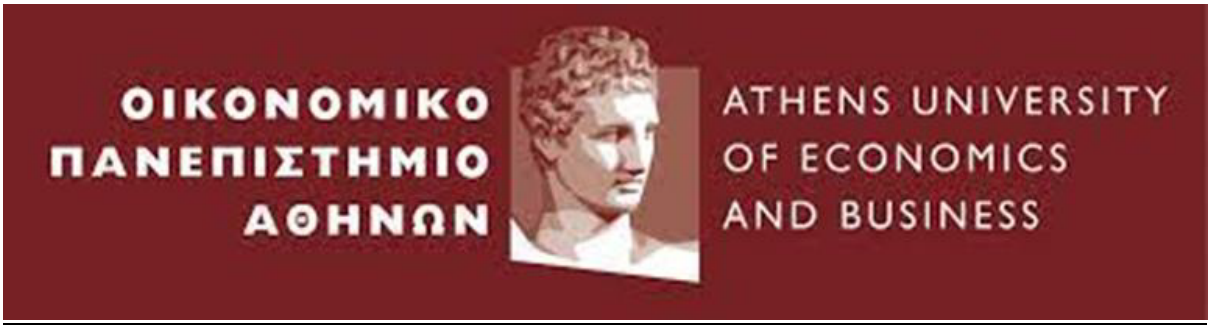


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
DEPARTMENT OF MANAGEMENT, SCIENCE & TECHNOLOGY
MSc BUSINESS ANALYTICS

"From raw data to temporal graph structure exploration"

Course: Social Network Analysis - PT

Full Name: ATHANASIOS ALEXANDRIS
Register Number: p2822202

Table of Contents

1. Task 1 – Twitter mention graph

In the first task of this assignment, we were given raw data from Twitter of July 2009, and we had to manipulate them in order to create csv files for each one of the first five days of July 2009. We had to create two kind of csv files. The one kind describes the flow of the tweets and how many times has a user mentioned another one (from, to, weight), and the second contains for each one of the users that posted each day defined as the mostly-used hashtag (#) for each user, across all his/her tweets (user, topic_of_interest).

Creating the files with Python

For the creation of the 10 csv files described above, we used Python programming language, that helped us to extract the files in the proper format.

Our code snippet processes a collection of tweet data stored in a file. It extracts mentioned users and hashtags from the tweet text and performs various calculations and aggregations on the data.

First, the code defines two functions to extract mentioned users and hashtags from the tweet text using regular expressions. It then reads and processes the tweet data from a file. For each tweet, it extracts the timestamp, username, and tweet text. It keeps track of unique users for each day and counts the number of mentions between users and the frequency of hashtags.

The code then writes the mentioned counts to separate CSV files for each date, where each row represents a mention between two users along with the weight of the mention. It also writes the top topics for each user to separate CSV files for each date. If a user has multiple top topics, one is randomly selected as their topic of interest.

Creating igraph graphs in R

Having created the mentioned csv files, we loaded them in R in order to create the respective igraph graphs.

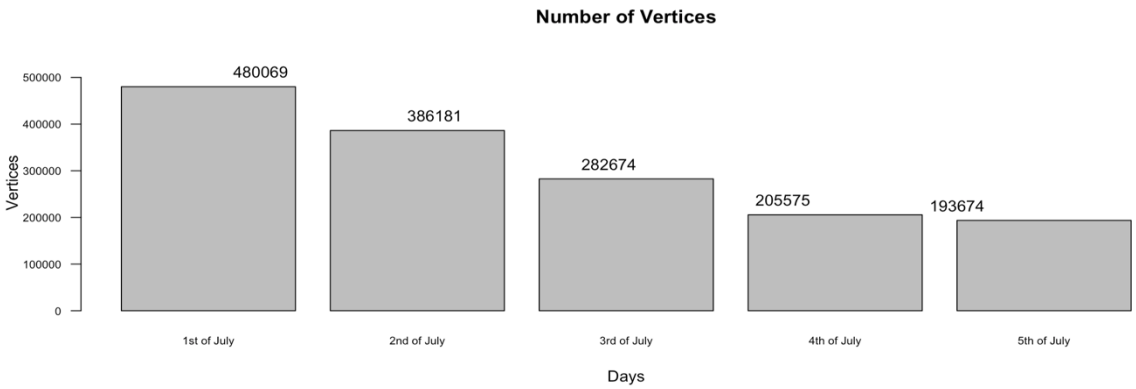
We first used the mentions .csv files to create 5 directed graphs (one for each day), using graph_from_data_frame function.

Then we added the topics to each one of the nodes (users) of our graphs.

2. Task 2 – Average degree over time

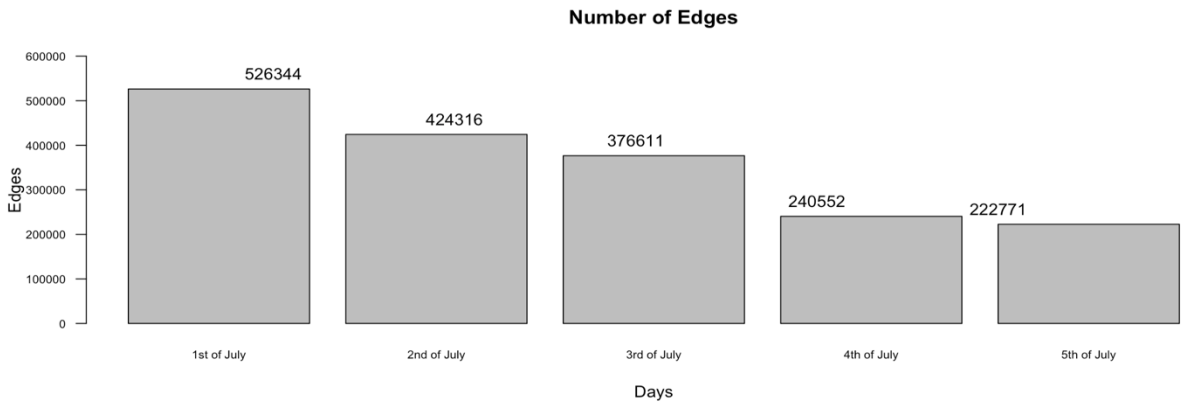
In the second task of our assignment we had to calculate various metrics of our graphs, and plot their 5-day evolution based on the below metrics.

• Number of vertices



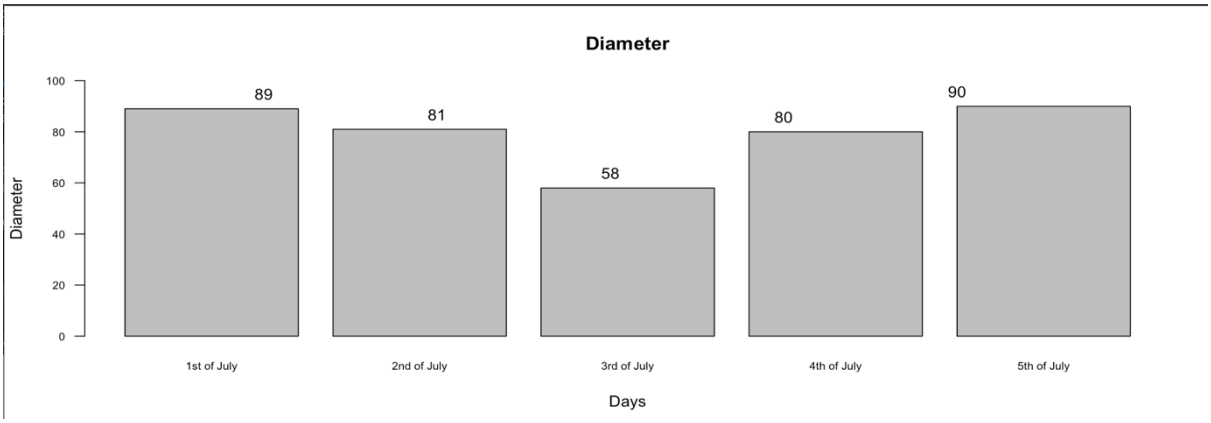
As we observe the number of vertices is higher in the 1st of July graph, with 480069 vertices, and then every day the number of vertices of graphs is being reduced with high flactuation. 4th and 5th of July graphs have the fewer number of vertices with small difference.

• Number of edges



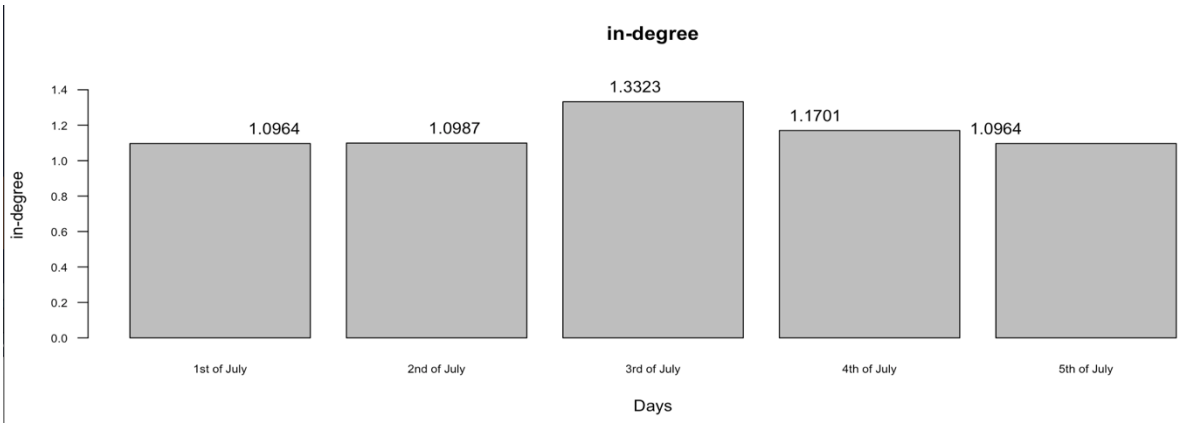
Similarly, the number of edges is higher in the 1st of July graph, with 526344 edges, and then every day the number of edges of graphs is being reduced with high flactuation. 4th and 5th of July graphs have the fewer number of edges with small difference.

• Diameter of the graph



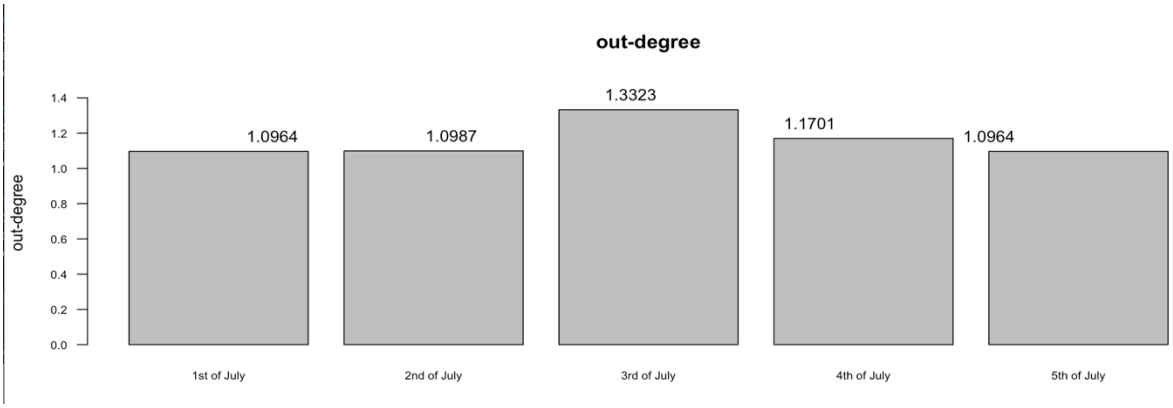
Regarding diameter metric, we observe that fifth day of July has higher value, with only one point difference from the first day of July. The lowest value with big difference from the others, is observed in third of July.

• Average in-degree



Regarding the average in-degree metric, we observe that 3rd of July has the highest value with 1.33 and then 4th of July with 1.17. Then follows the other three days with minus differences between them.

• Average out-degree



Exactly the same view we have for average out-degree metric, since in a graph, the number of in-degree is equal to outdegree.

It is important to mention that different metrics , provide different views regarding the importance of the graphs.

3. Task 3 – Important nodes

In this task we had to identify the top-10 Twitter users of each one of the 5 days regarding In-degree, Out-degree and PageRank metrics.

• In-degree

	Node_01.07	Node_02.07	Node_03.07	Node_04.07	Node_05.07
1	tweetmeme	tweetmeme	tweetmeme	BreakingNews	davidmmasters
2	mashable	ddlovato	souljaboytellem	addthis	iamdiddy
3	addthis	mashable	addthis	tweetmeme	addthis
4	smashingmag	cnnbrk	mashable	iamdiddy	tweetmeme
5	mileycyrus	cnn	BreakingNews	mileycyrus	mashable
6	BreakingNews	addthis	cnnbrk	cnnbrk	BreakingNews
7	cnn	souljaboytellem	moontweet	mashable	moontweet
8	GuyKawasaki	OfficialTila	lilduval	lilduval	mileycyrus
9	aplusk	officialtila	PhillyD	souljaboytellem	rainnwilson
10	rafinhabastos	mileycyrus	adamlambert	TheOnion	AKGovSarahPalin

As we see the most important user for the first three days is 'tweetmeme' that is ranked in the third place the fourth day and in the fourth place the fifth day. We observe that the 5 days graphs have many common nodes in their top 10 with the other days graphs, like 'tweetmem', 'BreakingNews', 'mileycyrus', 'addthis' etc, means that same users tend to be the most mentioned users in twitter those 5 days.

• Out-degree

	Node_01.07	Node_02.07	Node_03.07	Node_04.07	Node_05.07
1	dudebrochill	dudebrochill	drejones71	swbot	swbot
2	failbus	wootboot	deana1981	dudebrochill	twiprodigy008
3	tsliquidators	failbus	killah360dhh	wootboot	twiprodigy005
4	the_sims_3	the_sims_3	imbeeyo	fxxxyourlife	twiprodigy007
5	wootboot	dvdbot	java4two	andreapuddu	twiprodigy009
6	vaguetweetstest	takeyourpin	ohmichael	azandiamjbb	wildingp
7	lmaobot	teamqivana	nachhi	hoboprophet	dudebrochill
8	drharvey	luvorhate	dudebrochill	failbus	wootboot
9	luvorhate	modelsupplies	wootboot	herpescure	hoboprophet
10	help_echo	rt_thursday	medic_ray	twiprodigy009	the_sims_3

As we see the is 'dudebrochill', and 'swbot' have been placed two out of five times in the first place. User 'dudebrochill' is important to mention that is placed in the top 10 all 5 days. We observe that the 5 days graphs have some common nodes in their top 10 with the other days graphs, like 'dudebrochill', 'mileycyrus', 'wootboot etc. However the common users between the five days, are less fewer than in in-degree metric.

• **PageRank**

	Node_01.07	Node_02.07	Node_03.07	Node_04.07	Node_05.07
1	tweetmeme	ddlovato	tweetmeme	souljaboytellem	davidmmasters
2	mashable	drew_taubenfeld	souljaboytellem	addthis	iamdiddy
3	addthis	mashable	killerstartups	tweetmeme	addthis
4	smashingmag	tweetmeme	addthis	BreakingNews	aplusk
5	cnn	globalmanners	moontweet	lilduval	tweetmeme
6	mileycyrus	cnn	cnnbrk	mileycyrus	mashable
7	KISSmetrics	addthis	mashable	mashable	mrskutcher
8	CourageCampaign	souljaboytellem	BreakingNews	iamdiddy	moontweet
9	aplusk	cnnbrk	PhillyD	cnnbrk	BreakingNews
10	rafinhabastos	mileycyrus	adamlambert	garyvee	mileycyrus

As we see the most important users placed in the firs places of each day are ‘tweetmeme’, ‘ddlovato’, ‘souljaboytellem’ and ‘davidmasters’. We observe that the 5 days graphs have many common nodes in their top 10 with the other days graphs, like ‘tweetmem’, ‘BreakingNews’, ‘mileycyrus’, ‘addthis’ etc, means that same users tend to be the most mentioned users by the most important ones.

4. Task 4 – Communities

In this task we first had to perform community detection on the 5 mention graphs, by applying the methods fast greedy clustering, infomap clustering, and louvain clustering on the undirected versions of the 5 mention graphs. After executing the commands multiple times, only the Louvain clustering methods were able to run, as the others were running for hours and we had no results.

Then we had to pick a random user that has present to all five graphs, detect the evolution of the communities this user belongs to.

We found the user "JhonenV" that belongs to all five graphs. Then we found the communities of each graph that this user belongs to. His communities id's for each day are listed below.

```
Day 1 Community_id: 74
Day 2 Community_id: 25
Day 3 Community_id: 28
Day 4 Community_id: 63
Day 5 Community_id: 26
```

The community of day1 had 913 users, day2 1066 users, day3 538 users, day4 522 users and day5 856 users.

Then we calculated the common users between the communities. The most common users are observed between fourth and fifth day (8 users).

```
> print(num_common_users)
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    2    1    3    2
[2,]    2    0    1    2    3
[3,]    1    1    0    1    3
[4,]    3    2    1    0    8
[5,]    2    3    3    8    0
```

Then we found the most frequent topics of the users of the communities, that are presented below for each on of the five days.

```
[1] "#tcot"
[1] "#moonfruit"
[1] "#urwashed"
[1] "#moonfruit"
[1] "#poke"
```

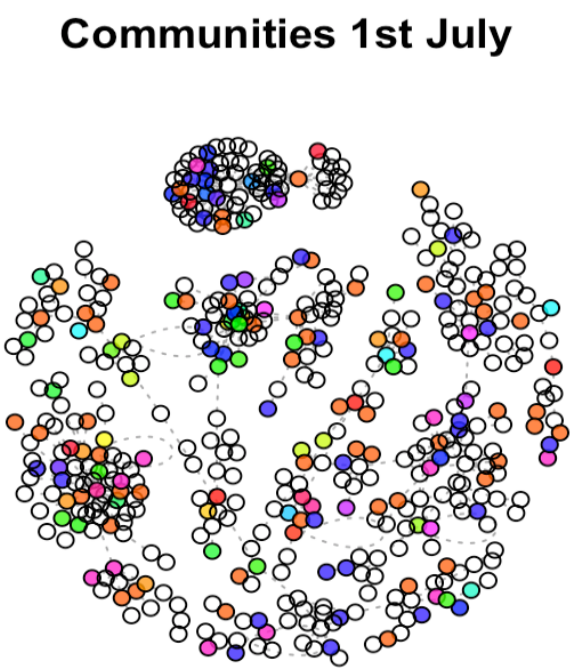
Also we found the number of common topics of the communities. Most common topics are in days 1 and 2 (26 common topics).

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    0   26   17   18   21
[2,]   26    0   25   19   25
[3,]   17   25    0   19   19
[4,]   18   19   19    0   29
[5,]   21   25   19   29    0
```

Finally we created plots for each one of the 5 days, keeping only the most important communities in order to create meaningful and aesthetically pleasing visualizations.

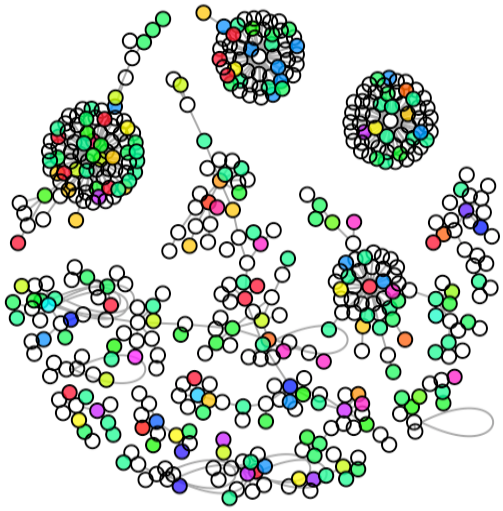
Visualizations for each day are presented below:

1st July



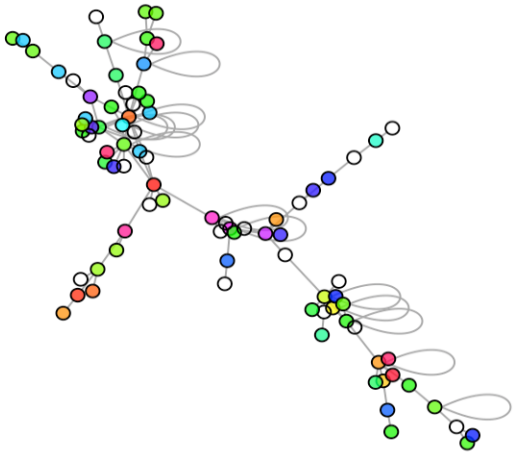
2nd July

Communities 2nd July

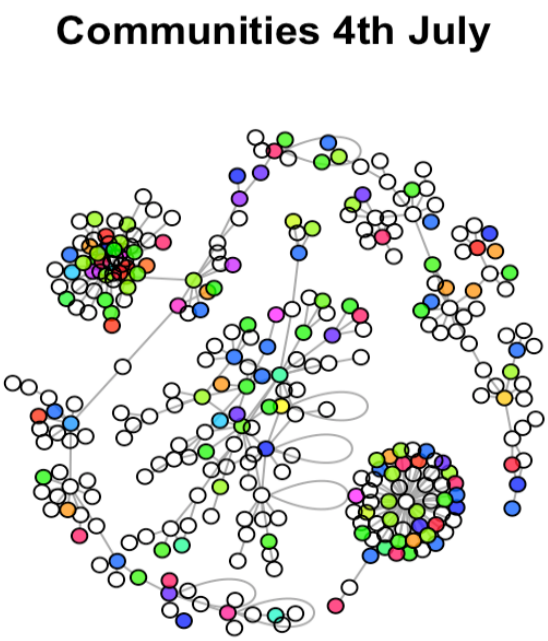


3rd July

Communities 3rd July



4th July



5th July