, with the goal of and attracting new followers.

Figure 1 plots the complementary cumulative distribution function (CCDF) of the number of followers (dotted blue line) and followings (solid red line). The followings case presents similar glitches as in 2009. According to Kwak et al. [11] the glitch on $x = 20$ is an inherent consequence of the Twitter initial recommendation of 20 people to follow, when a user first creates an account. The observation of the same glitch in the network snapshot taken 6-years later, shown by the first vertical line in Figure 1, intrigue us to investigate this further. We analyze the group of users who follow exact 20 other accounts, and we see that on average they have less than 19 followers while their average number of tweets is less than 32. With this in mind, we conclude that these are inactive users who did not use Twitter after their initial sessions.

The next glitch that we observe is the one at around $x = 2000$. Twitter imposed a limit at 2,000 followings, for each user. After that number any increase in the number of followers is correlated with the follower-following ratio and it is account specific. Myers et al. [24] in 2014, examined this limit and concluded that the platform does not allow users to follow more than 2,000 accounts unless they themselves have more than 2,200 followers. This threshold has been updated to 5,000 followings, according to Twitter[5]. For an account to get 2,200 followers is a difficult process, needing a lot of effort to interest people. Most accounts will never be able attract that many followers and for this reason will remain in the limit imposed by the service. However, this does not mean that the set of followings of an account remains unchanged through the years. Users may select to remove accounts that are not interesting anymore, in favor of new more interesting accounts.

The follower distribution in Figure 1 shows the presence of several celebrity users in Twitter. These users attract more than 10^5 followers. The large majority of users has less followers and fits to a power-law distribution with an exponent of 2.101. This value is lower than the 2.276 observed by Kwak et al., however it remains in the boundaries between 2 and 3 that characterize the majority of real-world networks, including OSN [11].

[5]https://support.twitter.com/articles/66885 (Last accessed: Jun. 2016)

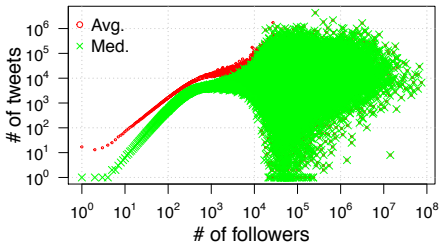


Figure 2. The number of *followers* and that of *tweets* per user.

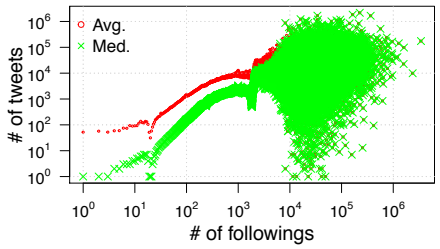
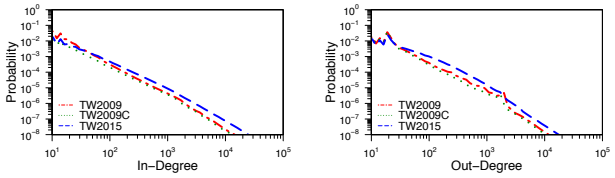


Figure 3. The number of *followings* and that of *tweets* per user.

B. Followers vs. Tweets

A common perception regarding Twitter is that the more active a user is (that is the more content she shares), the more followers (attention) she gets. Kwak et al. observed this perception to be true only for users with up to 5,000 followers. After that point there was no obvious correlation between the number of followers and that of the tweets of a user. Figure 2 presents the results of the same analysis for the TW2015 snapshot. The figure plots the number of followers as a function of the median (green cross) and average (red circle) number of tweets for each user. Comparing with the TW2009 study, we can see that the results are similar for users who have less than 10 followers; the majority has never tweeted or did just once, maintaining a median value of 1. Similarly, the existence of outliers who tweeted much more than the expected, based on their followers counter, preserve an average value always higher than the median in regards to the number of tweets. Furthermore, the flat line observed between the values of 100 and 1000 followers in 2009, has been moved to 1000 and 1500, as the number of followers seems to be increasing.

Figure 3 examines the relationship between the activity of a user (number of tweet she posts) as the number of the people she follows increases. It plots the number of followings and median and average number of tweets per



(a) In-Degree (b) Out-Degree

Figure 4. In-degree and Out-degree of the 3 different Twitter snapshots.

	25%	50%	75%	100%
In-TW2009	2	8	17	2.99M
Out-TW2009	4	9	21	770K
In-TW2009C	2	4	8	2.57M
Out-TW2009C	3	9	20	662K
In-TW2015	2	5	19	3.21M
Out-TW2015	6	20	69	608K

Table II
STATISTICS OF THE AVERAGE DEGREE DISTRIBUTIONS FOR THE 3 NETWORKS.

each user in our dataset. The two irregularities at $x = 20$ and $x = 2000$ observed in Figure 1 also appear in this plot. Furthermore, the additional irregularities observed by Kwak et al. and attributed to spam accounts have disappeared, an expected consequence since Twitter removed these accounts.

C. Degree Distribution

Twitter follower graph is a directed graph $G = (V, E)$, where each vertex $v \in V$ represents a user in the network while each edge $e \in E$ represents a directed follower relationship between the 2 vertices $\{v_s \rightarrow v_d\}$. Thus, each vertex has an in-degree, which represents the number of its followers, and an out-degree which represents the number of its followings.

In the previous section we observe that the number of connections between the Twitter users has almost doubled during the time period separating the two collections. In this section we examine the degree distribution of the three networks to answer whether this increase can be attributed to a small number of nodes that increased their incoming or outgoing connections tremendously or whether the majority of users has participated in this increase. Since Twitter is a directed network we examine both the in-degree, the number of the user's followers, and out-degree, the number of a user's followings.

Table II shows the 25th, 50th, 75th and 100th percentiles of the degree distribution in the different Twitter snapshots. With an exception for the popular users (100th percentile), in all other cases the in-degree is smaller than out-degree. In Twitter terms, this means that the average user is being followed by less users that the ones she follows. This observation holds for both 2009 and 2015 graphs, which reveals a non-era binded finding. Additionally, we can observe a rich-get-richer phenomenon, as the in-degree of the most popular users increases. On the other hand, less popular users show an out-degree that almost triples in some cases. This observation leads us to conclude that the increase in the number of edges observed from TW2009 to TW2015 is due to popular users getting more follow relationships coming from the rest of the network.

Figure 4(a) plots the in-degree distribution for the 3 different snapshots. As expected, we observe a heavy tail power-law distribution in all cases. From Figure 4(b) we can see that the out-degree also follows a heavy tail power-law distribution but not at the same extent as in-degree; a fact also described by [24].

Figure 4(b) presents a spike at the value of 2000 out-degree nodes for both 2009 snapshots. We examine the reasoning behind it in Twitter mechanisms and found that the platform applies an anti-spam/bots strategy regarding the number of accounts that an individual user can follow. In recent studies spammers and bots have been characterized by very low values of $\#Followers/\#Followings$ ratio [5]. For that reason Twitter sets a limitation of 2000 followings for each account who has less than 2200 followers [24]. In the 2015 snapshot we do not observe similar spike, as the threshold strategy has been changed during recent years[6].

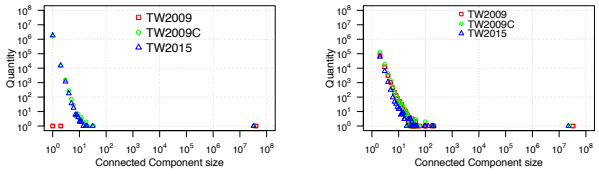
Discussion: The results suggest that the increase in the number of edges observed from TW2009 to TW2015 is due to the increasing number of connections that popular users attract, coming from the rest of the network. Establishing a relationship in Twitter denotes that a user follows another and is able to receive notifications and read the content the latter publishes. However, the large number of out-degree, and thus followings, would reasonably be a problem for a user to easily access and read information of interests. However, as reported in [11], Twitter looks like more an information network instead of a social network. Furthermore, the platform of Twitter provides to its users the functionality to 'Mute' an account; the following relationship remains but the content that the muted user publishes does not appear in the tweet feed of the one who muted her.

D. Connected Components

We now turn our attention in examining how the connectivity of the network changes, as this can be seen through the number of Strongly and Weakly Connected Components. A Strongly Connected Component in a directed graph is a subgraph where there exists a path from every node to every other node in the subgraph. A Weakly Connected component is a sub-graph where all nodes are connected with some path, ignoring the direction of the edges.

Figure 5 plots the distribution of the size of the Weakly and Strongly Connected Components for the 3 Twitter snapshots. As we can see from Figure 5(a), in all cases a large connected component maintains an enormous size compared

[6]https://support.twitter.com/articles/66885 (Last accessed: Jun. 2016)



(a) Weakly Connected Components (b) Strongly Connected Components

Figure 5. Strongly and Weakly Connected Components of the 3 different Twitter snapshots.

to the rest components. In Twitter 2009 graph a single Weakly Connected component covers more than 99.9% of the nodes. Despite the fact that the number of Weakly Connected Components has been increased in the graph of 2015, the coverage of the largest WCC is still very high, as it contains 94.58% of the graph nodes. We also see that 5.52% of the nodes change Weakly Connected Components in 2015 snapshot. Finally, if we exclude removed users from 2009 graph, we observe an increase on the quantity of WCC, while the largest contains 95.58% of the nodes.

Studying the Strongly Connected Components enable us to extract more interesting insights for the case of Twitter, as in such components the direction of the edge is not ignored. Due to the fact that Twitter graph is directed the metric has different meaning than WCC. From Figure 5(b) we observe that in all 3 cases the largest Strongly Connected Component covers the largest portion of the graph, while several others maintain a much smaller size. In 2009 the largest SCC covered a large percentage of the graph, 83.90%, a higher value compared to the 65.56% of the largest SCC in 2015. The largest SCC in 2009 seems to be closer to the coverage observed in other social graphs, such the ones of MSN messenger [13] and Facebook [29], which show a coverage of more than 99%.

An important observation from both cases is that despite the fact that the network is becoming denser, it seems to be disconnecting. While the largest WCC in 2009 included almost all the network, in 2015 the number of WCC increases, showing a number of sub-graphs that are disconnected from the rest of the network. Furthermore, the largest SCC size decreases by almost 20%, and we observe an increased number of smaller SCC. Taking into account that popular users are the ones that actually increase their incoming connections, we might consider that Twitter users decide to remove edges from non-popular users to target more popular ones. This change limits the paths that connect users between them, resulting in more groups of fewer nodes that can be reached by each other.

To put this into perspective, we calculate the percentage of users who appear in different Connected Components in 2009 and 2015 snapshots. The comparison regarding Weakly Connected Components show that 5.52% of the nodes change component in the evolving snapshot, while none of them left the largest WCC. In contrast, in the case of the Strongly Connected Components we observe that 72.43% of the vertices have moved to a different one during the 6 years period, having 22.94% leaving and 4.60% joining the largest.

Discussion: From the results we derive the insight that the structure of the network has changed significantly regarding the Strongly Connected Component. We observe a decrease of about 20% in the coverage of the largest SCC between 2009 and 2015 snapshots. One possible reasoning of this fact is that Twitter in its early years was used as a social networking platform. As the years past, the network has been evolved and changed to a more information dissemination platform, where users connect with accounts who post content that lies in their interests instead of the one with whom they share physical-world relationship. This resulted to a more sparse network, where clusters between users who share similar interests have been created.

E. Reciprocity

As presented in [11], the reciprocity of the 2009 Twitter snapshot does not exceed 22.1%, meaning that only this amount of user pairs follow each other. The rest 77.9% of the pairs are single sourced, thus they only share one relationship. The reciprocity in 2015 Twitter snapshot increases to 29.2%, showing that more and more users tend to follow back the ones who follow them. However, this number is still much smaller than the reported values of 68%, 79% and 84% for Flickr [4], Yahoo!360 [8] and YouTube [21] respectively.

Reciprocity in directed Online Social Networks is often considered as a measure of a stronger connection between two users [2]. However, in Twitter the follow-back mechanism is also used for more practical reasons, such as to recruit more followers. A number of users mentions in their description fields that they follow back, in order to attract others to follow them so as to increase their number of followers. As we observe, that mechanism in not popular as only 10,656 users have added this information in their profile description fields.

F. Edges Comparison

The study of the users behavior in OSN platforms has gain the attention of researchers during past years. The majority of the studies has been focused on the content that individuals publish and how their behavior change in time [7, 9, 16]. Having collected two different snapshots with a difference of several years, we study the behavior of users regarding the connections created and removed over the time.

Figure 6 plots the average value of the ratio between newly created and removed out-going connections in relation to the number of out edges of each node (out-degree). Users

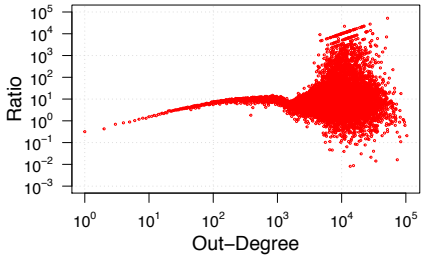


Figure 6. The Out-Degree and the ratio between newly created and removed out-going edges.

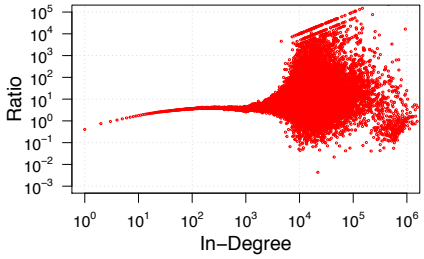
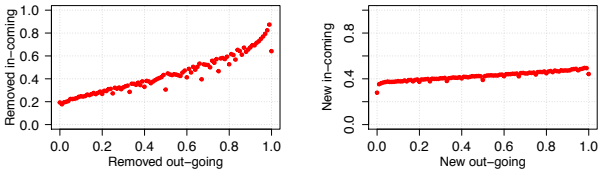


Figure 7. The In-Degree and the ratio between newly created and removed in-coming edges.

with an out-degree of less than 5,000 tend to remove 1 edge for every 6.33 created. For users with an out-degree of less than 100 this ratio decreases to 3 on average. The latter result is similar with the one estimated by Myers et al., who observe 1 removal for every 3 created [23]. Furthermore, for users who have an out-degree of more than 500, the ratio between newly created and removed edges is 5.73. The results suggest that users tend to create edges in a higher rate than removing. However, this ratio does not exceed the value of 8 for any case when $x <= 10,000$.

With Figure 7 we examine the average value of the ratio between newly created and removed in-coming connections in relation to the number of in-coming edges of each node (in-degree). As we can see, the ratio is increasing steadily for users with less than 100 followers until it reaches a value of 3.21, while it remains almost stable at 3.51 between $x = 100$ and $x = 1000$. Users who maintain an in-degree between 1,000 and 5,000 edges tend to lose 1 follower for every 4.24 new in-coming connections, while for more popular users, between 5,000 and 10,000 followers, the ratio increases to 5.44. Furthermore, for users who maintain an in-degree of more than 10,000 edges, which are mostly celebrities, the ratio between newly created and removed edges decreases to 3.66.

Figure 8(a) plots the fraction of removed out-going connections as a function of the average fraction of removed in-coming connections. As we can derive, the fraction of out-going connections removal increases linearly with the



(a) Fraction of removed out-going and in-coming edges.

(b) Fraction of newly created out-going and in-coming edges.

Figure 8. Fractions of removed and newly created edges.

one of the in-coming connections. From this result we can conclude that users un-follow other accounts in similar rate as their followers remove the connections towards them. In Figure 8(b) we plot the fraction of newly created out-going connections as a function of the average fraction of newly created in-coming connections. As we can see, there is a slight increase of new in-coming connections related with the increase of new out-going connections. However, this increase is not at the same scale as the one in out-going connections.

Discussion: These results enrich the hypothesis that in-coming connections are not directly related with a user's action; a user cannot increase her in-coming relations by direct actions, unlike the out-going relations. In order to attract more followers a user should post content that fits the interests of others, and trigger them to follow her account, instead of following other to follow her back.

G. Degree of Separation

The small world phenomenon refers to the surprisingly small distance that actually separates two users in a social network. Kwak et al. [11], examined the full Twitter graph as it appeared in 2009 and their analysis on the structural properties of the graph shows an average shortest path length of 4.12.

The calculation of the average shortest path between all pairs of vertices is computationally infeasible, due to the large scale of the collected dataset. Thus, we employ a sampling procedure similar to the one performed by [11]; we randomly retrieve a group of 2,000 users, that we call *seeders* and calculate the shortest path between them and all other vertices in the graph. The calculations have been performed using the single source shortest path algorithm, which we have developed on GraphChi [12]. Figure 9 presents the results on the degrees of separation between the seeders and all other vertices in the network. Our results show that the distance between two nodes in the graph is 4.05, while the median is 4.29 intermediates. As we can observe, the average shortest path value has been slightly decreased in the past 6 years.

Discussion: Despite the fact that a large number of Twitter users have been disappeared from the network, as presented

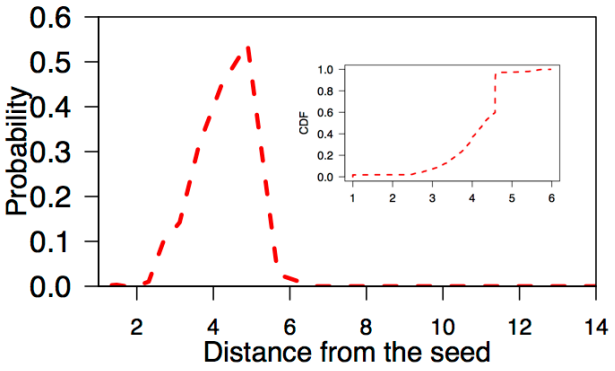


Figure 9. Distribution of degrees of separation between 1000 random chosen users and the rest of the network. Inner plot shows the cumulative distribution function for the same shortest paths.

in Section V, the length of the average shortest path has been reduced. At the time of Kwak's et al. study [11], Twitter graph had an average value much smaller than Facebook; users where separated by 4.12 and 4.74 intermediaries on average respectively. However, a recent study from Facebook[7] shows that the average degree of separation in the network gets smaller and reaches an average value of 3.57 intermediaries while within the US, people are connected to each other by an average of 3.46 nodes. Compared to our observation of 4.05, we conclude that the average shortest path in Twitter decreases in time but with much smaller coefficient than Facebook. This result could be explained by the different type of the two graphs, as Twitter is directed while Facebook is undirected.

IV. Rankings

In this section we present the results regarding two different popularity metrics on Twitter users. For each profile, Twitter maintains a counter that reveals the number of users who follow the corresponding profile. As reported by Kwak et al. [11], this metric does not reflect the topological influence of the node; i.e. the number of influential users who follow her. Thus, we proceed to another ranking procedure, using the widely used PageRank algorithm on the collected social graph [25].

A. By Followers

We use the straightforward approach of ranking users by descending order based on the number of their followers. As shown on Table III, the users contained in this list are very different than the one published in 2009 [11], as we observe 65% new entries. From the rest 7 users, only Barack Obama manage to improve his corresponding 2009 position, while Oprah Winfrey and CNN Breaking News accounts, who appeared in top-5, are now outside of the top-15 rankings.

[7]Three and a half degrees of separation, https://research.facebook.com/blog/three-and-a-half-degrees-of-separation (Last accessed: Jun. 2016)

B. By PageRank

We apply the PageRank algorithm on Twitter 2015 graph, which contains 34.6M users, connected with 2.05B directed edges. Each node of this network represents a user, while each edge a following relationship. We calculate the PageRank value using the GraphChi cpp implementation. Table III shows the top-20 list regarding this value, with a column that describes the updates regarding their difference between the 2015 and 2009 rankings. As we can see, despite the fact that users of 2015 list are by 80% the same, only 10% of them maintain the same rank position. Furthermore, we observe that 35% of the top-20 entries have improved their rankings, while the same fraction appears in a position lower than in 2009.

C. Discussion:

From the comparison between the 2009 and 2015 top-20 rankings lists we observe significant differences. We find this differences related with physical world events, i.e Barrack Obama maintains or improves his rankings as he also upholds his influence in the physical world. On the other hand, Ashton Kutcher appeared as 1st in both Followers and PageRank 2009 rankings before his famous divorce with Demi Moore; for the 2015 rankings he is outside top-20 and in 5th position, respectively.

Comparing the two lists of 2015 we can see that only less than half of the users is presented in both. As we observe, the top-2 users in followers rankings do not belong in the PageRank list, while the complete top-4 PageRank list maintain a position in top-20 followers list, having 3 of them in top10. Kwak et al. observe in 2009 that although the two lists do not match exactly, users are ranked similarly by the number of followers and PageRank. However, from our 2015 study we conclude that the two rankings lists show significant differences. For example, Katy Perry has the most followers, but does not belong to the top-20 PageRank list, while 'CNN Breaking News', ranked 17th in followers, is ranked 3rd in PageRank. This fact could imply that Katy Perry's followers are mostly teenagers or average individuals with low PageRank, while 'CNN Breaking News' has many heavy-weight followers.

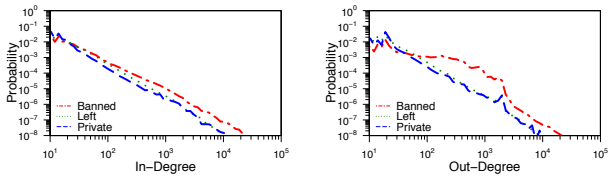
With these results we conclude that the number of followers does not provide us with strong insights regarding the topological influence of a user in the Twitter network.

V. Removed Users

Several studies have been performed on the characterization of the Twitter graph topology [11, 24] and users demographics [20]. Moreover, Liu et al. perform a study on the evolution of users behavior and highlight the rise of spammers and malicious behavior [16]. Thomas et al.

	Ranking by Followers			Ranking by PageRank		
Rank	Screen Name	Name	Change	Screen Name	Name	Change
1	katyperry	KATY PERRY	New	TheEllenShow	Ellen DeGeneres	+3
2	justinbieber	Justin Bieber	New	BarackObama	Barack Obama	=
3	BarackObama	Barack Obama	+4	cnnbrk	CNN Breaking News	=
4	taylorswift13	Taylor Swift	New	twitter	Twitter	+5
5	YouTube	YouTube	New	aplusk	ashton kutcher	-4
6	ladygaga	Lady Gaga	New	britneyspears	Britney Spears	-1
7	jtimberlake	Justin Timberlake	New	Oprah	Oprah Winfrey	-1
8	TheEllenShow	Ellen DeGeneres	-5	jimmyfallon	jimmy fallon	+4
9	twitter	Twitter	-3	nytimes	The New York Times	+7
10	britneyspears	Britney Spears	-8	KimKardashian	Kim Kardashian West	+10
11	KimKardashian	Kim Kardashian West	-1	RyanSeacrest	Ryan Seacrest	-1
12	shakira	Shakira	New	TheOnion	The Onion	+7
13	selenagomez	Selena Gomez	New	SHAQ	SHAQ	-6
14	ArianaGrande	Ariana Grande	New	lancearmstrong	Lance Armstrong	-3
15	ddlovato	Demi Lovato	New	taylorswift13	Taylor Swift	New
16	Oprah	Oprah Winfrey	-11	StephenAtHome	Stephen Colbert	New
17	cnnbrk	CNN Breaking News	-13	stephenfry	Stephen Fry	+1
18	jimmyfallon	jimmy fallon	New	mashable	Mashable	New
19	Pink	P!nk	New	google	Google	New
20	Drake	Drizzy	New	justdemi	Demi Moore	-6

Table III
TOP-20 USERS RANKED BY THE NUMBER OF FOLLOWERS AND PAGERANK IN THE TWITTER 2015 SOCIAL GRAPH. USERS WHO BELONG IN BOTH LISTS ARE HIGHLIGHTED. COLUMN *Change* REPORTS THE UPDATE FROM TW2009 POSITION IN TOP-20 RANKINGS.



(a) In-Degree (b) Out-Degree

Figure 10. In-degree and Out-degree of the 3 different categories of removed users.

Removed Reason	*Size*	*In-Degree*	*Out-Degree*
Banned from Twitter	1,042,060	61.77	263.68
Intentionally Left	4,365,923	28.12	28.88
Privacy Settings	179,800	21.37	23.48

Table IV
SAMPLE SIZE, AVERAGE IN-DEGREE AND OUT-DEGREE VALUES FOR THE 3 DIFFERENT CATEGORIES OF REMOVED USERS.

analyze the behavior of suspended accounts and present insights regarding their OSN behavior [28].

In this section we present a study on the graph structure of users who were part of the graph in 2009 but do not belong in the Twittersphere anymore. In this study we consider all removed accounts and not only users who have been suspended from Twitter mechanism as in [28]. We divide these users in different groups, based on the reasoning behind their disappearance. We conclude to the following categories of users: (i) who intentionally removed their account, (ii) who updated their ego-network visibility settings to private, (iii) who have been banned from Twitter due to their OSN behavior (e.g. bots, spammers). In the rest of this section we describe the graph metrics of these groups and extract insights on the comparison between them.

A. Degree Distribution

Studying the degree distributions (Figure 10) enable us to extract insights regarding the position of a removed user in the network before its disappearance and correlate it with the reasoning behind the latter.

Table IV presents the average numbers of the in-degree and out-degree for each one of the 3 categories. As we can derive, users who have been banned from Twitter maintain a much larger out-degree, about 9-times more than the other categories, while the value for the rest two categories differ by only 5.4 nodes on average. Furthermore, the latter two categories maintain a ratio between out-degree and in-degree of about 1, which is an indication that these were maintained mostly by individuals. However, the case for the users who have been banned from Twitter monitoring services is completely different; despite the fact that these users maintain a larger value of in-degree than other categories, the out-degree/in-degree ratio has a value higher than 4.2.

Discussion: From the results we can conclude that users who have been banned from Twitter have showed a degree distribution which has been observed on bots and/or spammers in past studies [5]. Regarding the rest two categories, we can see that they showed similar degree characteristics before their disappearance and can be related with an average non-active Twitter user. From these findings we can conclude that Twitter social graph eventually gets cleaned from non-active users, as they tend to disappear from the network either by intentionally removing their accounts or by maintaining strict privacy settings.

B. Connected Components

For the 3 categories we examine the Weakly and Strongly Connected Components that they participate in 2009. In the

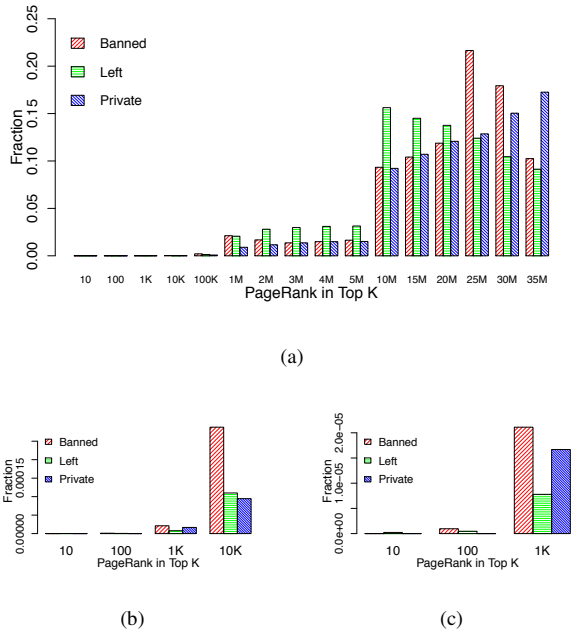


Figure 11. PageRank in highest ranking lists for the 3 different categories of removed users.

Weakly Connected Components (WCC) the direction of the connectivity is ignored, while in the Strongly is taken into consideration. Studying the connected components of the nodes who left twitter enable us to derive insights about their involvement in the social graph.

Regarding the users who have been banned from Twitter, 80% of them participate in the largest strongly connected component. Similarly, 81% of users who have update their privacy settings participate in the same Strongly Connected Component (SCC). However, the fraction of participation increases in the case of users who intentionally deactivate their profile, as 88% of them participates in the largest SCC. The percentages observed for the 3 categories of removed users are much higher than the corresponding values for the average twitter users presented in Section III and also observed by Myers et al. [24].

For the case of the Weakly Connected Components, we observe that in the 3 cases of removed users, all of them (100%) participate in the largest component. As presented in Section III, more than 99.9% of the overall Twitter users participate in the largest WCC in the 2009 graph.

C. PageRank

The degree distribution and connected components metrics give us an overview on the graph activity of a node in the network. In order to have a better overview on the influence of a node in the topology we use PageRank algorithm. We apply PageRank algorithm on the complete Twitter network of 2009 and extract insights regarding the different categories of removed users.

Figure 11 presents the fraction of removed users in the top rankings for each category. As we can see in Figure 11(a), the largest fraction of users who update their privacy settings (blue) appears in the lower ranking lists. Regarding the highest rankings, we observe that users who have been banned from Twitter hold the largest fraction until the top-2M list, were users who intentionally left take over.

Figures 11(c) and 11(b) present the results of the highest ranking lists. Surprisingly, we observe 60 users in total, who belong in the top-1K lists and are not part of the network today. The majority of these users have been banned from Twitter, while several others highly ranked users have changed their privacy settings. Furthermore, several users who intentionally left the network appeared in the highest rankings; 3 were in the top-100 lists, while one of them held a position in the top-10 higher ranked users of the network.

VI. Related Work

Networks constructed by people and their relationships in the physical world have been extensively studied. During the past years scientists have performed different experiments where they form hypotheses regarding the relationship between people and the overall structure of the underlying networks, with the most famous being the Milgram's experiment which lead to the famous small-world phenomenon [19].

The nature of Online Social Networking platforms enables Leskovec et al. [13] to study the Milgram's theory, using data generated by users of MSN messenger. In specific, they study the mean distance between users who interact through this platform. Their results show that the average degree of separation is 6 intermediaries and people who share similar physical world attributes, such as age and locations, tend to create connections and maintain a more frequent communication between each other than with users whom their characteristics differ.

Furthermore Leskovec, Kleinberg and Faloutsos [14] perform a study on the evolution of 4 real-graphs (ArXiv citation, U.S Patents citation, Autonomous Systems-AS and affiliation graphs) aiming in identifying the growth patterns of such networks. The assumptions that the average node degree remains constant and the diameter is slowly growing over time are examined. Surprisingly, they observe that the aforementioned assumptions and common-truths do not hold. Contrariwise, they show that the networks are becoming denser over time, with the average degree increasing; these results show that the number of edges grows superlinearly in the number of nodes. Furthermore, they show that the effective diameter of the networks is decreasing as the network grows over time.

VII. Conclusions

In this study we revisit the Twitter network as it appeared in 2009 and re-collect the users full characteristics as of late-2015. In total we retrieve 34.66M users connected by 2.06B social connections. We perform a comprehensive study of the 2009 and 2015 social graph snapshots and present the results regarding various metrics in the topology of the social graph. In specific, we compare the two network snapshots and study the distributions of followers and followings, the relation between followers and tweets, reciprocity, degrees of separation, connected components and differences in newly created and removed edges. Our results show a denser network with increased reciprocity but lower connectivity, as shown by the decrease in the networks largest strongly connected component. The average shortest path of the network also slightly decreases to 4.05 hops. We then examine the influential users of the network, as these can be defined by the number of followers and PageRank metrics. Our results show a significant change of these users between the years.

Having access to the entire 2009 Twittersphere, we identify users who do not belong in this directory anymore and investigate the reasoning behind their disappearance. We group removed users based on the reason they left the network and present a detailed comparison of the topological characteristics. We show that they have significant differences from the remaining set of users regarding their degree distributions, participation in Weakly and Strongly Connected Components, and their influential position in the social graph using their PageRank rankings. The results suggest that users who have been banned from Twitter showed different degree distributions than other categories, while the participation in WCC and SCC is much lower than the rest of the users. To the best of our knowledge this work is the first quantitative study on the entire Twittersphere, which compares the evolution of the network in such a large scale. We also introduce the study on removed users, where we group them in different fields and investigate their position in the social graph before their disappearance.

Acknowledgments

This work was partially supported by the iSocial EU Marie Curie ITN project (FP7-PEOPLE-2012-ITN).

References

[1] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao. Semantics+ filtering+ search= twitcident. exploring information in social web streams. In *Proceedings of the ACM HT 2012*.

[2] S. Adalı and J. Golbeck. Predicting personality with social behavior: a comparative study. *Social Network Analysis and Mining*, 4(1):1–20, 2014.

[3] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *ICWSM 2011*.

[4] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *Proceedings of the International Conference on WWW 2009*.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: Human, bot, or cyborg? In *Proceedings of the ACSAC 2010*.

[6] H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos. Identification of key locations based on online social network activity. In *Proceedings of the IEEE/ACM ASONAM 2015*.

[7] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of the International Conference on WWW 2013*.

[8] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proceedings of the ACM SIGKDD 2006*.

[9] S. Kumar, X. Hu, and H. Liu. A behavior analytics approach to identifying tweets from crisis regions. In *Proceedings of the ACM HT 2014*.

[10] S. Kumar, F. Morstatter, and H. Liu. *Twitter data analytics*. Springer, 2014.

[11] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the International Conference on WWW 2010*.

[12] A. Kyrola, G. Blelloch, and C. Guestrin. Graphchi: Large-scale graph computation on just a pc. In *USENIX Symposium on OSDI 2012*.

[13] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the International Conference on WWW 2008*.

[14] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM TKDD*, 1(1), Mar. 2007.

[15] Y.-R. Lin, B. Margolin, A. Keegan, A. Baronchelli, and D. Lazer. # bigbirds never die: Understanding social dynamics of emergent hashtag. In *ICWSM 2013*.

[16] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *ICWSM 2014*.

[17] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, et al. The arab spring| the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31, 2011.

[18] J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM Trans. Intell. Syst. Technol.*, 5(3):47:1–47:21, July 2014.

[19] S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[20] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist. Understanding the demographics of twitter users. In *ICWSM 2011*.

[21] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the ACM SIGCOMM Conference in IMC 2007*.

[22] J. S. Morgan, C. Lampe, and M. Z. Shafiq. Is news sharing on twitter ideologically biased? In *Proceedings of the CSCW 2013*.

[23] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the International Conference on WWW 2014*.

[24] S. A. Myers, A. Sharma, P. Gupta, and J. Lin. Information network or social network?: The structure of the twitter follow graph. In *Proceedings of the International Conference on WWW 2014 Companion*.

[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[26] M. J. Paul and M. Dredze. You are what you tweet: Analyzing twitter for public health. In *ICWSM 2011*.

[27] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the International Conference on WWW 2010*.

[28] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: An analysis of twitter spam. In *Proceedings of the ACM SIGCOMM Conference in IMC 2011*.

[29] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

[30] I. Weber, V. R. K. Garimella, and A. Batayneh. Secular vs. islamist polarization in egypt on twitter. In *Proceedings of the IEEE/ACM ASONAM 2013*.

[31] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann. Who, where, when and what: Discover spatio-temporal topics for twitter users. In *Proceedings of the ACM SIGKDD 2013*.