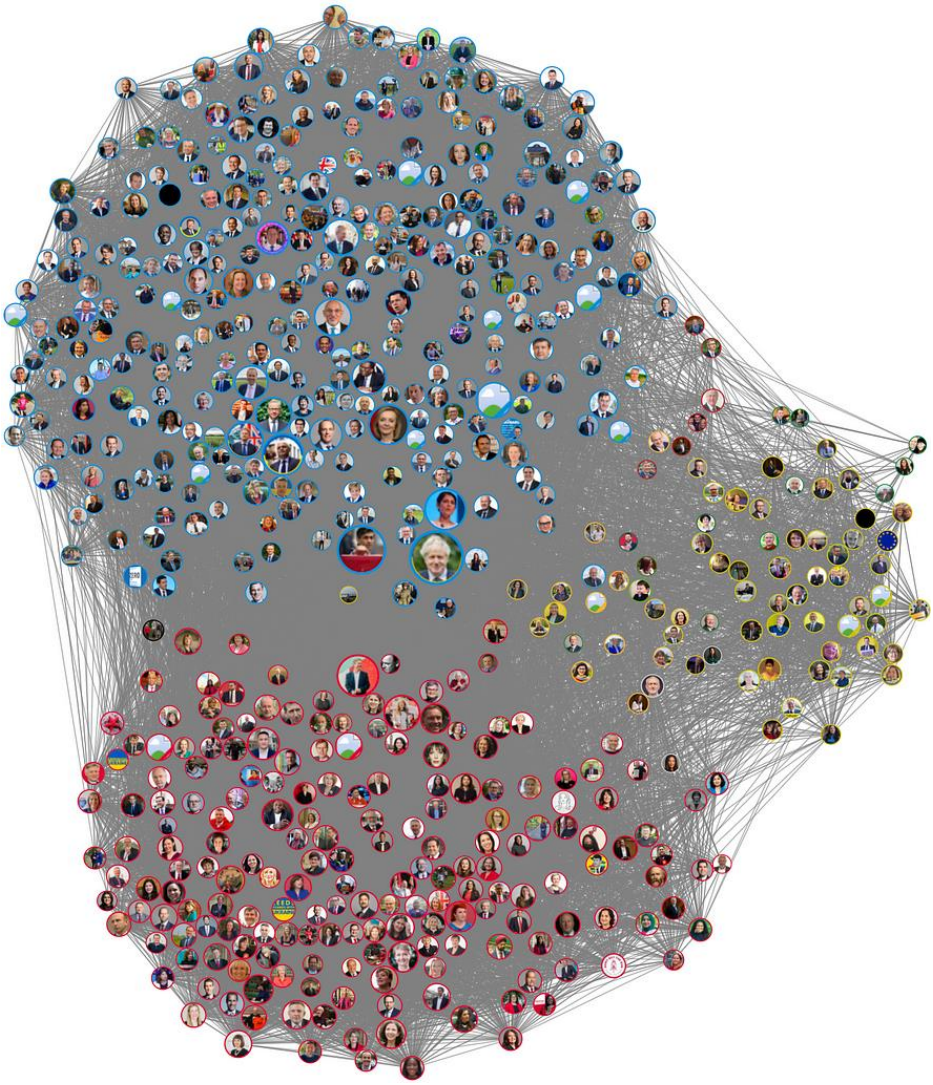


From raw data to temporal graph structure exploration

HOMEWORK 2

SOCIAL NETWORK ANALYSIS
MSc in Business Analytics Part Time (2022-2024)
Athens University of Economics and Business
Instructor: Dr. Katia Papakonstantinopoulou

08/07/2023

Vaidomarkakis Panagiotis | p2822203

Table of Contents

Introduction

In this assignment, I am tasked with analyzing Twitter data from July 2009 to gain insights into the dynamics of the Twitter network during that time. The data consists of tweets posted in July 2009 and includes information such as timestamps, users, and tweet content. My goal is to create a weighted directed graph representing the mention relationships between users, identify the most important topic for each user based on their hashtags, and analyze the evolution of various graph metrics over five days.

Firstly, I will process the raw data and create five separate CSV files, each representing a day in July 2009. These files will contain the weighted mention graph, indicating the source user, target user, and the weight of the mention. Additionally, I will identify the most important topic for each user by analyzing the hashtags used in their tweets.

Using the generated CSV files, I will create igraph graphs in R and update the graph vertices to include the users' most important topics as attributes. This will enable further analysis and visualization of the network. I will then plot the evolution of metrics such as the number of vertices, number of edges, graph diameter, average in-degree, and average out-degree to identify any significant fluctuations or trends in the network's characteristics over the five-day period.

In addition, I will analyze the top 10 Twitter users each day based on their in-degree, out-degree, and PageRank scores. This analysis will provide insights into the most influential and active users during different time periods. Furthermore, I will perform community detection using three algorithms: fast greedy clustering, infomap clustering, and Louvain clustering. By examining the communities within the mention graphs, I can identify shared interests and patterns. Finally, I will visualize the graph, coloring each community differently, and filter out nodes from very small or large communities for a more meaningful representation.

Overall, this analysis aims to provide a comprehensive understanding of the Twitter network in July 2009 by examining mention relationships, identifying important topics, analyzing graph metrics, and detecting communities.

1) Twitter mention graph

ANALYSIS

All the analysis used to extract both edgelist and topic_of_interest for each of the 5 days is located in the .ipynb file which is located inside the zip with comments about want the file is doing in each step so no further comment will be written here. I tried to follow all the instructions to extract the right info. From my point of view, I have decided to exclude all hashtags that contains less than 3 numbers because they don't seem to mean something. For example, #1 can mean anything and it is something very common. 4 or more digits are acceptable, because of year dates (for example #1995). I kept all hashtags that have letters or numbers, for example #1music is an acceptable hashtag. All the other extracts are based on the instructions.

IGRAPH

We will move on to the igraph library in R which we started creating and analyzing the graphs.

First, we will read the edgelist files.

df1 = read.csv("CSV Files/edgelist_2009_07_01.csv")

```
df2 = read.csv("CSV Files/edgelist_2009_07_02.csv")
df3 = read.csv("CSV Files/edgelist_2009_07_03.csv")
df4 = read.csv("CSV Files/edgelist_2009_07_04.csv")
df5 = read.csv("CSV Files/edgelist_2009_07_05.csv")
```

Then the topic_of_interest files.

```
dftopic1 = read.csv("CSV Files/topic_of_interest_2009_07_01.csv")
dftopic2 = read.csv("CSV Files/topic_of_interest_2009_07_01.csv")
dftopic3 = read.csv("CSV Files/topic_of_interest_2009_07_01.csv")
dftopic4 = read.csv("CSV Files/topic_of_interest_2009_07_01.csv")
dftopic5 = read.csv("CSV Files/topic_of_interest_2009_07_01.csv")
```

Now, we will create the graphs.

```
dfedgelist1 = df1[, c("from","to","weight")]
g1 <- graph_from_data_frame(dfedgelist1[,c("from","to","weight")], directed = TRUE)
dfedgelist2 = df2[, c("from","to","weight")]
g2 <- graph_from_data_frame(dfedgelist2[,c("from","to","weight")], directed = TRUE)
dfedgelist3 = df3[, c("from","to","weight")]
g3 <- graph_from_data_frame(dfedgelist3[,c("from","to","weight")], directed = TRUE)
dfedgelist4 = df4[, c("from","to","weight")]
g4 <- graph_from_data_frame(dfedgelist4[,c("from","to","weight")], directed = TRUE)
dfedgelist5 = df5[, c("from","to","weight")]
g5 <- graph_from_data_frame(dfedgelist5[,c("from","to","weight")], directed = TRUE)
```

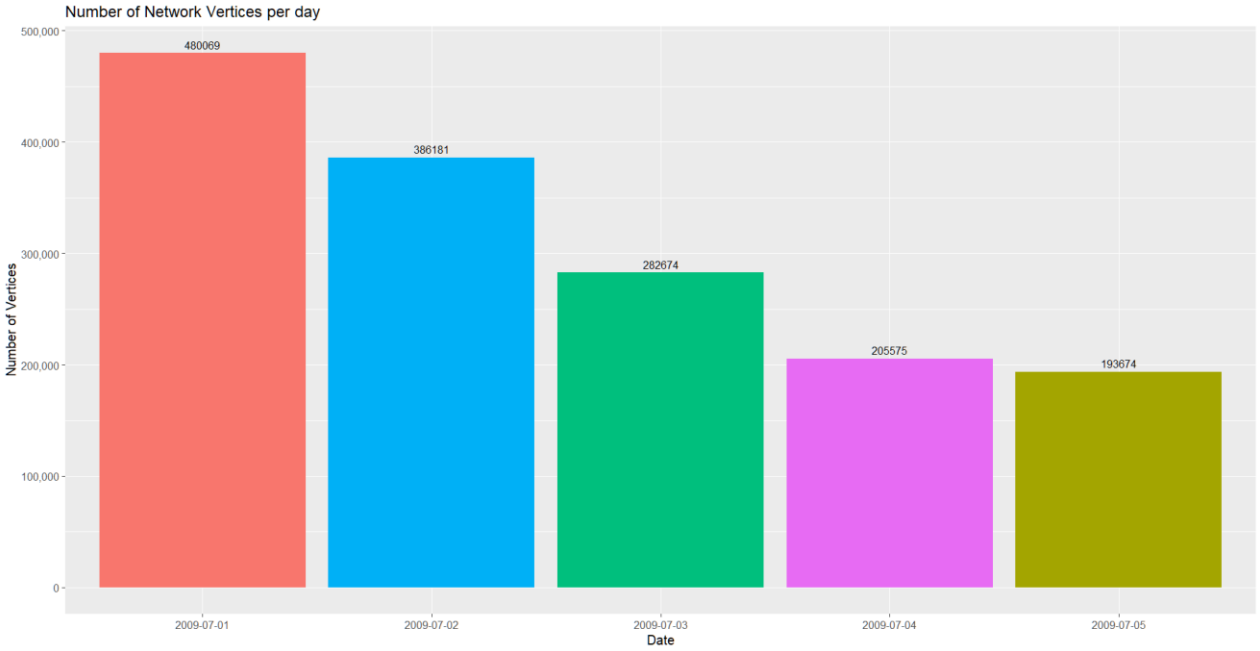
We inject now the subject of interest for each user for each day.

```
V(g1)$topic_of_interest <- dftopic1$topic_of_interest[match(V(g1)$name, dftopic1$user)]
V(g2)$topic_of_interest <- dftopic2$topic_of_interest[match(V(g2)$name, dftopic2$user)]
V(g3)$topic_of_interest <- dftopic3$topic_of_interest[match(V(g3)$name, dftopic3$user)]
V(g4)$topic_of_interest <- dftopic4$topic_of_interest[match(V(g4)$name, dftopic4$user)]
V(g5)$topic_of_interest <- dftopic5$topic_of_interest[match(V(g5)$name, dftopic5$user)]
```

2) Average degree over time

VERTICES

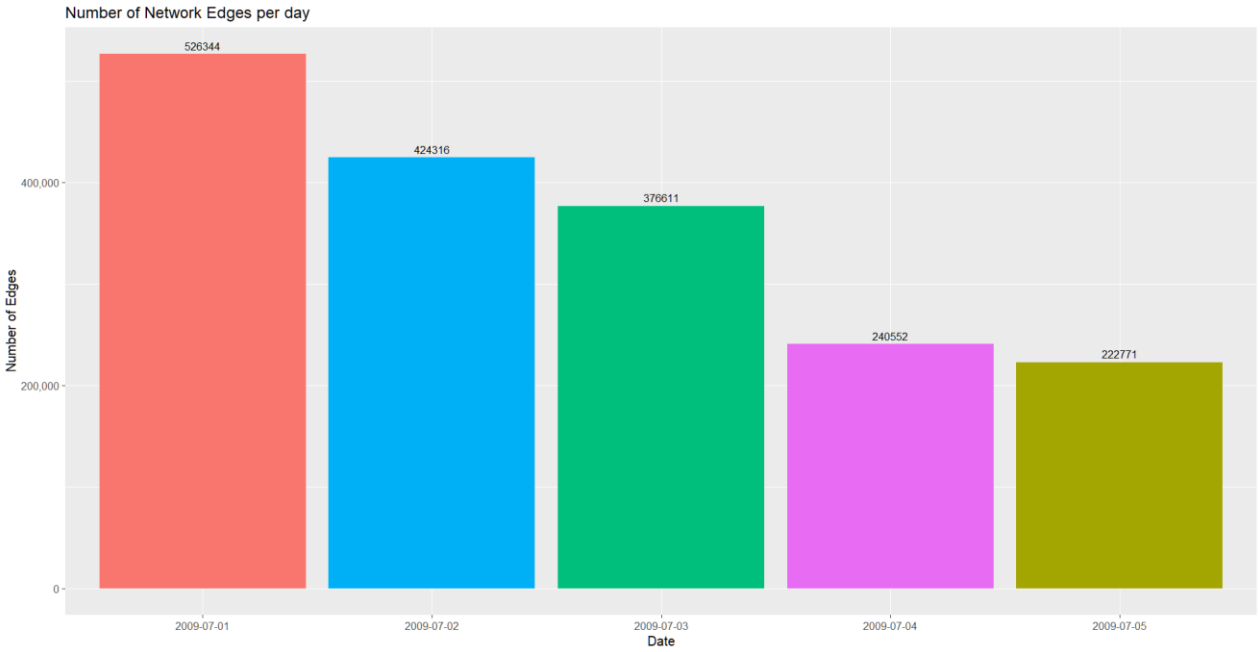
We now want to compare the number of nodes of each day graph. Below we can see a chart for doing that:



As we can see, the 1st day of July seems to have a huge jump and in the next days, it starts to fall. That means that something huge happened in the beginning of July which triggered many people to spread the news to another by mentions.

EDGES

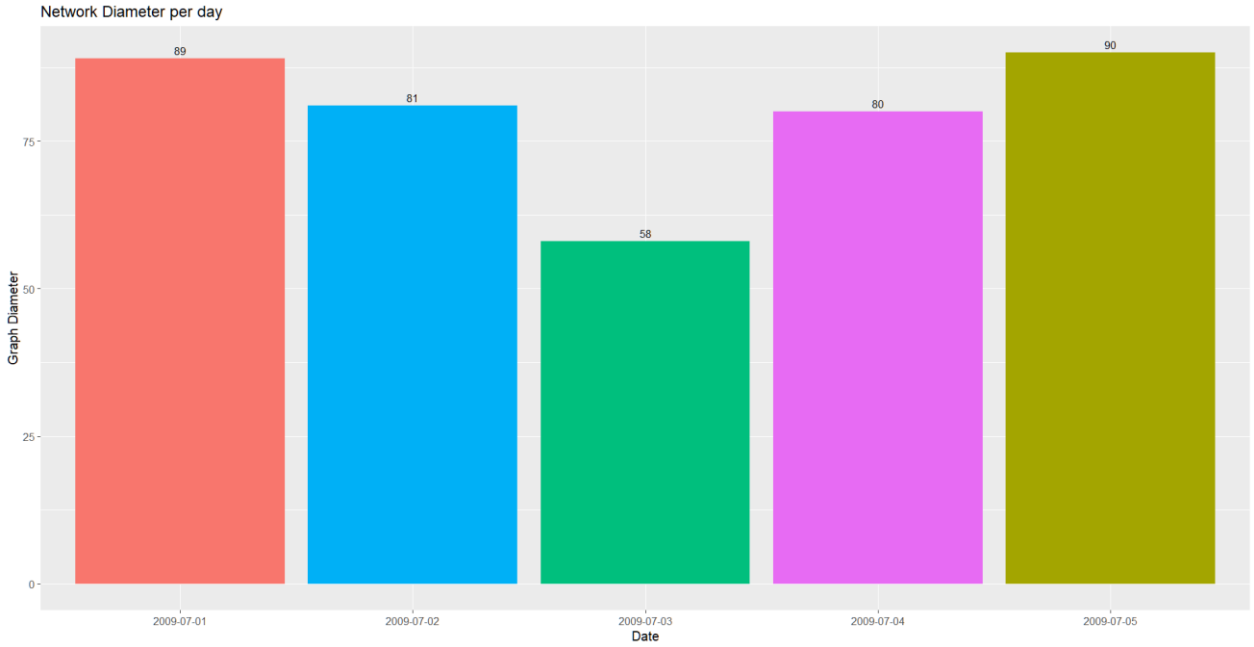
We now want to compare the number of edges of each day graph. Below we can see a chart for doing that:



As we can see, the 1st day of July seems to follow the same pattern.

DIAMETER

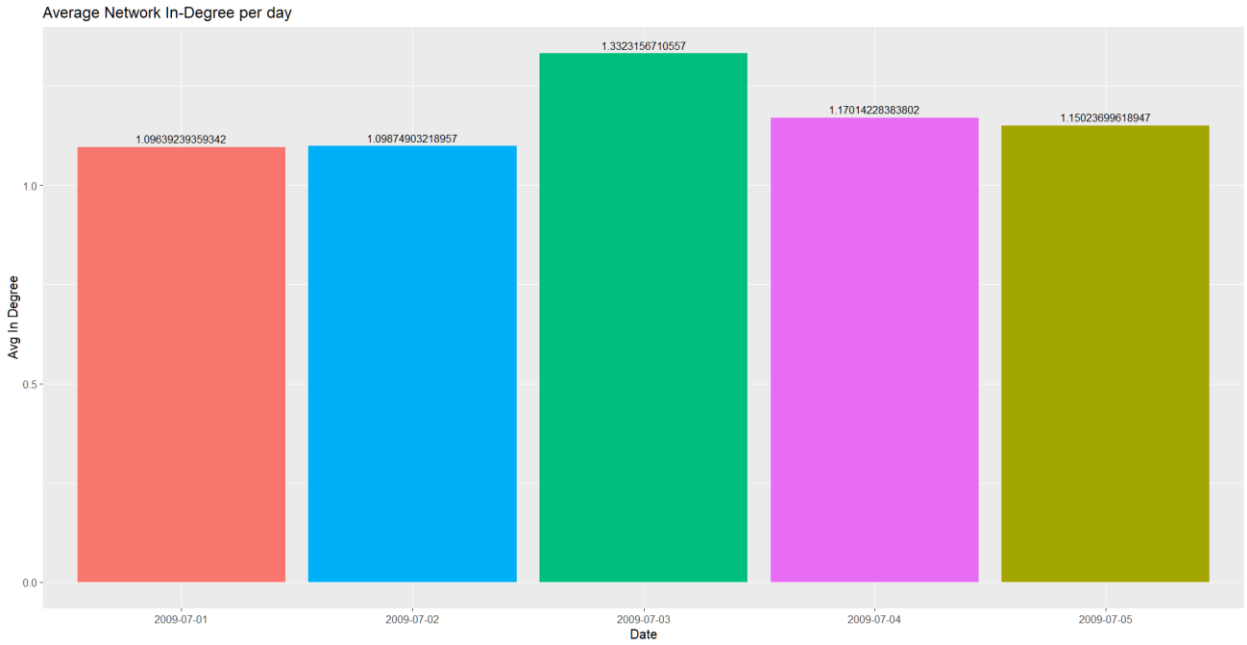
We now want to compare the diameter of each day graph. Below we can see a chart for doing that:



As we can see, in the beginning of the month, we had a huge graph in which the next days, started to shrink and then it started to rise again at 5th of July.

IN DEGREE

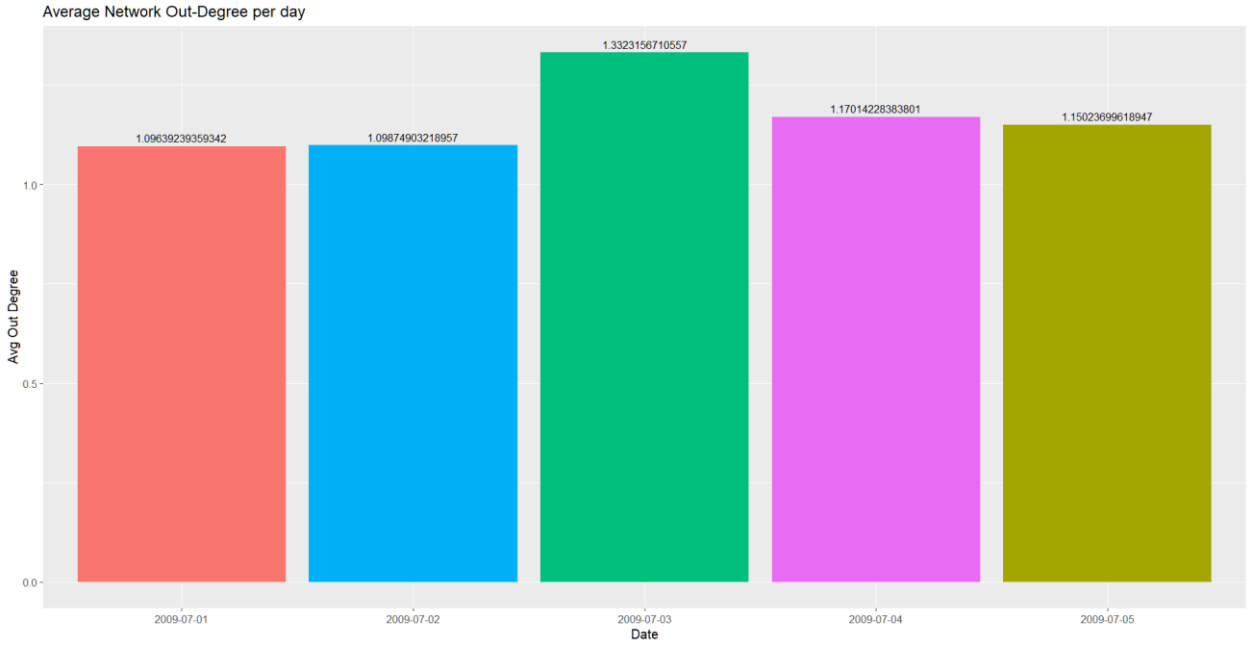
We now want to compare the average in-degree of each day graph. Below we can see a chart for doing that:



As we can see, the in-degree in the 3rd day of July is higher than the rest and that is logical because this graph has the lowest diameter, meaning that it is a smaller graph than the others. Because of that, the average in-degree seems to be higher.

OUT DEGREE

We now want to compare the average out-degree of each day graph. Below we can see a chart for doing that:



As we can see, nothing really changed compared to average in-degree.

3) Important nodes

IN DEGREE

	User	In_Degree	Date
1	tweetmeme	2522	2009-07-01
2	mashable	1627	2009-07-01
3	addthis	1212	2009-07-01
4	smashingmag	965	2009-07-01
5	mileycyrus	778	2009-07-01
6	BreakingNews	763	2009-07-01
7	cnn	746	2009-07-01
8	GuyKawasaki	679	2009-07-01
9	aplusk	669	2009-07-01
10	rafinhabastos	629	2009-07-01
11	tweetmeme	2478	2009-07-02
12	ddlovato	2242	2009-07-02
13	mashable	1996	2009-07-02
14	cnnbrk	1300	2009-07-02
15	cnn	1219	2009-07-02
16	addthis	1118	2009-07-02
17	souljaboytellem	898	2009-07-02
18	OfficialTila	748	2009-07-02
19	officialtila	738	2009-07-02
20	mileycyrus	680	2009-07-02
21	tweetmeme	1826	2009-07-03
22	souljaboytellem	1379	2009-07-03
23	addthis	1002	2009-07-03
24	mashable	940	2009-07-03
25	BreakingNews	874	2009-07-03
26	cnnbrk	856	2009-07-03
27	moontweet	720	2009-07-03
28	lilduval	428	2009-07-03
29	PhillyD	365	2009-07-03
30	adamlambert	362	2009-07-03
31	BreakingNews	949	2009-07-04
32	addthis	816	2009-07-04
33	tweetmeme	762	2009-07-04
34	iamdiddy	543	2009-07-04
35	mileycyrus	535	2009-07-04
36	cnnbrk	516	2009-07-04
37	mashable	456	2009-07-04
38	lilduval	454	2009-07-04
39	souljaboytellem	443	2009-07-04
40	TheOnion	350	2009-07-04
41	davidmmasters	1914	2009-07-05
42	iamdiddy	1147	2009-07-05
43	addthis	861	2009-07-05
44	tweetmeme	746	2009-07-05
45	mashable	550	2009-07-05
46	BreakingNews	490	2009-07-05
47	moontweet	360	2009-07-05
48	mileycyrus	353	2009-07-05
49	rainnwilson	339	2009-07-05
50	AKGovSarahPalin	332	2009-07-05

Above we can se 50 registrations of users, with the top 10 from each date for 5 days. We can see that many users are appearing more than one date. As an example, BreakingNews usually is at 6th place, except 3rd date which want 1st place. This make sense because BreakingNews seems to be an information channel and that's why many users have mentioned it.

OUT DEGREE

	User	Out_Degree	Date
1	dudebrochill	245	2009-07-01
2	failbus	215	2009-07-01
3	tsliquidators	215	2009-07-01
4	the_sims_3	202	2009-07-01
5	wootboot	200	2009-07-01
6	vaguetweetstest	193	2009-07-01
7	lmaobot	165	2009-07-01
8	drharvey	142	2009-07-01
9	luvorhate	119	2009-07-01
10	help_echo	106	2009-07-01
11	dudebrochill	279	2009-07-02
12	wootboot	240	2009-07-02
13	failbus	185	2009-07-02
14	the_sims_3	166	2009-07-02
15	dvdbot	158	2009-07-02
16	takeyourpin	147	2009-07-02
17	teamqivana	143	2009-07-02
18	luvorhate	127	2009-07-02
19	modelsupplies	125	2009-07-02
20	rt_thursday	119	2009-07-02
21	drejones71	624	2009-07-03
22	deana1981	605	2009-07-03
23	killah360dhh	438	2009-07-03
24	imbeeyo	431	2009-07-03
25	java4two	383	2009-07-03
26	ohmichael	347	2009-07-03
27	nachhi	340	2009-07-03
28	dudebrochill	305	2009-07-03
29	wootboot	277	2009-07-03
30	medic_ray	271	2009-07-03
31	swbot	830	2009-07-04
32	dudebrochill	391	2009-07-04
33	wootboot	353	2009-07-04
34	fxxxyourlife	257	2009-07-04
35	andreapuddu	246	2009-07-04
36	azandiamjbb	244	2009-07-04
37	hoboprophet	240	2009-07-04
38	failbus	239	2009-07-04
39	herpescure	216	2009-07-04
40	twiprodigy009	202	2009-07-04
41	swbot	876	2009-07-05
42	twiprodigy008	808	2009-07-05
43	twiprodigy005	672	2009-07-05
44	twiprodigy007	644	2009-07-05
45	twiprodigy009	588	2009-07-05
46	wildingp	339	2009-07-05
47	dudebrochill	331	2009-07-05
48	wootboot	319	2009-07-05
49	hoboprophet	255	2009-07-05
50	the_sims_3	225	2009-07-05

Above we can se 50 registrations of users, with the top 10 from each date for 5 days. We can see that many users are appearing more than one date. As an example, dudebrochill started at 1st place and ending up 7th place in 5th date. That means that he wanted to express his opinion in many people in the first dates.

PAGERANK

	User	PageRank	Date
1	tweetmeme	0.0017889774	2009-07-01
2	mashable	0.0012591438	2009-07-01
3	addthis	0.0011849832	2009-07-01
4	smashingmag	0.0011813695	2009-07-01
5	cnn	0.0007182473	2009-07-01
6	mileycyrus	0.0007096070	2009-07-01
7	KISSmetrics	0.0006783605	2009-07-01
8	CourageCampaign	0.0006260832	2009-07-01
9	aplusk	0.0005397417	2009-07-01
10	rafinhabastos	0.0005195846	2009-07-01
11	ddlovato	0.0028156243	2009-07-02
12	drew_taubenfeld	0.0023948782	2009-07-02
13	mashable	0.0021490222	2009-07-02
14	tweetmeme	0.0021307449	2009-07-02
15	globalmanners	0.0018296943	2009-07-02
16	cnn	0.0015276583	2009-07-02
17	addthis	0.0013608868	2009-07-02
18	souljaboytellem	0.0012128067	2009-07-02
19	cnnbrk	0.0011659017	2009-07-02
20	mileycyrus	0.0007575637	2009-07-02
21	tweetmeme	0.0024601761	2009-07-03
22	souljaboytellem	0.0023079639	2009-07-03
23	killerstartups	0.0020972462	2009-07-03
24	addthis	0.0017641800	2009-07-03
25	moontweet	0.0012360145	2009-07-03
26	cnnbrk	0.0011727753	2009-07-03
27	mashable	0.0011172055	2009-07-03
28	BreakingNews	0.0010180932	2009-07-03
29	PhillyD	0.0007181871	2009-07-03
30	adamlambert	0.0006171114	2009-07-03
31	souljaboytellem	0.0056384024	2009-07-04
32	addthis	0.0019969251	2009-07-04
33	tweetmeme	0.0016726163	2009-07-04
34	BreakingNews	0.0016722975	2009-07-04
35	lilduval	0.0012235658	2009-07-04
36	mileycyrus	0.0011959503	2009-07-04
37	mashable	0.0011092977	2009-07-04
38	iamdiddy	0.0010881197	2009-07-04
39	cnnbrk	0.0010306820	2009-07-04
40	garyvee	0.0009087917	2009-07-04
41	davidmmasters	0.0034208343	2009-07-05
42	iamdiddy	0.0029246273	2009-07-05
43	addthis	0.0022322762	2009-07-05
44	aplusk	0.0021665916	2009-07-05
45	tweetmeme	0.0016896555	2009-07-05
46	mashable	0.0010695413	2009-07-05
47	mrskutcher	0.0009199140	2009-07-05
48	moontweet	0.0008531593	2009-07-05
49	BreakingNews	0.0007359969	2009-07-05
50	mileycyrus	0.0007279574	2009-07-05

As far as PageRank, mashable is a nice example of it. He has high score in the first dates, because he was talking to important users but in the next days, he stayed behind.

4) Communities

FAST GREEDY CLUSTERING

This algorithm is particularly slow and it is hard to execute.

INFOMAP CLUSTERING

This algorithm practically never executed because it tries to find the best partition of the network based on information flow. This means that it need to compare each node with each other so in order to be executed, huge amount of resources are needed.

LOUVAIN CLUSTERING

This algorithm was the only one which was executed in fairly normal time so this algorithm is used in the next sections.

Below, you can see a comparison between Fast Greedy and Louvain clustering:

```
> compare(com_fg1, com_lvcl1)
[1] 2.028709
> compare(com_fg2, com_lvcl2)
[1] 2.051972
> compare(com_fg3, com_lvcl3)
[1] 2.487919
> compare(com_fg4, com_lvcl4)
[1] 1.982509
> compare(com_fg5, com_lvcl5)
[1] 1.841737
```

As we can see, they produce different communities with each other.

COMMUNITY ANALYSIS

We need to extract one pseudorandom user that appears in all 5 graphs and then detect the evolution of the community this user belongs to. With the below code, we can extract the pseudorandom user:

```
alldg<-c(degree(g1,mode='total',loops=FALSE),degree(g2,mode='total',loops=FALSE),
         degree(g3,mode='total',loops=FALSE),degree(g4,mode='total',loops=FALSE),
         degree(g5,mode='total',loops=FALSE))

common_user <- table(names(alldg))

common_user<- common_user[common_user == 5]

set.seed(42)

random_user <- names(sample(common_user, 1))
```

We have set a seed in order to be reproduceable. Our random user is: PeaceZicklin

Then, after finding the communities he belongs to, we compare these communities:

```
closeness :
Community 1 : 8.813925e-05
Community 2 : 1
Community 3 : 8.758565e-05
Community 4 : 0.01359627
Community 5 : 0.2611111

betweenness :
Community 1 : 4610.825
Community 2 : 0
Community 3 : 4600.757
Community 4 : 19.31898
Community 5 : 0.3333333

in_degree :
Community 1 : 2.136947
Community 2 : 1
Community 3 : 2.554873
Community 4 : 6.2
Community 5 : 1.333333
```

```
out_degree :
Community 1 : 2.136947
Community 2 : 1
Community 3 : 2.554873
Community 4 : 6.2
Community 5 : 1.333333

modularity :
Community 1 : 0.9001346
Community 2 : 0.9025843
Community 3 : 0.847891
Community 4 : 0.8846259
Community 5 : 0.898963
```

Based on my analysis, Community 5 stands out with the highest closeness centrality value of 0.261, indicating that the nodes within this community are relatively close to each other compared to other communities. Additionally, Community 1 has a closeness centrality value of 1, suggesting that all nodes within this community are directly connected.

In terms of betweenness centrality, Community 1 shows the highest value of 4610.825, indicating that the nodes in this community play a crucial role in connecting different parts of the network as important intermediaries. Conversely, Community 2 has the lowest betweenness centrality value of 0, indicating that the nodes within this community do not serve as significant intermediaries.

Looking at the in-degree and out-degree distributions, all communities have similar average values. However, Community 4 slightly exceeds other communities in both in-degree and out-degree values, implying a relatively balanced flow of connections both into and out of the nodes in each community.

Regarding modularity, Community 2 demonstrates the highest modularity value of 0.9025843, suggesting a strong division of the network into distinct communities. Conversely, Community 3 has the lowest modularity value of 0.847891, indicating that the division of this community is not as well-defined.

TOPIC DETECT

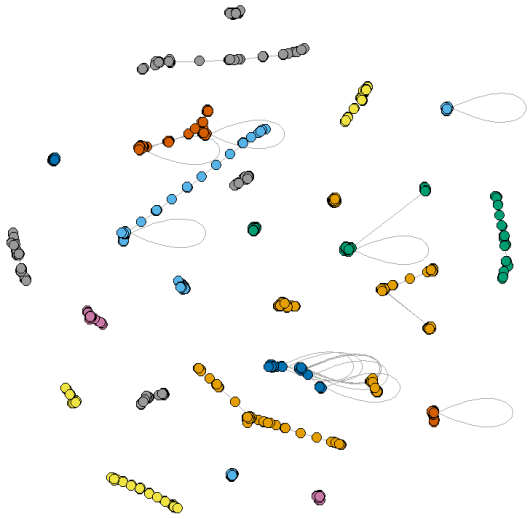
Below are the top 5 topics for each community PeaceZicklin belongs to:

```
Top 5 Community Topics for Community 1 are: #quote #woofwednesday #moonfruit #voss #fb
Top 5 Community Topics for Community 2 are: #moonfruit #MMOT #michaeljackson #p2 #quotes
Top 5 Community Topics for Community 3 are: #quote #tcot #moonfruit #fb #voss
Top 5 Community Topics for Community 4 are: #quote #moonfruit #fb #voss #tcot
Top 5 Community Topics for Community 5 are: #FringeTO #Maddie #1chiphttp #alice #AmericanGirl
```

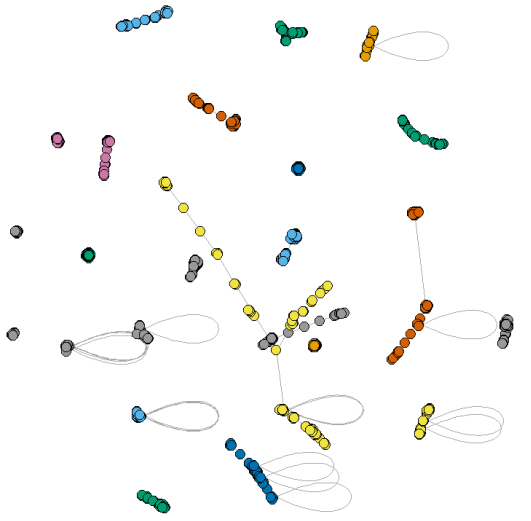
Below, are the top 5 topics among all communities that PeaceZicklin belongs to:

```
"#quote"     "#moonfruit" "#voss"      "#fb"          "#quotes"
```

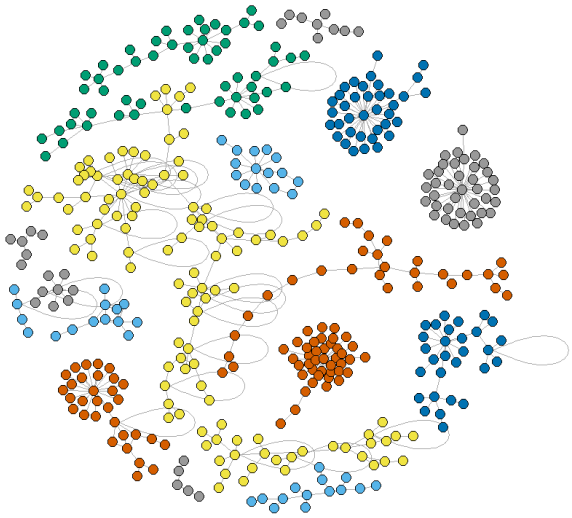
COMMUNITY VISUALIZATION



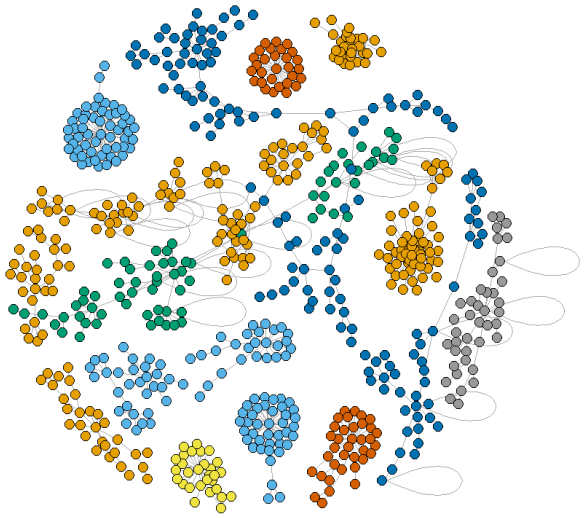
Communities - Day: 2009-07-01



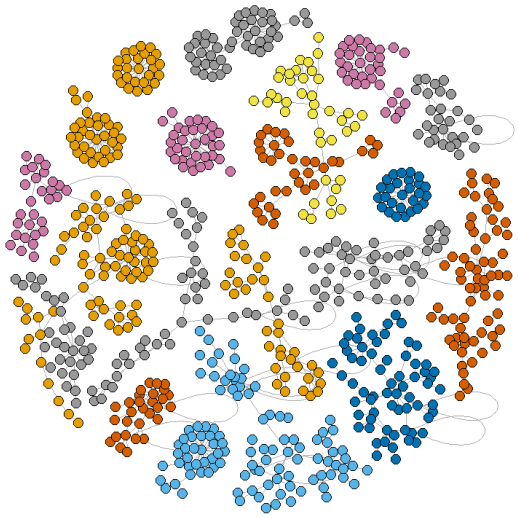
Communities - Day: 2009-07-02



Communities - Day: 2009-07-03



Communities - Day: 2009-07-04



Communities - Day: 2009-07-05