Homework 2:
From raw data to temporal graph structure exploration

Piermattia Schoch

1. Create a weighted directed graph with igraph, using raw data from Twitter.

Basically the approach used is the following:

Since the raw data scraped are quite large in size (6.94Gb) and R reads the entire dataset into RAM all at once it will inevitably face issues in reading this file. Given the fact that we are interested just in a small portion of the dataset (the first 5 days of July) and that it is ordered chronologically by date, i decided to split this files into smaller chunks (800Mb) and keep and load just the first one obtained (which contains the dates of interest).

After some cleaning procedure, which is possible to find in "Cleaning.R" file attached in the submission, I obtained 5 csv files, each for the first 5 days of July.

Once created, it is possible to load them directly in R and to start the graph analysis.

2. Average degree over time

A/B. Number of vertices and Edges:

Here the nodes (vertices) represent users who tweet or were mentioned. Since the graph is directed, there is an edge from a user A to user B if B is being mentioned by user A.

From the graph is possible to see a strong negative trend regarding the people who tweet and the mentions made during these 5days. It is interesting to see that the number of edges are always grater than the number of vertices, meaning that for the same bunch of users there are many of them mentioned more than 1 time. The biggest gap between vertices and edges is registered the 3° July.

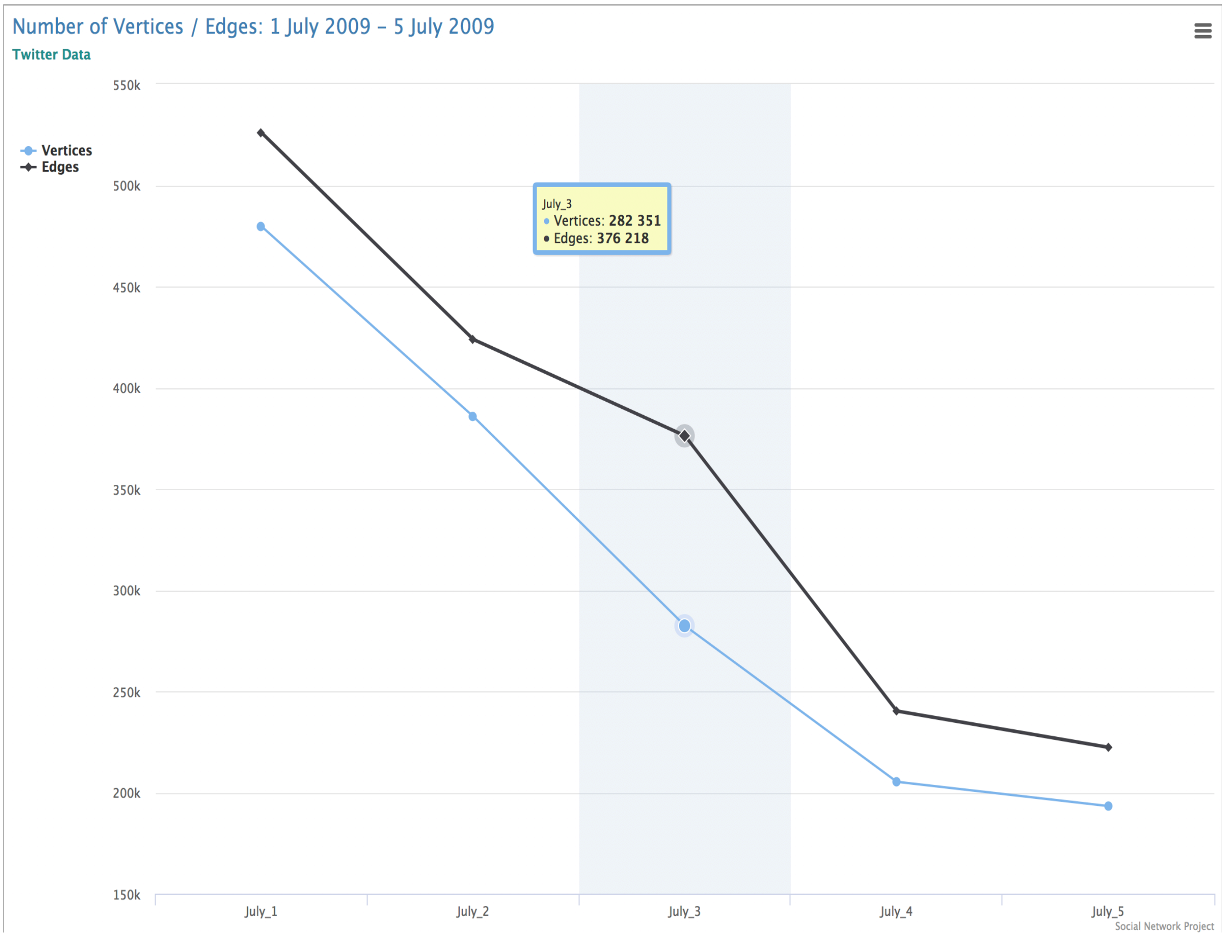


Fig.1 : Vertices and Edges (1/07/2009 - 5/07/2009)

C. Diameter

In Fig.2 it is possible to see the value of diameter. The lowest value is found in the third day of July, meaning that the in this day the users are more connected between each other (the network is more dense). As a consequence, the number of mentions made are greater compared to other days in percentange with respect to the number of users in the network.

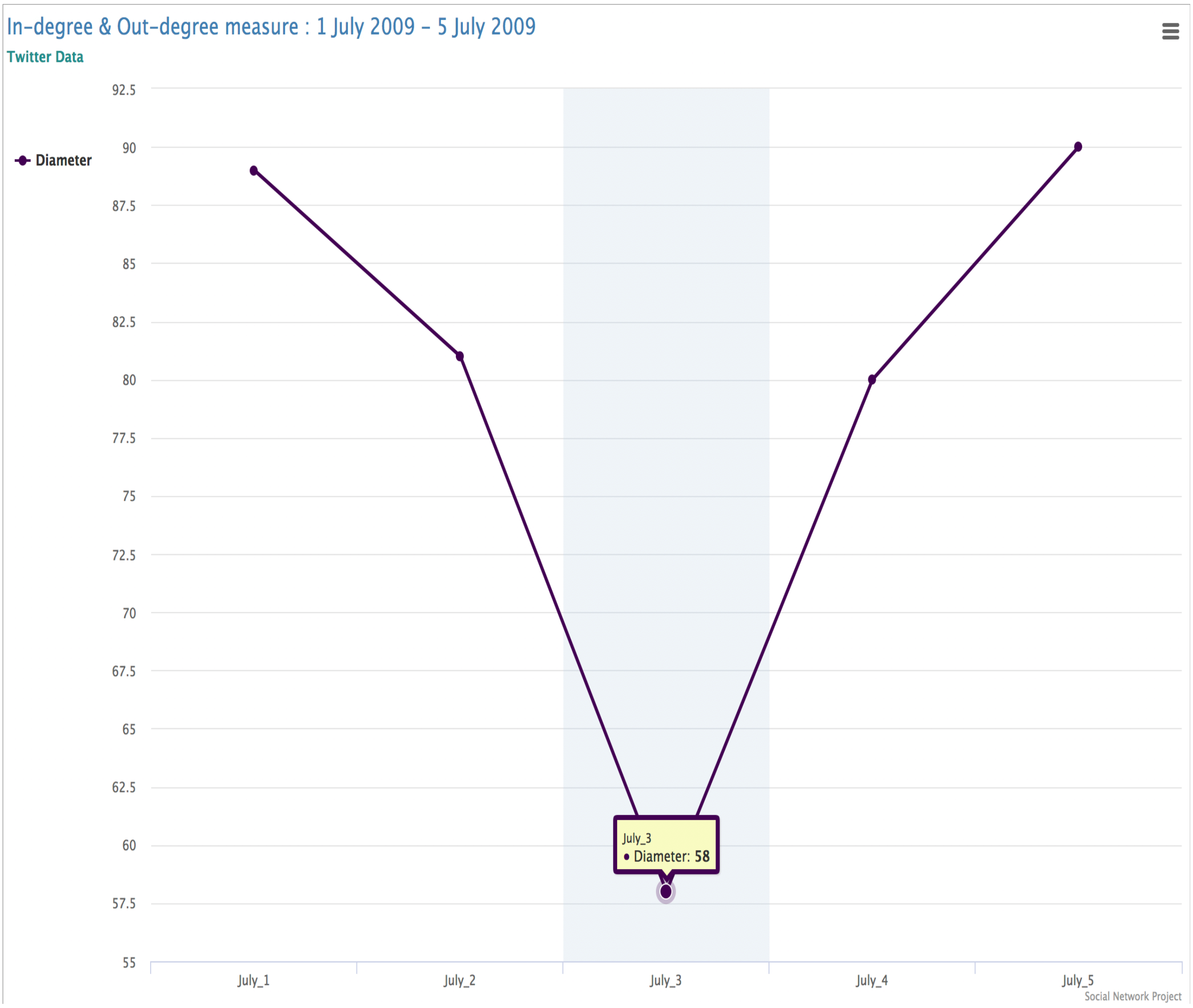


Fig.2 : Diameter (1/07/2009 - 5/07/2009)

D/E. Average In-Degree and Out-Degree

In Fig.3 it is possible to see the value of average in-degree and out-degree during the 5days.

Since the graphs are directed, the sum of in degree and out degree is equal, as well as their average. The peak is reached on July 3, which is the expected result considering the result in Fig.1 and Fig.2.

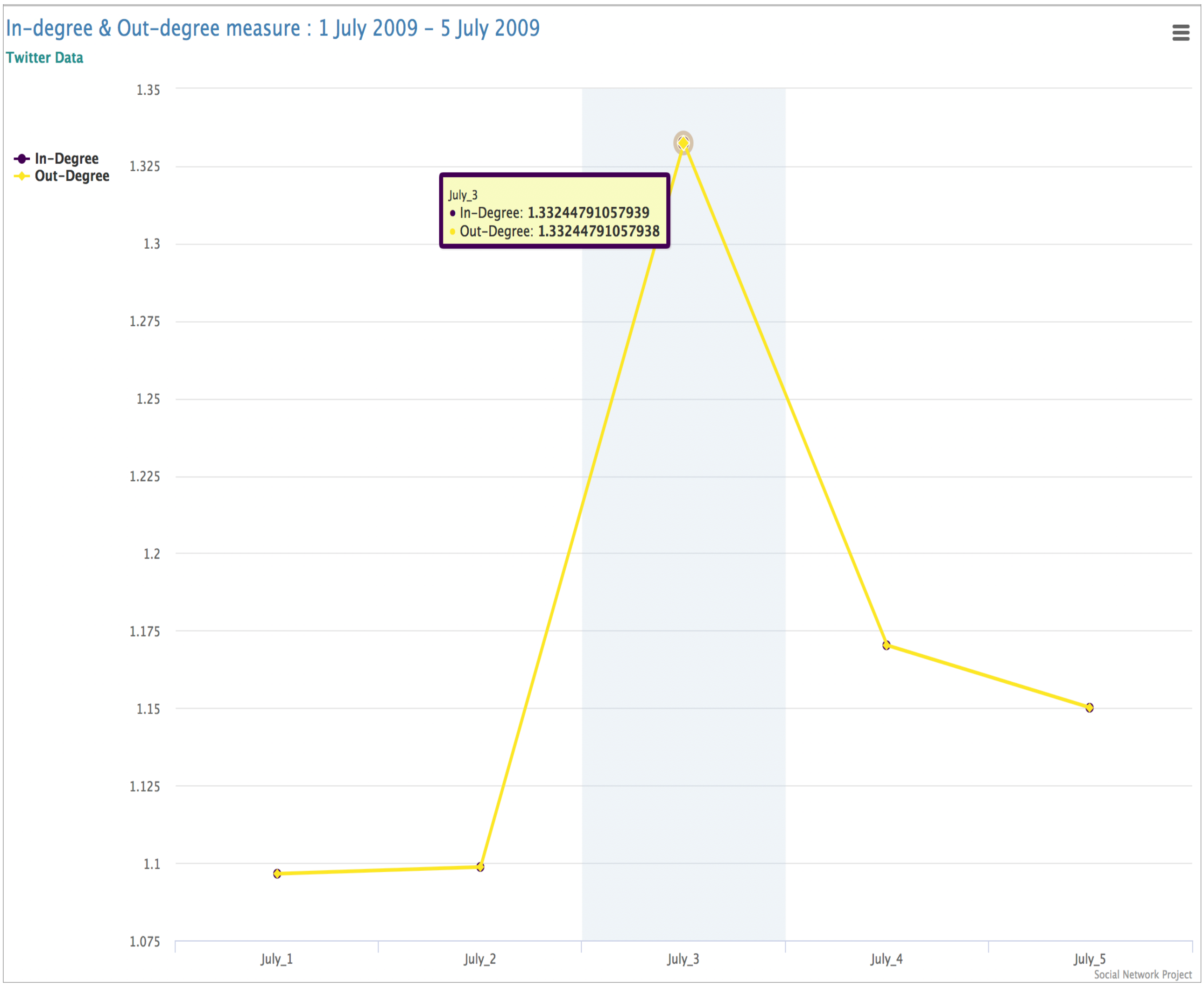


Fig.3 : Average In-Degree, Out-Degree (1/07/2009 - 5/07/2009)

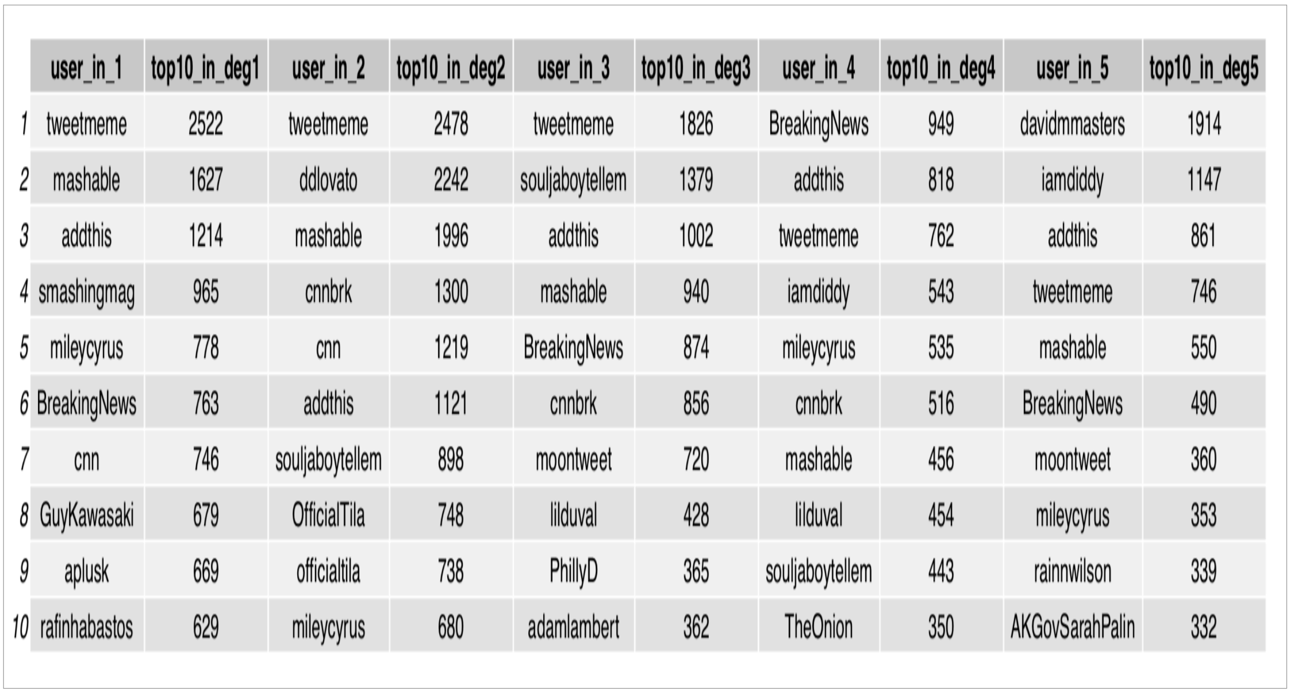
3. Import nodes

A. Indegree:

According to in-degree measure, which gives information about all the people that have been mentioned, it is possible to see that some user are in ranked in the top10 each of the 5 days (Tweetmeme, Mashable, Addthis).

Most of the top ranked accounts are not individual user, but social pages that spread specific information (from news to funny pages). In the table, there are also celebrities (iamdiddy, ddlovato, ...).

Davidmmasters was the user most mentioned during July5, while in the other days he is not present in the top10 list.



	user_in_1	top10_in_deg1	user_in_2	top10_in_deg2	user_in_3	top10_in_deg3	user_in_4	top10_in_deg4	user_in_5	top10_in_deg5
1	tweetmeme	2522	tweetmeme	2478	tweetmeme	1826	BreakingNews	949	davidmmasters	1914
2	mashable	1627	ddlovato	2242	souljaboytellem	1379	addthis	818	iamdiddy	1147
3	addthis	1214	mashable	1996	addthis	1002	tweetmeme	762	addthis	861
4	smashingmag	965	cnnbrk	1300	mashable	940	iamdiddy	543	tweetmeme	746
5	mileycyrus	778	cnn	1219	BreakingNews	874	mileycyrus	535	mashable	550
6	BreakingNews	763	addthis	1121	cnnbrk	856	cnnbrk	516	BreakingNews	490
7	cnn	746	souljaboytellem	898	moontweet	720	mashable	456	moontweet	360
8	GuyKawasaki	679	OfficialTila	748	lilduval	428	lilduval	454	mileycyrus	353
9	aplusk	669	officialtila	738	PhillyD	365	souljaboytellem	443	rainnwilson	339
10	rafinhabastos	629	mileycyrus	680	adamlambert	362	TheOnion	350	AKGovSarahPalin	332

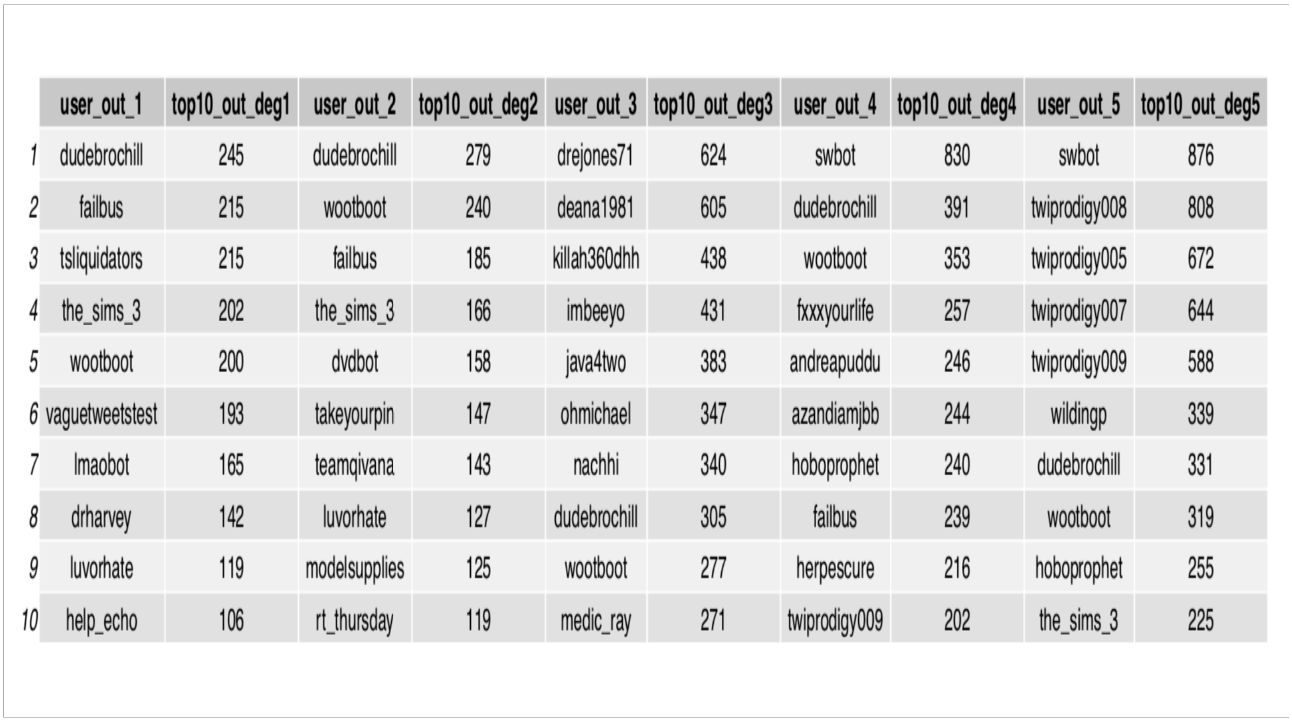
Fig.4 : Top 10 user according to In-Degree(1/07/2009 - 5/07/2009)

B. Out-Degree

Overall, looking the dataframe of the top10 user according to out-degree measure, which tells us how many mentions a user made in each day, the values are lower compared to the in-degree measure. This is an implicit effect of the structure of the network. Indeed, there are nodes (users/pages) that receive mentions from many other users in the same day.

Overall, the trend during the 5 days is increasing, also due to some user which can be defined as bot:

(swbot, wootbot, twiprodigy005/7/8/9)



	user_out_1	top10_out_deg1	user_out_2	top10_out_deg2	user_out_3	top10_out_deg3	user_out_4	top10_out_deg4	user_out_5	top10_out_deg5
1	dudebrochill	245	dudebrochill	279	drejones71	624	swbot	830	swbot	876
2	failbus	215	wootboot	240	deana1981	605	dudebrochill	391	twiprodigy008	808
3	tsliquidators	215	failbus	185	killah360dhh	438	wootboot	353	twiprodigy005	672
4	the_sims_3	202	the_sims_3	166	imbeeyo	431	fxxxyourlife	257	twiprodigy007	644
5	wootboot	200	dvdbot	158	java4two	383	andreapuddu	246	twiprodigy009	588
6	vaguetweetstest	193	takeyourpin	147	ohmichael	347	azandiamjbb	244	wildingp	339
7	lmaobot	165	teamqivana	143	nachhi	340	hoboprophet	240	dudebrochill	331
8	drharvey	142	luvorhate	127	dudebrochill	305	failbus	239	wootboot	319
9	luvorhate	119	modelsupplies	125	wootboot	277	herpescure	216	hoboprophet	255
10	help_echo	106	rt_thursday	119	medic_ray	271	twiprodigy009	202	the_sims_3	225

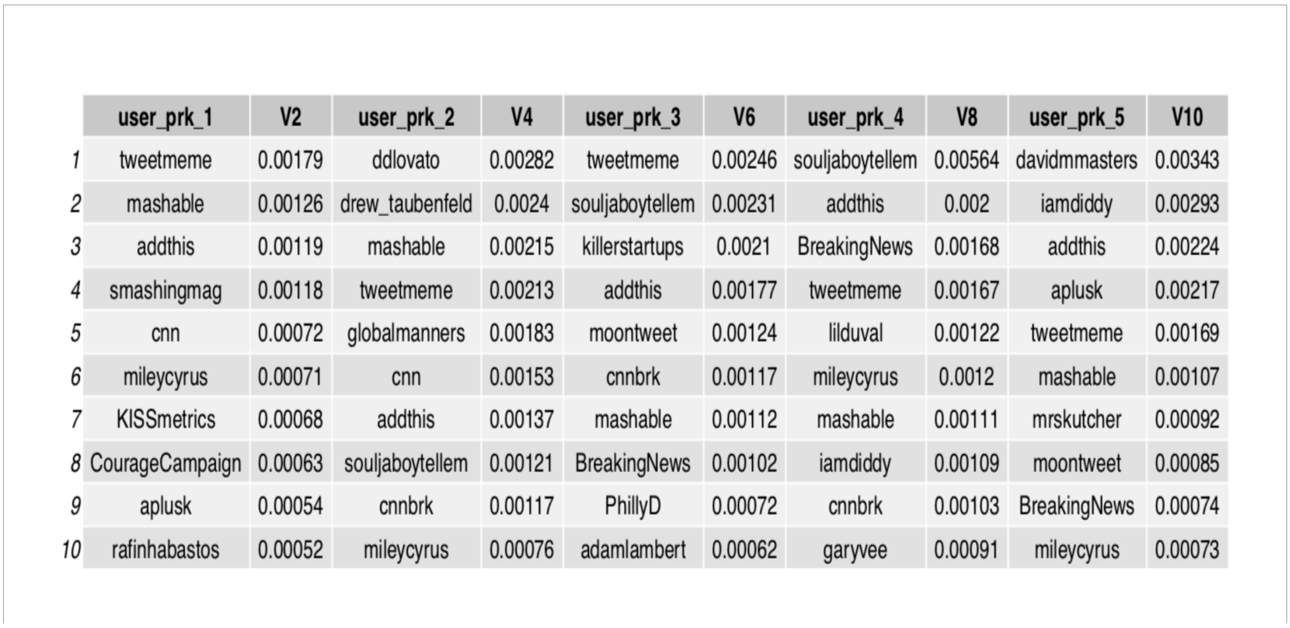
Fig.5 : Top 10 user according to Out-Degree(1/07/2009 - 5/07/2009)

C. Page Rank

The idea is that if a user A has a link to user B , the owner of A is giving some measure of importance to page B.

As we might expect, tweetmeme has the highest PageRank the first and the third day of July, because it has incoming links from all other pages. It's important to remember that it's not only the number of incoming links that is important, but also the importance of the pages behind those links.

Overall, the highest PageRank values are obtained by user with high in-degree values.



	user_prk_1	V2	user_prk_2	V4	user_prk_3	V6	user_prk_4	V8	user_prk_5	V10
1	tweetmeme	0.00179	ddlovato	0.00282	tweetmeme	0.00246	souljaboytellem	0.00564	davidmmasters	0.00343
2	mashable	0.00126	drew_taubenfeld	0.0024	souljaboytellem	0.00231	addthis	0.002	iamdiddy	0.00293
3	addthis	0.00119	mashable	0.00215	killerstartups	0.0021	BreakingNews	0.00168	addthis	0.00224
4	smashingmag	0.00118	tweetmeme	0.00213	addthis	0.00177	tweetmeme	0.00167	aplusk	0.00217
5	cnn	0.00072	globalmanners	0.00183	moontweet	0.00124	lilduval	0.00122	tweetmeme	0.00169
6	mileycyrus	0.00071	cnn	0.00153	cnnbrk	0.00117	mileycyrus	0.0012	mashable	0.00107
7	KISSmetrics	0.00068	addthis	0.00137	mashable	0.00112	mashable	0.00111	mrskutcher	0.00092
8	CourageCampaign	0.00063	souljaboytellem	0.00121	BreakingNews	0.00102	iamdiddy	0.00109	moontweet	0.00085
9	aplusk	0.00054	cnnbrk	0.00117	PhillyD	0.00072	cnnbrk	0.00103	BreakingNews	0.00074
10	rafinhabastos	0.00052	mileycyrus	0.00076	adamlambert	0.00062	garyvee	0.00091	mileycyrus	0.00073

Fig.6 : Top 10 user according to Page-Rank (1/07/2009 - 5/07/2009)

4. Communities

Identifying communities in these network is a crucial step for gaining an in-depth understanding on network structure, dynamics and interactions. Informally, a good community is a densely-connected group of nodes that is sparsely connected to the rest of the network

The "igraph" package implements a variety of network clustering methods, most of which are based on Newman-Girvan modularity.

- The first algorithm that I used is the "fast greedy" method. This function tries to find dense subgraph, also called communities in graphs via directly optimizing a modularity score. Time complexity: O(|E||V|log|V|) in the worst case, O(|E|+|V|log^2|V|) typically, |V| is the number of vertices, |E| is the number of edges.

- Secondly I performed "cluster_infomap". This function find community structure that minimizes the expected description length of a random walker trajectory. Time Complexity: none given; looks worst case like |V|(|V| + |E|) based on quick reading.

- The last algorithm used is "cluster_louvain". This function implements the multi-level modularity optimization algorithm for finding community structure.. Time Complexity: None given.

Considering that we are dealing with very large graphs (1° July csv contains 479640 Vertices and 525932 Edges) i firstly tried the algorithms with the smaller networks (5°- 4°- 3°- 2°- 1° July).

Fast Greedy algorithm was able to produce results only for the networks of 5°- 4° - 3° July, while for the other the algorithm was too computational expensive to obtain results.

Regarding Cluster Infomap i did not get result for any of the 5 graphs (as we can see the Time Complexity is higher (linear) than Fast-Greedy algorithm (logarithmic)).

Considering time efficiency, the best algorithm, among the ones tried, is Louvain clustering. It's much faster compared to the other 2 algorithms and was the only to provide results, which are displayed in Fig 7. For each day are illustrated the number of communities identified by the algorithm and the ratio between the largest community size and the total number of user.

Day	Number of communities	Largest community size / ratio
July 1	81045	0,04868443
July 2	63471	0.03258069
July 3	31426	0.03798891
July 4	28659	0.02447669
July 5	28013	0.01757777

Fig.7 : Louvian Clustering result

In the next step of the analysis I picked a random user ("zz23377737") which appeared in all the five networks and I analyzed the evolution of the communities this user belong to.

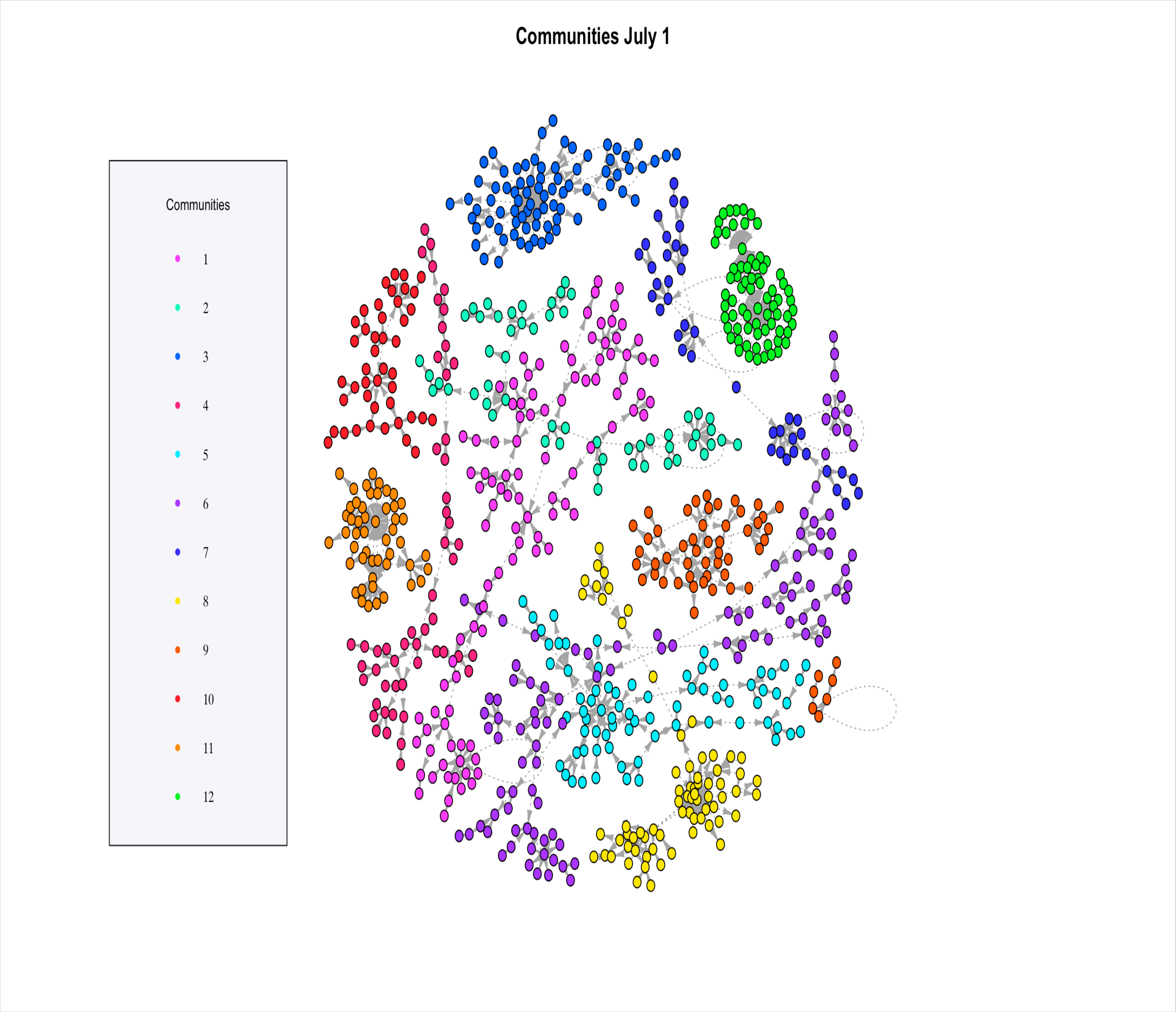
In order to have an idea about the evolution of communities in terms of similarities, an important metric, which has a significant meaning in the context of social network, is the community size.

According to Anthropologist Dunbar[1] suggests that the size of communities with strong ties in both traditional social networks and Internet-based social networks should be limited to 150 (called Dunbar's number) due to the cognitive constraint and time constraint of human being. Large communities of size over 150 contain weak connections among their members therefore are not stable, while small communities of size 2 or 3 cannot provide the strong sense of team or community. Therefore, we refer to communities of size 4 to 150 as desirable community.

The first two days this user belongs to large density communities (with 2813 and 2734 other users). Surprisingly, in the remaining days he belongs to tiny communities, made up by 2-2-3 other users. Given that, on one hand, groups of size 1, 2, and 3 are too small to be called a community and on the other hand, large communities cannot facilitate communication or interaction (therefore members in the community have limited influence upon each other) the insight that can be deduced looking this random user are not well defined.

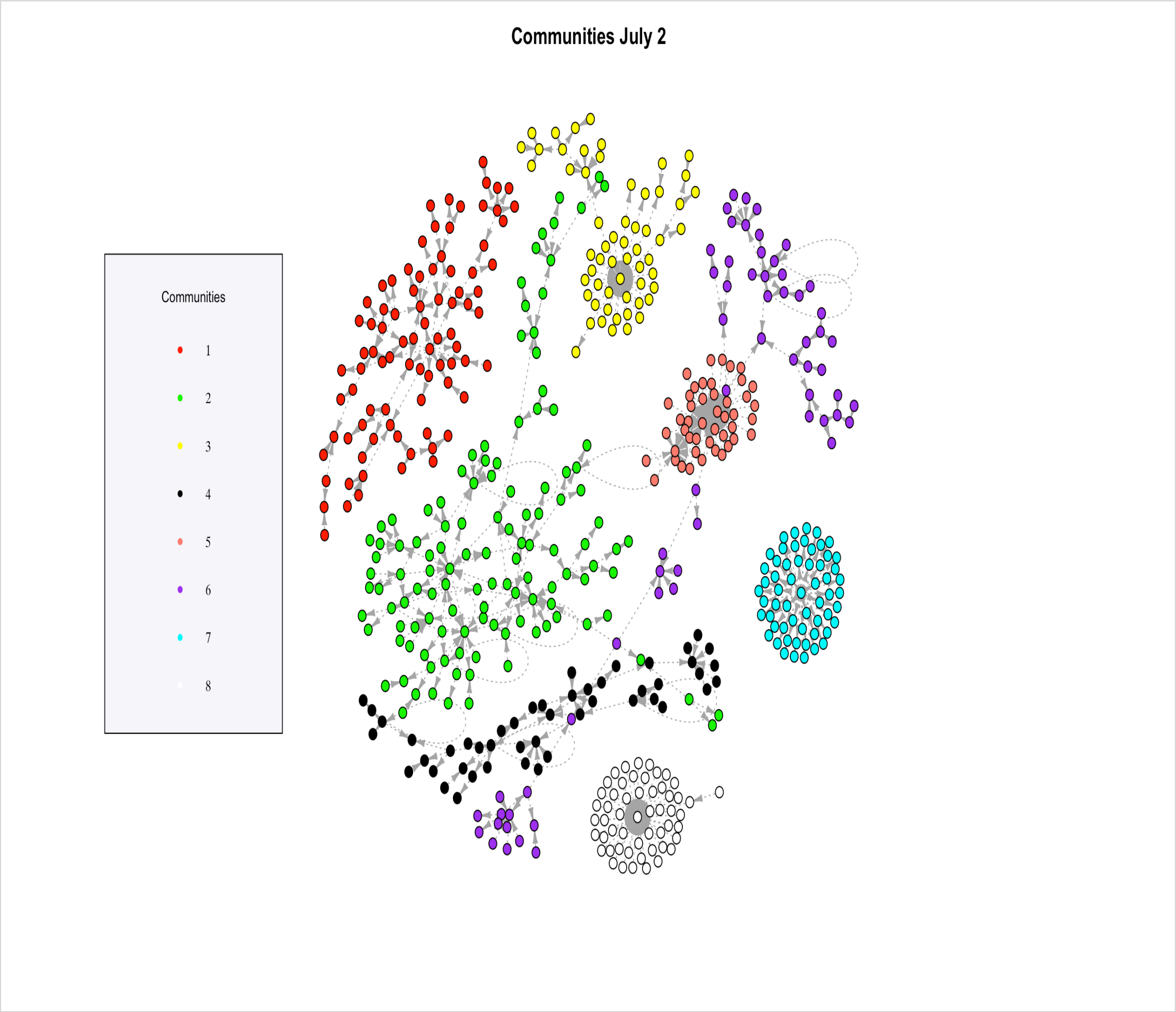
In the final stage I created a visualization of the graph using a different color for each community. In order to have neater graphs I decided to plot communities that have at least 40 members and at maximum 150.

In Fig 8, 9, 10, 11, 12 are illustrated the result of this analysis.


Communities July 1

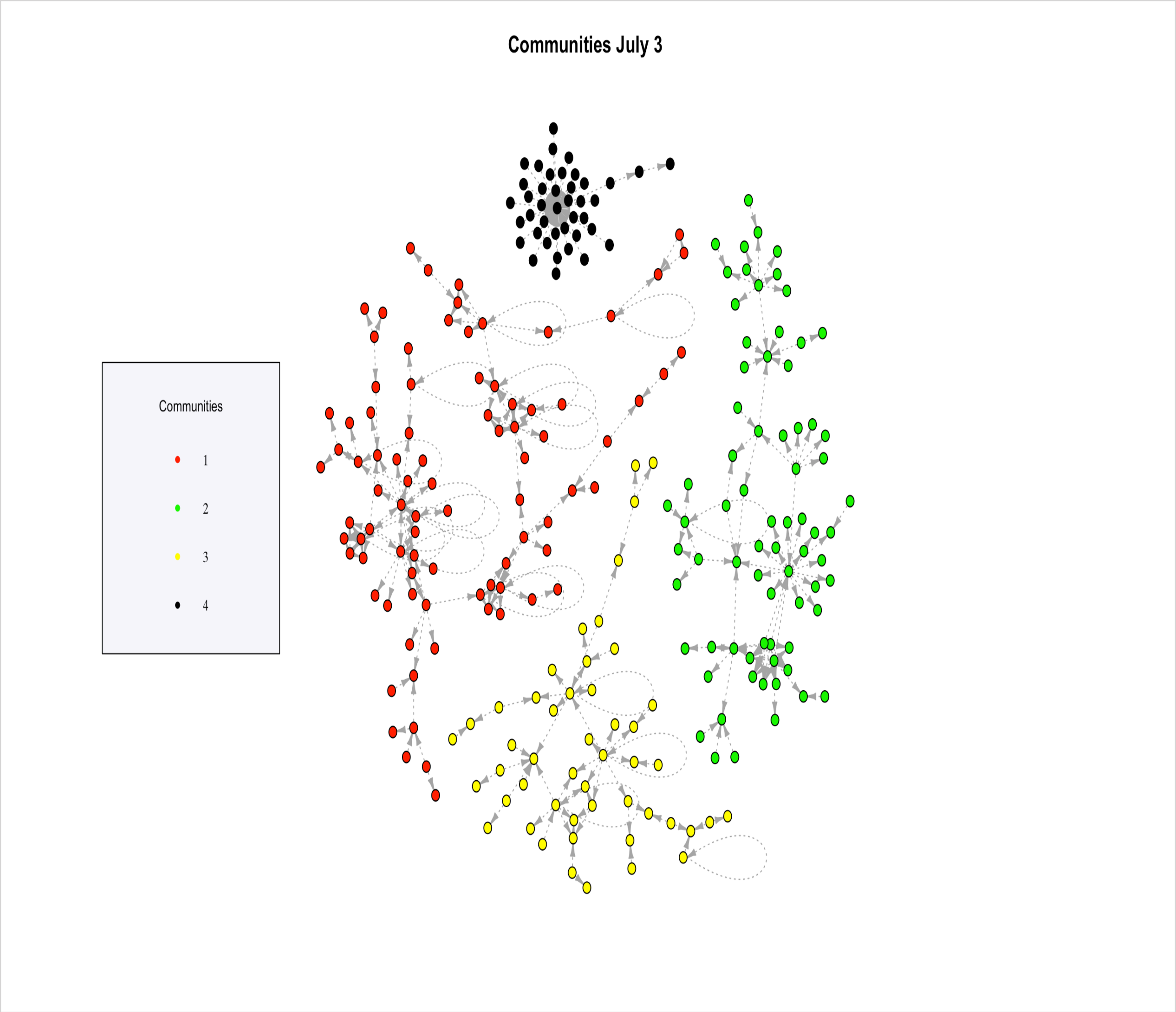
iGRAPH

twitter



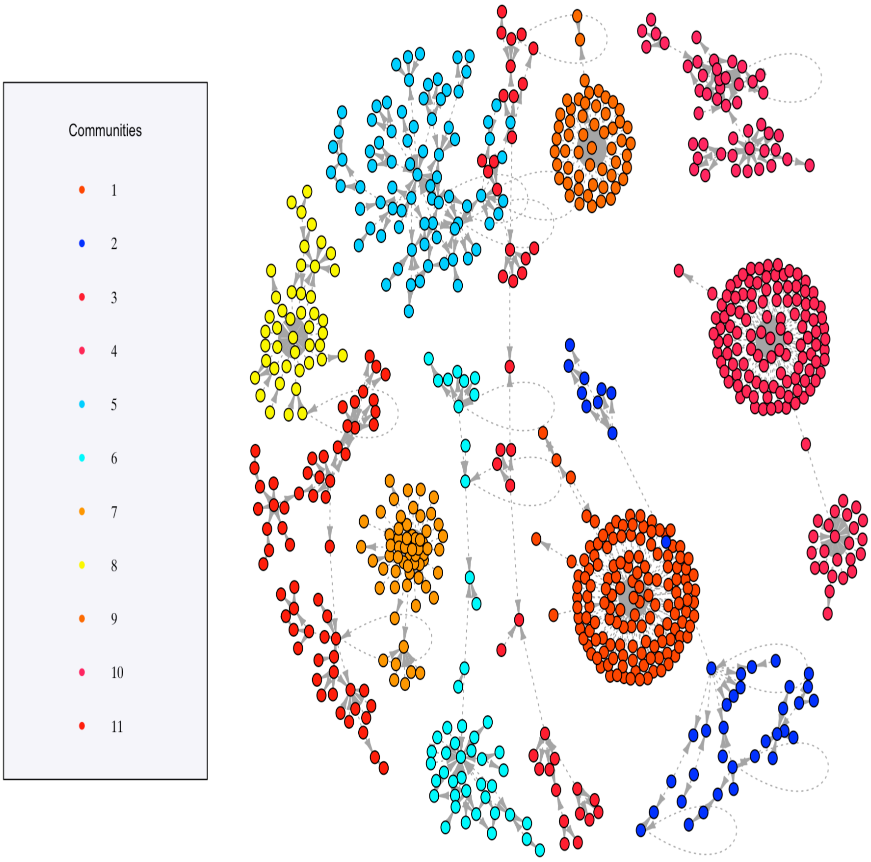
iGRAPH

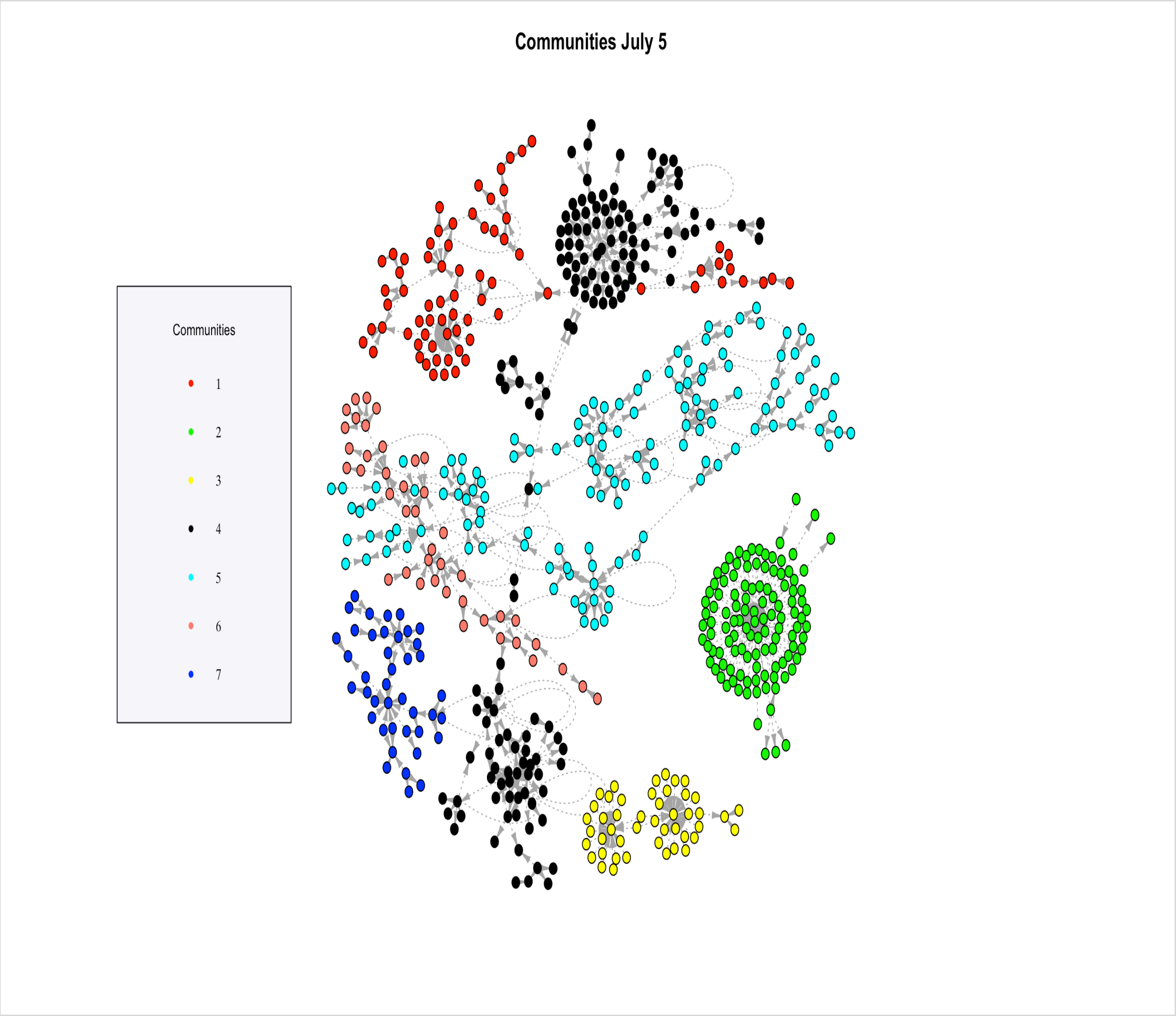
twitter



Communities July 3

Communities July 4




Communities July 5
Communities
● 1
● 2
● 3
● 4
● 5
● 6
● 7