Athens University of Economics and Business

School of Business

Department of Management Science & Technology

Master of Science in Business Analytics

Program:	Full-time
Quarter:	3rd (Spring Quarter)
Course:	Social Network Analysis
Instructor:	Dr. Katia Papakonstantinopoulou
Assignment №:	Project 2
Students (Registration №):	Souflas Eleftherios - Efthymios (f2822217)

Table of Contents

Homework 2: From Raw Data to Temporal Graph Structure Exploration

Introduction

In this Homework, we used raw data from Twitter to create a weighted directed graph using R programming language and its *igraph* library. In fact, we created five graphs from the evolution of tweets posted during the first five days of July 2009 and then utilized them to firstly visualize the 5-day evolution of different graph metrics, then explore the most important users during these 5 days and finally perform community detection on the forementioned graphs.

Problem 1: Twitter mention graph

The aim of this problem was to analyze a Twitter dataset from July 2009 and create a weighted directed graph using the $igraph$ library in R. Additionally, the most important topic for each user was identified based on the hashtags used in their tweets.

The raw data ("tweets2009-07.txt") was stored in a compressed file, which was downloaded, decompressed, and manipulated using R programming language. The data was processed in chunks to optimize memory usage, because its size after decompression was approximately 6.8 Gigabytes. The raw data was in a specific format with T indicating the time the tweet was posted, U indicating the user who posted it, and W containing the text of the tweet. Each chunk was parsed to extract the relevant information, including the timestamp, user, and tweet text only for the first five days of July. Mentions (@) and hashtags (#) in the tweet's text were identified using regular expressions.

After processing the data, five separate CSV files were created, each representing one of the first five days in July 2009. The files followed the format "$from, to, weight$" to describe the weighted directed mention graph. Each mention between users was assigned a weight of 1, indicating the strength of the connection. The edges from a user to the same user (loop-edges), indicating self-mentioning tweets, were filtered out. The CSV files were named "edgelist_July_X.csv," where X represents the respective day.

To determine the most important topic for each user, the hashtags used in their tweets were analyzed. For each user, the hashtag with the highest frequency across all their tweets was considered their most important topic. In cases where multiple hashtags had the same frequency, the topic was chosen randomly (the first that was parsed was chosen – not sorted). Five separate CSV files were created, one for each day, following the format "$user, topic\_of\_interest$". The files were named "nodelist_July_X.csv," where X represents the respective day.

Please note that the execution time for processing the data and generating the CSV files lasted for approximately three quarters of an hour (45 minutes). However, this time might vary depending on the computing power of the operating system.

Problem 2: Average degree over time

Then, having created the forementioned graphs, we analyzed the 5-day evolution of various graph metrics. The metrics included the number of vertices, number of edges, diameter of the graph, average

in-degree, and average out-degree. By creating plots for each metric, we were able to observe the trends and fluctuations over the five days.

Firstly, we loaded the necessary libraries and created an empty list to store the graphs. Then, using a loop, we read the edge and node lists from the CSV files exported from the previous problem and created directed weighted graphs for each day. The number of vertices and edges were calculated for each graph and displayed as output to verify the correct creation of the graphs. The resulting graph objects were stored in the previously created list for further analysis.

Next, we focused on the 5-day evolution of various metrics over time. We counted the number of vertices and edges, measured the diameter of the graph, and calculated the in- and out-degree, by taking into consideration the weights of the edges in the calculation with the use of the $igraph's$ "strength" function. Since calculating the diameter of large graphs is computationally intensive, all graph metrics were previously computed and exported to a CSV file, named "metrics.csv". We read the metrics data from the file into a data frame for further analysis. The metrics included the number of vertices, number of edges, diameter, average in-degree, and average out-degree for each day. Using the $ggplot2$ library, we created plots to visualize the evolution of each metric. Here are the key observations for each metric:

1. **Number of Vertices**: There is a gradual decrease in the number of vertices from day 1 to day 5, indicating a reduction in the size of the graph over time (Figure 1).

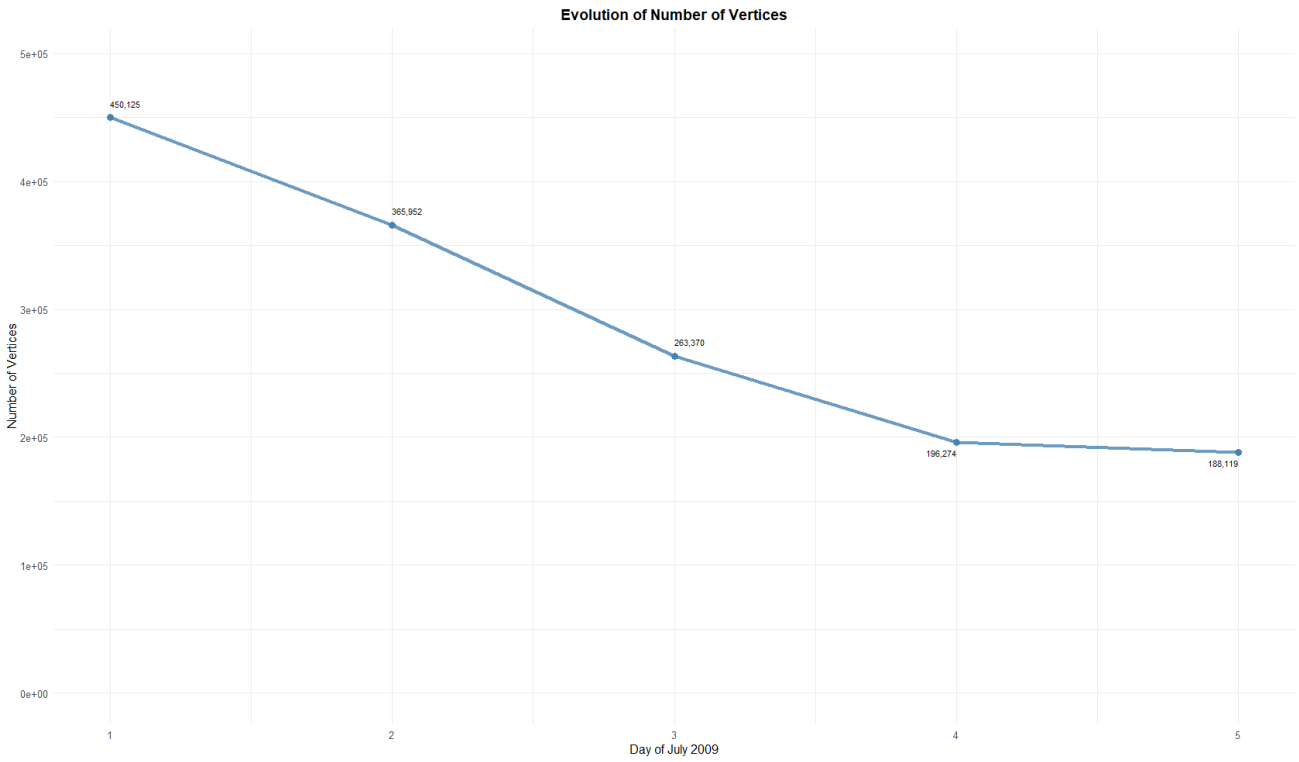

Figure 1 - Evolution of the number of vertices for the first 5 days of July 2009

2. **Number of Edges**: There is a gradual decrease in the number of edges across the five days, especially in the last two days, indicating less social interaction during that weekend (Figure 2).
3. **Diameter**: The diameter of the graph fluctuates during the five days. There is a significant decrease in diameter from day 1 to day 3, followed by a gradual increase over the remaining days, indicating a more connected or compact graph during July 3rd (Figure 3).

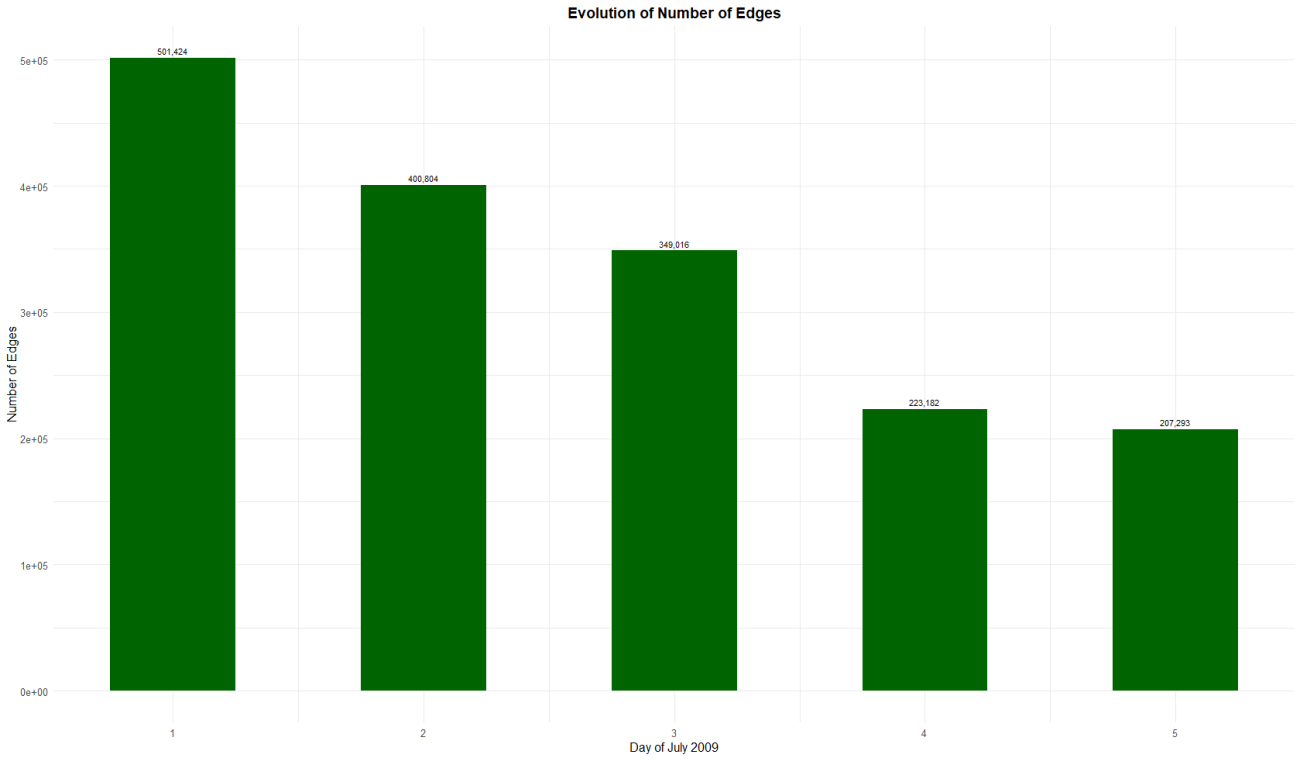

Figure 2 - Evolution of the number of edges for the first 5 days of July 2009

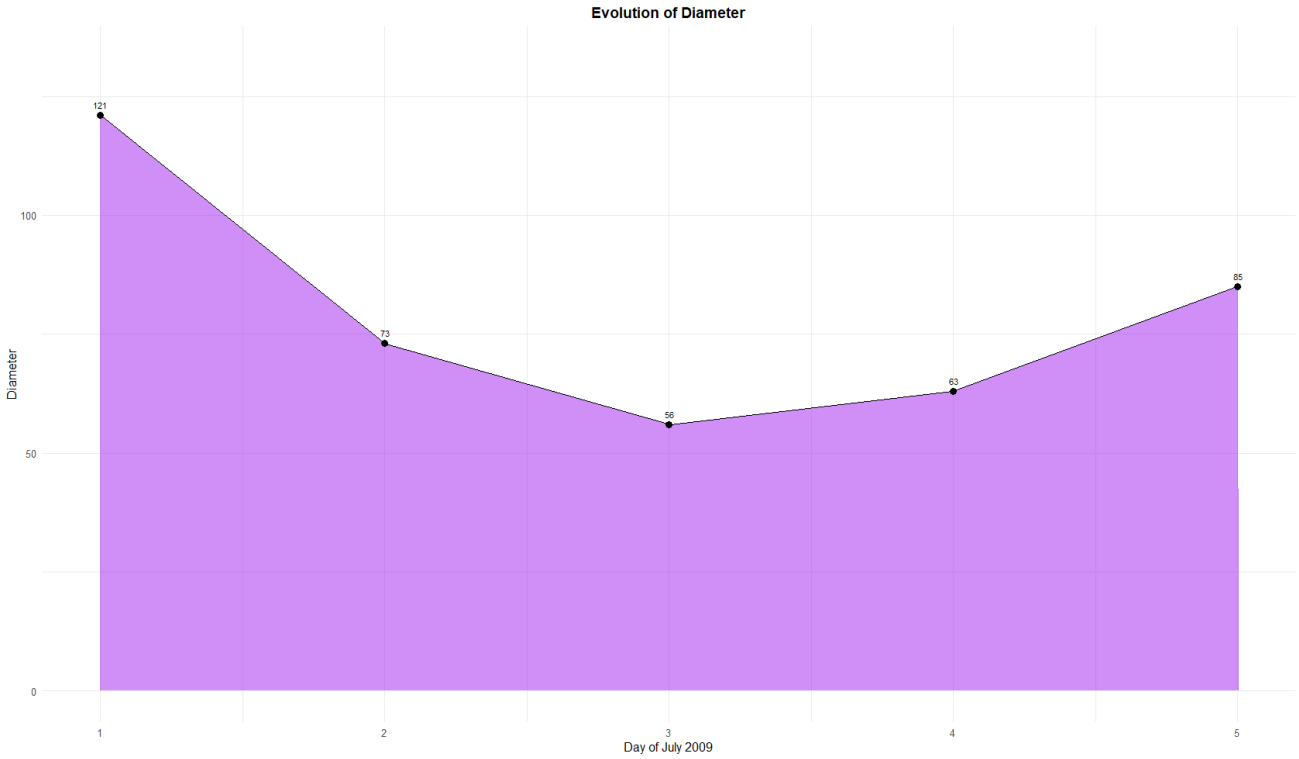

Figure 3 - Evolution of the graph's diameter for the first 5 days of July 2009

4. **Average In-degree**: The average in-degree remains relatively constant over time, showing only minor fluctuations and reaching its top value on July 3rd, indicating that certain individuals are becoming

slightly more influential or popular, receiving a slightly higher number of mentions during that day (Figure 4).

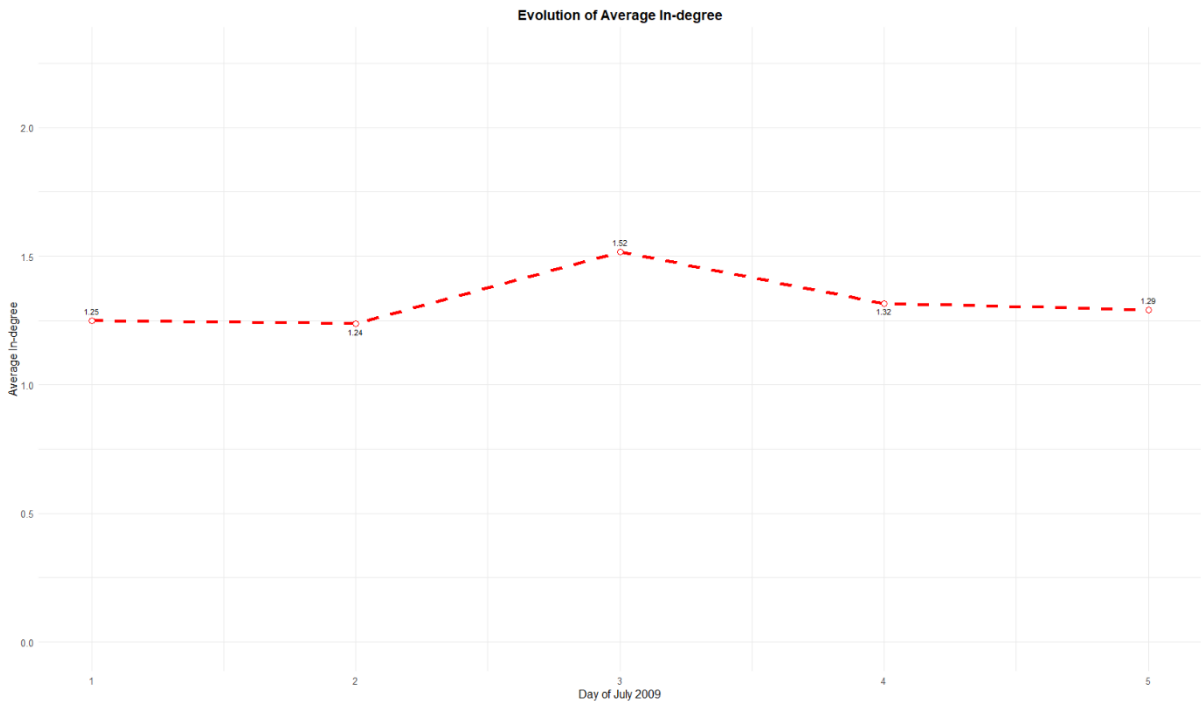

Figure 4 - Evolution of the graph's average in-degree for the first 5 days of July 2009

5. **Average Out-degree**: Similar to the average in-degree, the average out-degree also remains stable with minimal variations, reaching its top value in July 3rd, indicating that individuals are reaching out to slightly more people or forming a little greater number of connections with others, during this day, leading to a slightly higher level of activity or information dissemination (Figure 5).

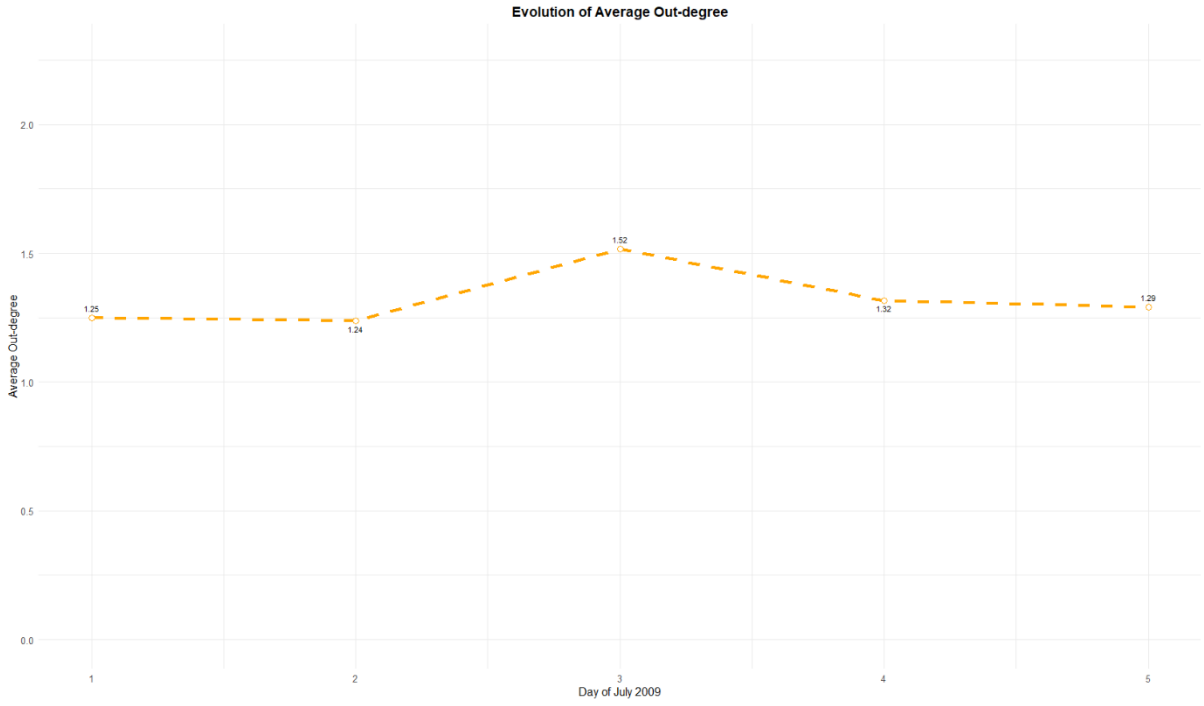

Figure 5 - Evolution of the graph's average out-degree for the first 5 days of July 2009

Note: Because for every edge in the network there is a sender and a receiver, any attempt to calculate the average in- or out-degree will result in the same answer as the average degree calculation (Lizardo & Jilbert, 2023) i.e., 1.52 is the average in- and out-degree simultaneously for day 3. That is because of the handshaking theorem which mentions that the total number of in-degree is equal to the total number of out-degree (Manwani, 2023). More specifically:

$$\sum_{u \in V} deg^{-}(u) = \sum_{u \in V} deg^{+}(u) = |E|, \text{ where}$$
$G = (V, E)$ a directed graph, u a vertex of the graph, $deg^{-}(u)$ its in-degree, and $deg^{+}(u)$ its out-degree

Thus, the plots of the forementioned graph metrics (average in- and out-degree) are identical and reveal the same trend (Figure 4 and Figure 5).

Problem 3: Important nodes

The next goal of our analysis was to examine the evolution of the top-10 Twitter users over a 5-day period in July 2009. Three network metrics were considered: in-degree, out-degree, and PageRank. We created three data frames to store the top-10 users for each metric on different days and looped through the five days to calculate the respective metrics which were then stored in each data frame's columns for each day's graph, taking into consideration the weights of the graph's edges. The top-10 users for each metric on each day were determined by sorting the users based on their corresponding scores (in-degree, out-degree, and PageRank) in descending order.

Here are the findings from the analysis of the Top-10 Twitter Users based on:

1. **In-Degree**: Across the five days, the user "*tweetmeme*" consistently had the highest in-degree, indicating a high number of incoming connections. Users like "*mashable*", "*addthis*", "*mileycyrus*", and "*breakingnews*" were also present in the top-10 on multiple days, suggesting a sustained level of attention from other users (Table 1).

	Day_1	Day_2	Day_3	Day_4	Day_5
1	*tweetmeme*	*officialtila*	*tweetmeme*	*tweetmeme*	*iamdiddy*
2	*mashable*	*ddlovato*	*souljaboytellem*	*songzyuuup*	*davidmmasters*
3	*smashingmag*	*tweetmeme*	*addthis*	*breakingnews*	*tweetmeme*
4	*addthis*	*mashable*	*mashable*	*addthis*	*addthis*
5	*mileycyrus*	*cnnbrk*	*breakingnews*	*souljaboytellem*	*breakingnews*
6	*breakingnews*	*cnn*	*moontweet*	*iamdiddy*	*mashable*
7	*aplusk*	*addthis*	*phillyd*	*mileycyrus*	*mileycyrus*
8	*guykawasaki*	*souljaboytellem*	*cnnbrk*	*mashable*	*moontweet*
9	*cnn*	*mileycyrus*	*officialtila*	*lilduval*	*akgovsarahpalin*
10	*cardoso*	*jeepersmedia*	*jeepersmedia*	*cnnbrk*	*imeem*

Table 1 - Top-10 Twitter Users based on In-Degree for the first 5 days of July 2009

2. **Out-Degree**: The users "*dudebrochill*" and "*wootboot*" appeared frequently in the top-10 out-degree list over the five days, indicating a high number of outgoing connections. However, we observed more variation in this list compared to the in-degree list, with different users dominating on different days (Table 2).

	Day_1	Day_2	Day_3	Day_4	Day_5
1	*teamqivana*	*dudebrochill*	*drejones71*	*swbot*	*swbot*
2	*dudebrochill*	*penishunter*	*deana1981*	*andreapuddu*	*twiprodigy009*
3	*failbus*	*wootboot*	*imbeeyo*	*dudebrochill*	*twiprodigy007*
4	*tsliquidators*	*failbus*	*killah360dhh*	*hoboprophet*	*twiprodigy008*
5	*the_sims_3*	*modelsupplies*	*java4two*	*itz_cookie*	*twiprodigy005*
6	*wootboot*	*the_sims_3*	*nachhi*	*wootboot*	*wildingp*
7	*vaguetweetstest*	*thickdecadence*	*andreapuddu*	*dj_fresh*	*bilbo232*
8	*jamokie*	*dvdbot*	*thickdecadence*	*azandiamjbb*	*apeeescape*
9	*juliesearser*	*jamokie*	*ohmichael*	*nachhi*	*hoboprophet*
10	*drharvey*	*takeyourpin*	*beatbean_*	*modelsupplies*	*dudebrochill*

Table 2 - Top-10 Twitter Users based on Out-Degree for the first 5 days of July 2009

3. **PageRank**: "*tweetmeme*" again appeared consistently in the top-10 users based on PageRank score, indicating a high influence in the network. "*mashable*" and "*addthis*" also showed notable influence, being present in the top 10 across several days (Table 3).

	Day_1	Day_2	Day_3	Day_4	Day_5
1	*tweetmeme*	*ddlovato*	*tweetmeme*	*addthis*	*davidmmasters*
2	*mashable*	*mashable*	*killerstartups*	*breakingnews*	*iamdiddy*
3	*addthis*	*drew_taubenfeld*	*souljaboytellem*	*tweetmeme*	*aplusk*
4	*smashingmag*	*tweetmeme*	*addthis*	*mashable*	*addthis*
5	*liambowers*	*globalmanners*	*moontweet*	*mileycyrus*	*tweetmeme*
6	*kissmetrics*	*cnn*	*mashable*	*songzyuuup*	*mashable*
7	*cnn*	*souljaboytellem*	*cnnbrk*	*iamdiddy*	*moontweet*
8	*bofu2u*	*addthis*	*breakingnews*	*cnnbrk*	*mrskutcher*
9	*mileycyrus*	*cnnbrk*	*phillyd*	*souljaboytellem*	*breakingnews*
10	*couragecampaign*	*officialtila*	*adamlambert*	*lilduval*	*mileycyrus*

Table 3 - Top-10 Twitter Users based on PageRank for the first 5 days of July 2009

We noticed variations in the top-10 lists for the different days, particularly in the out-degree metric. While some users demonstrated consistent influence over the entire period, others appeared more sporadically. This variability could be influenced by trending topics, the timing of user activity, or specific events that led to increased interactions with certain users on specific days.

Problem 4: Communities

Finally, we focused on performing community detection on the mention graphs. Three different algorithms, namely *fast greedy*, *infomap*, and *Louvain* clustering, were applied to the undirected versions of the five mention graphs. The performance of the three algorithms was assessed based on the modularity metric.

The *fast greedy* clustering algorithm achieved, during the 5-day period, a mean modularity of approximately 0.875, indicating a high quality of clustering. However, the execution time was relatively long, taking close to an hour (approximately 58 minutes) to complete.

The *infomap* clustering algorithm was applied with two different numbers of trials (number of attempts to partition the network). The mean modularity achieved with one trial was 0.774634, and it slightly increased to

0.7746889 with two trials (minor increase of 0.0000549). However, further increasing the number of trials did not lead to significant improvements in the clustering quality. On the contrary, it substantially increased the runtime of the algorithm. Therefore, it was concluded that using more than one trial reduced the efficiency of the algorithm. The execution time for two trials was close to 7 hours (approximately 6.9 hours), while only 26 minutes were spent for one trial, resulting in a similar quality of clustering solution.

The $Louvain$ clustering algorithm achieved a mean modularity of approximately 0.881, which was slightly better than the $fast\ greedy$ algorithm (increase of 0.0053). Also, it had a significantly faster execution time of approximately 15 seconds, making it the most efficient algorithm among the three.

To analyze the evolution of communities for a random user present in all five graphs, we applied a code to randomly select a user from a pool of common users among these days. The most frequent topic of interest (if there existed one) for this user was returned for each of these days. Across all code test runs that we made, the users selected at each run of the code belonged to communities that did not have the same set of users and differed in each consecutive day. There was no continuity in the user's community membership throughout the five-day period. However, there were some important, in terms of PageRank score, users that were grouped together with our selected users across more than one consecutive day.

Also, the communities that the user was partitioned into had distinct sets of important topics. The top 10 most frequent topics in each community for each day were returned. The communities had varied topics of interest, indicating a lack of topic consistency for the user. The common topics across all 5-day period were also returned.

For example, for the user "$fabioricotta$" his topics of interest were $\#seo$ for the first day; no topic was mentioned for the second day; $\#google\_street\_view$ for the third day; $\#design$ for the fourth one; and $\#guanacast$ for the fifth one. He was partitioned at community number 7 in the first day; number 3 in the second and third day; number 44 in the fourth day; and at community number 461 in the fifth day. He was partitioned together with the following important users (in terms of PageRank score):

- "$rafinhabastos$", "$marcelotas$", and "$cardoso$" (Day 1 – Day 4) across four consecutive days.
- "$huckluciano$" (Day 2 – Day 4) across three consecutive days.
- "$manomenezes$", "$danilogentili$", "$bgagliasso$", "$ticostacruz$" (Day 1 – Day 2), "$rubarrichello$" (Day 2 – Day 3) and "$kibeloco$", "$millorfernandes$" (Day 3 – Day 4) across two consecutive days.
- In day 5, he was partitioned in a community that hadn't in common any important user with the previous day.

The previous analysis was done because usually the important users determine the most frequent (important) topics of interest for the community. Then, we extracted the 10 most important topics of interest of the user's community for each day to check if the communities shared any topic. From the analysis, we found out that there existed two topics ($\#moonfruit$ and $\#ff$) that were common amongst all communities. In particular:

- $\#moonfruit$ and $\#ff$ were common topics of interest across all five communities, but they were generally famous topics across many other communities and users during this period.
- $\#iranelection$ was another common topic of interest found across four communities that the user was partitioned into each day, but it was also a famous topic during these days, due to Iranian Presidential Election on June 12th, 2009.
- $\#tcot$, $\#fb$, $\#spymaster$, $\#followfriday$, and $\#tinychat$ were common topics of interest found also across four communities that the user was partitioned into each day.
- $\#tweetmyjobs$ and $\#lastfm$ were common topics of interest that were shared across three communities.

At last, to visualize the graphs, a different color was assigned to each community. Additionally, to create a meaningful and aesthetically pleasing visualization, nodes belonging to very small or very large communities were filtered out. For each day, the vertices were colored based on their community membership, and edges were displayed as solid lines if they crossed between communities and as dotted lines otherwise. Mid-size communities, defined as those with sizes between 50 and 300, were selected for visualization.

The resulting plots showcased the community structure of the graphs for each day, providing visual insights into the network's organization and community relationships. In the figures below (Figure 6, Figure 7, Figure 8, Figure 9, and Figure 10), we can observe the 5-day evolution of Twitter network's mid-sized communities.

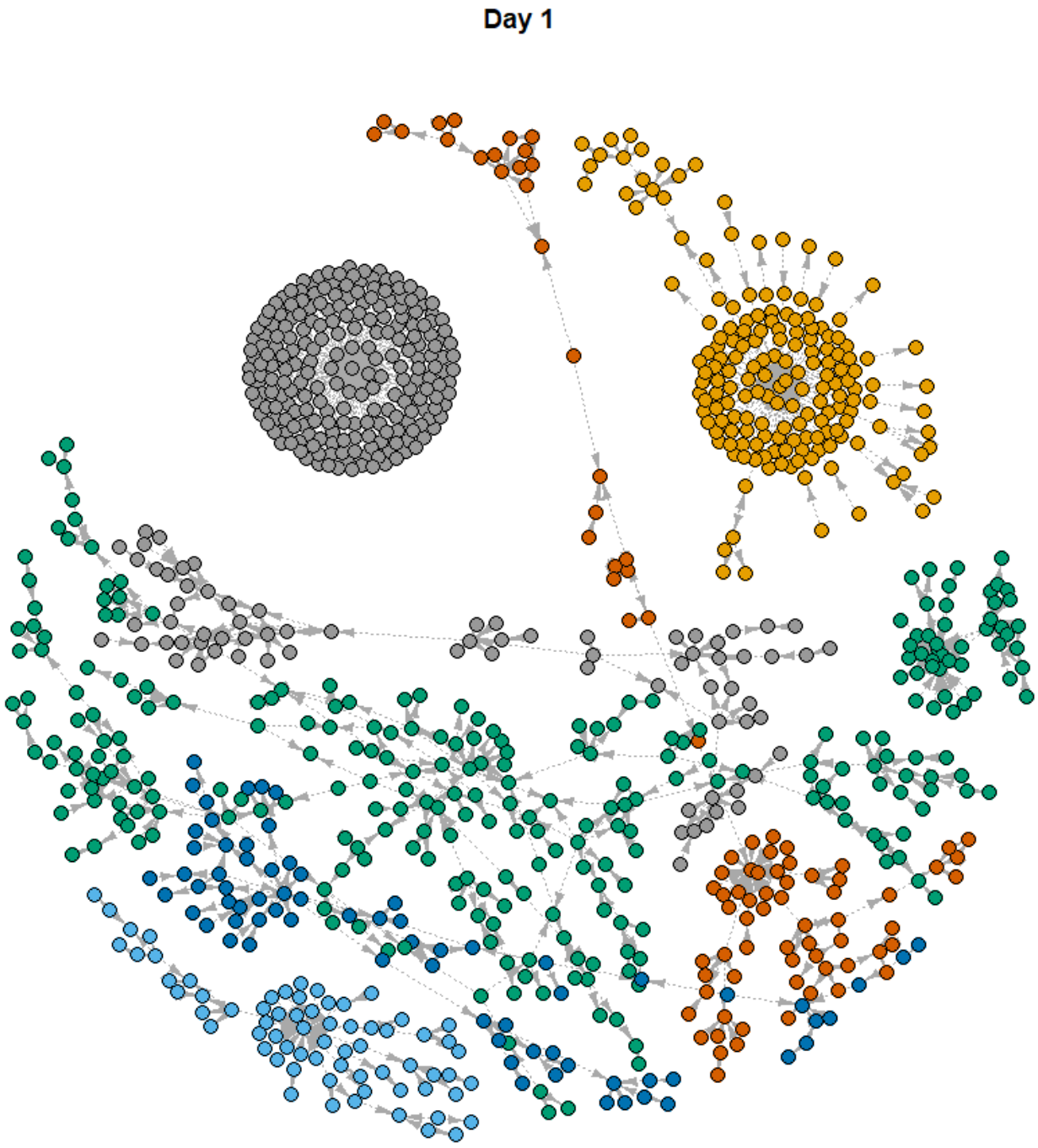

Figure 6 - Day 1 mid-sized communities

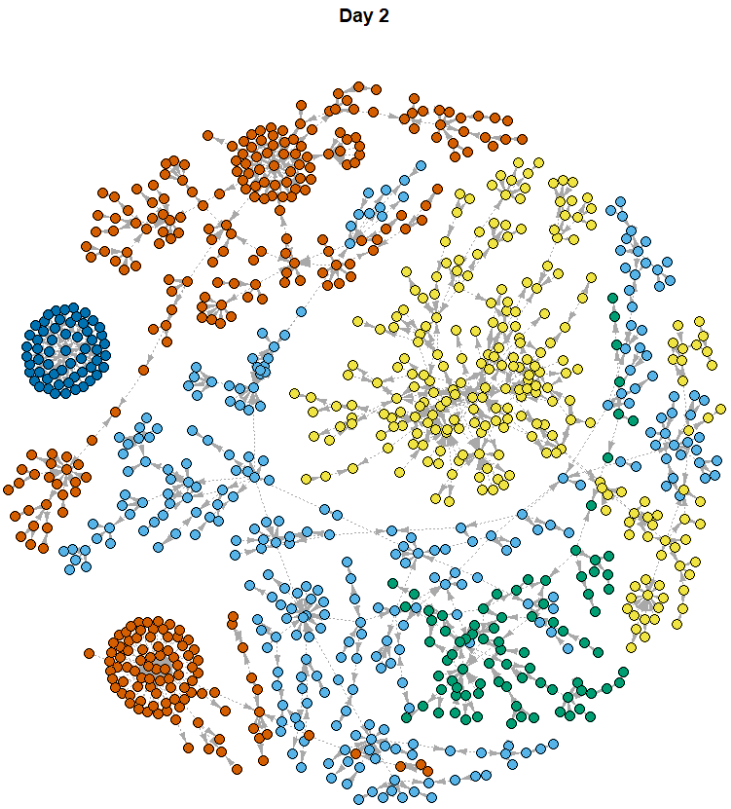

Figure 7 - Day 2 mid-sized communities

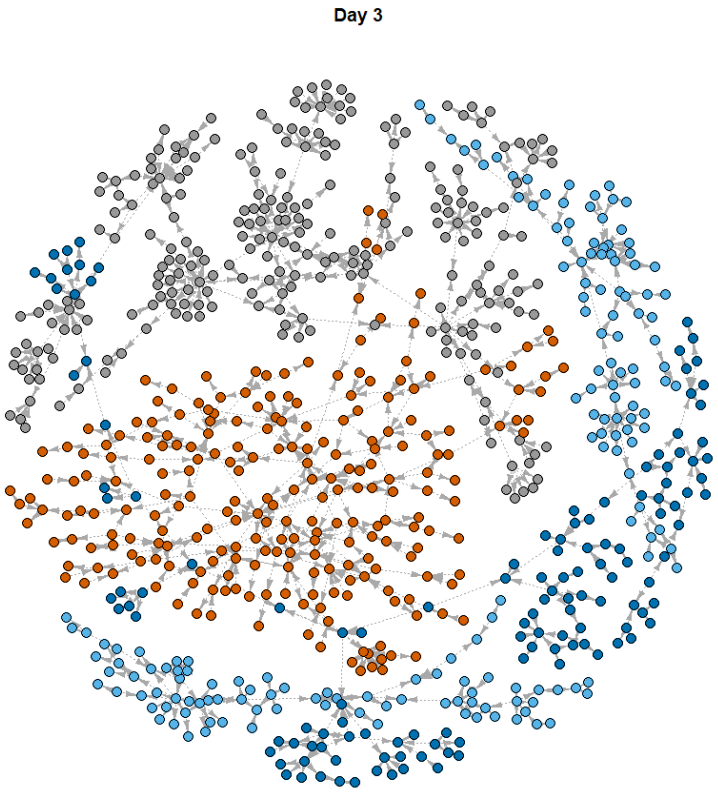

Figure 8 - Day 3 mid-sized communities

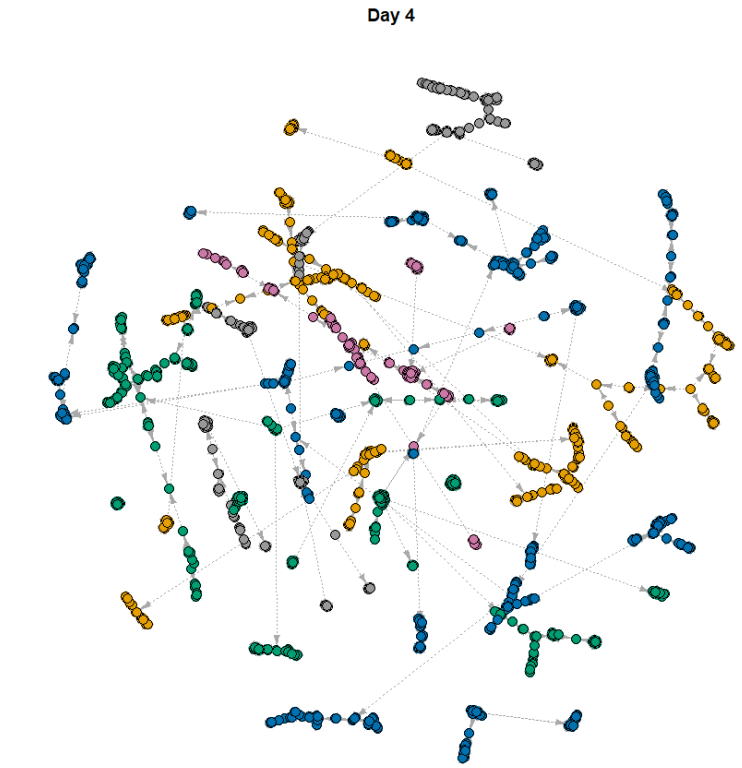

Figure 9 - Day 4 mid-sized communities

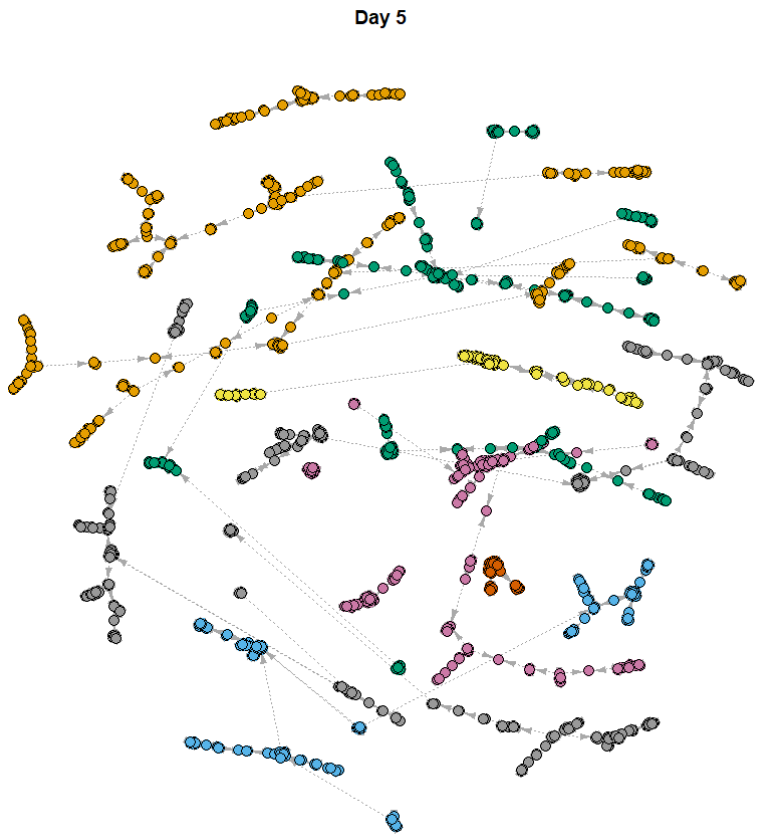

Figure 10 - Day 5 mid-sized communities

References

Lizardo, O., & Jilbert, I. (2023, July 3). *Social Networks: An Introduction*. Retrieved from bookdown.org: https://bookdown.org/omarlizardo/_main/2-7-average-degree.html

Manwani, C. (2023, July 3). *Mathematics | Graph Theory Basics – Set 2*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/mathematics-graph-theory-basics/