Social Network Analysis II

Athens University of Economics and Business

Master of Science in Business Analytics

Christos Vlassis

f2822204

Data Manipulation.

In the first part of the assignment, we used Python to extract the relative information in the format needed to create the Graphs. The fact that the file that was given was large and could not be uploaded in RAM we had to make use of different techniques. To cope with this issue we first made some small manipulations in the data, we stored them in SSD then we imported the new files in Python and finally we created the final files. We also, deleted at the same time any files that we did not need. More details can be found in the Python file that contains the pipeline.

I made use of the UI only to change the file names to something smaller for my convenience.

Average degree over time

For the rest parts of the assignments, we will use R. We used a for loop and lists to append the metrics and the data frames to lists.

Before continuing we must make a note. The day at 2009-07-01 was Wednesday and of course, the day at 2009-07-05 was Sunday. So, we will use the days as our labels since they provide more information.

We create the following table for each of the metrics and for each day of the week:

Wednesday					
avg in degree	avg out degree	avg PageRank	Diameter	Number of vertices	Number of edges
1.0963	1.0963	2.10E-06	89	480069	526344

Thursday					
avg in degree	avg out degree	avg PageRank	Diameter	Number of vertices	Number of edges
1.098	1.098	2.60E-06	81	386181	424316

Friday					
avg in degree	avg out degree	avg PageRank	Diameter	Number of vertices	Number of edges
1.332	1.332	3.50E-06	58	282674	376611

Saturday					
avg in degree	avg out degree	avg PageRank	Diameter	Number of vertices	Number of edges
1.1701	1.1701	4.90E-06	80	205575	240552

Sunday					
avg in degree	avg out degree	avg PageRank	Diameter	Number of vertices	Number of edges
1.1502	1.1502	5.20E-06	90	193674	222771

For the **in degree** and **out degree** (those metrics have the same values) we find some variation for each day of the week with Friday having the largest value. Wednesday and Thursday have the lower values. Finally, Saturday and Sunday appear to have almost the same values with Saturday having a larger value, but the difference is insignificant.

For the **PageRank,** we find overall small values. More specifically, Wednesday and Friday appear to have the lowest PageRank with Friday having a higher PageRank. Finally, the weekend appears to have the largest, overall, PageRank's.

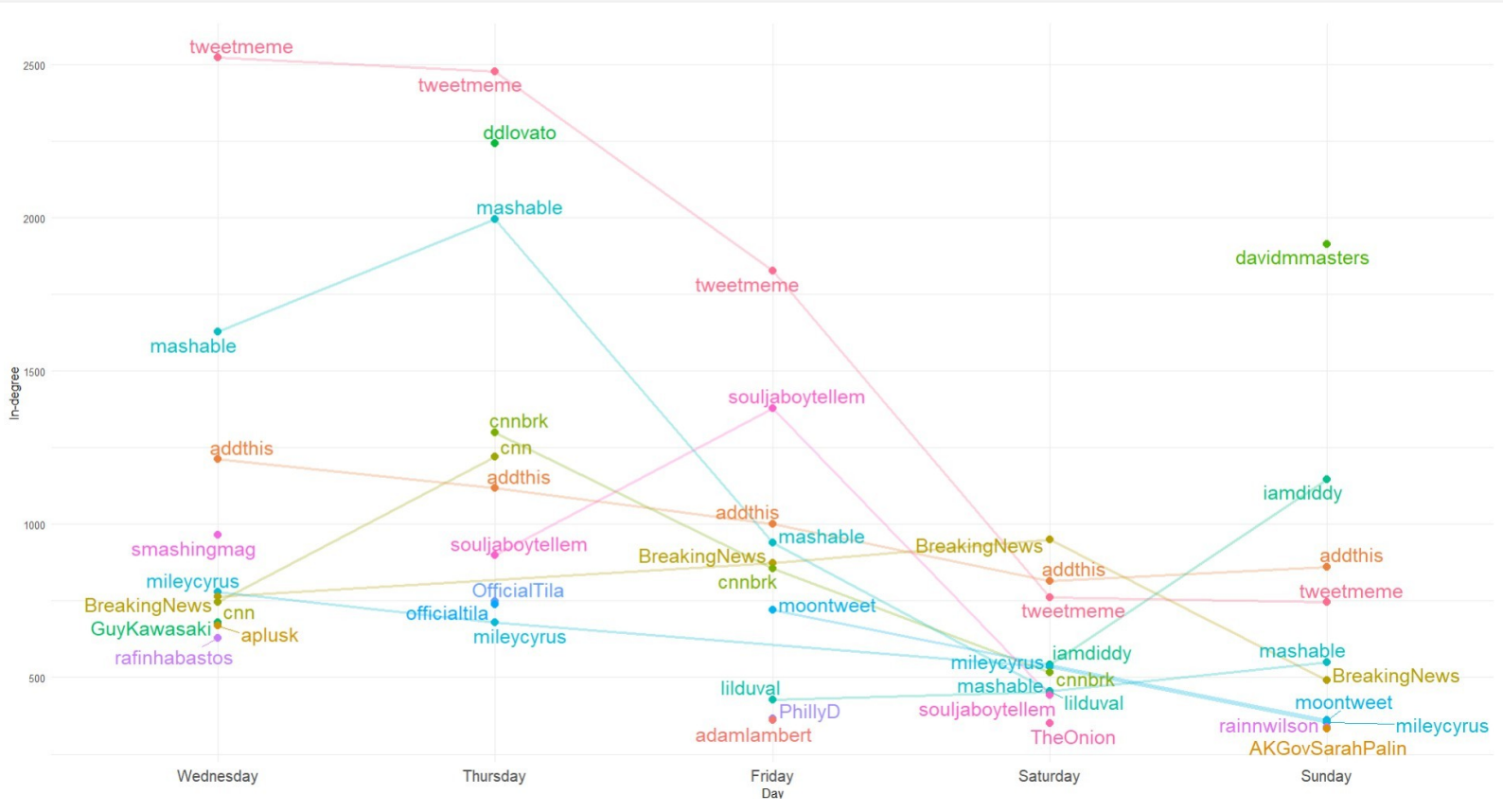
Finally, for the **Number of vertices and edges**, we found that as the week progresses these metrics are being reduced.

Important nodes

In Degree

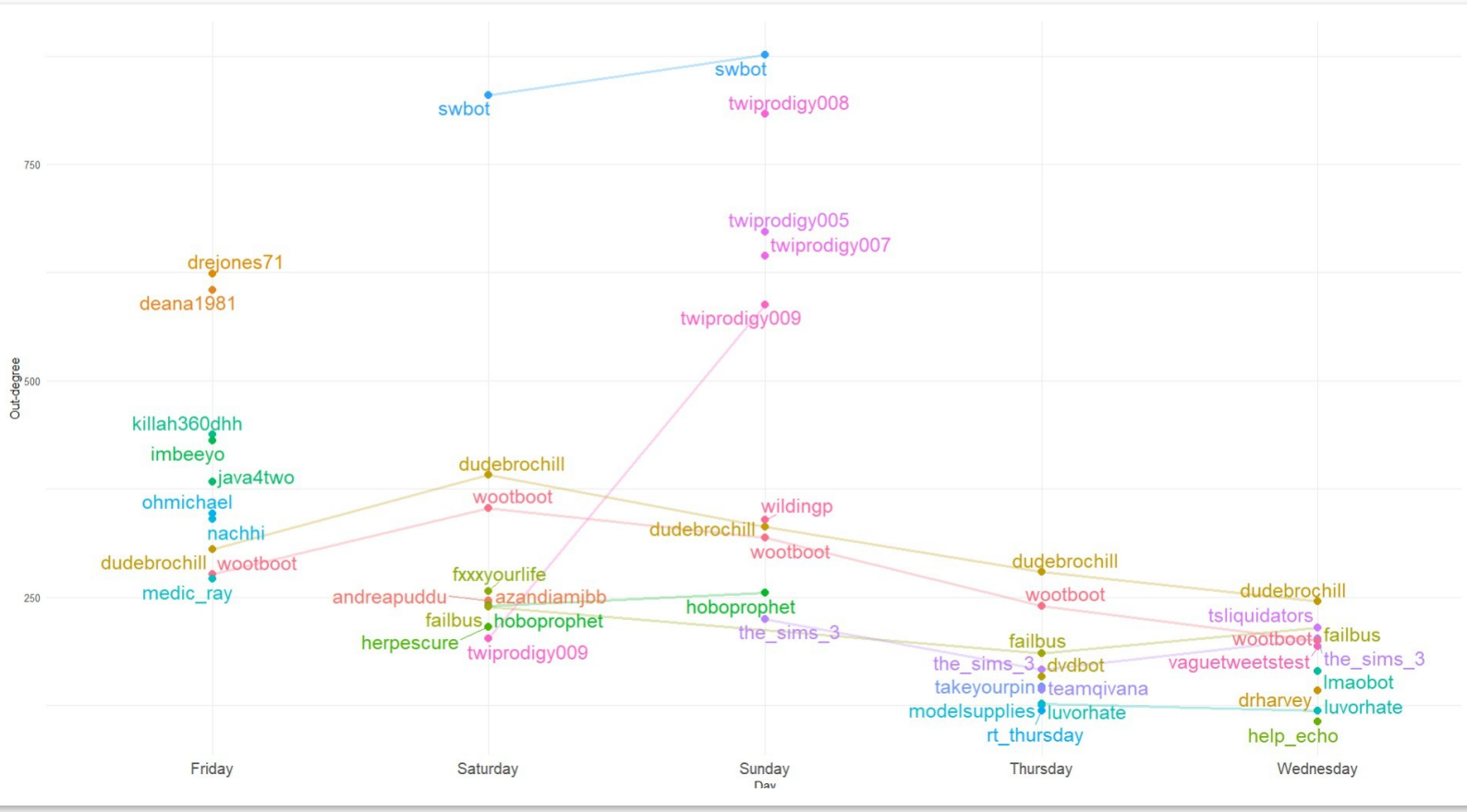
Note: The lines of the following plots are very important for the analysis. Nodes that are in the top ten continuously, for more than one day, will be connected with a line. Also, The points helps us identify nodes that are not in the top 10 for more than one day of the week.

The following plot is created to have a better understanding of how the in degree for the top 10 is changing thought time for, each day. The plot has on the x axis the days of the week that we have in our disposal and on the y axis the in-degree. It shows, for the top ten based on in degree, how they change over the time. We find that some nodes are in top 10 list for many days of the week. More specifically the tweetmeme, mashable, milicyrus, breaking news, add this, mootweet are most of the time present in the top 10. Other nodes (that do not have lines) are nodes that appear only for a particular day on the top 10. Such as ddvato and davidmmasters. For the trend analysis, we find that tweetmeme is descending as the week progresses. The same can be said for the mashable. In other cases such as breaking news we find that there is no big variation as the week progresses.



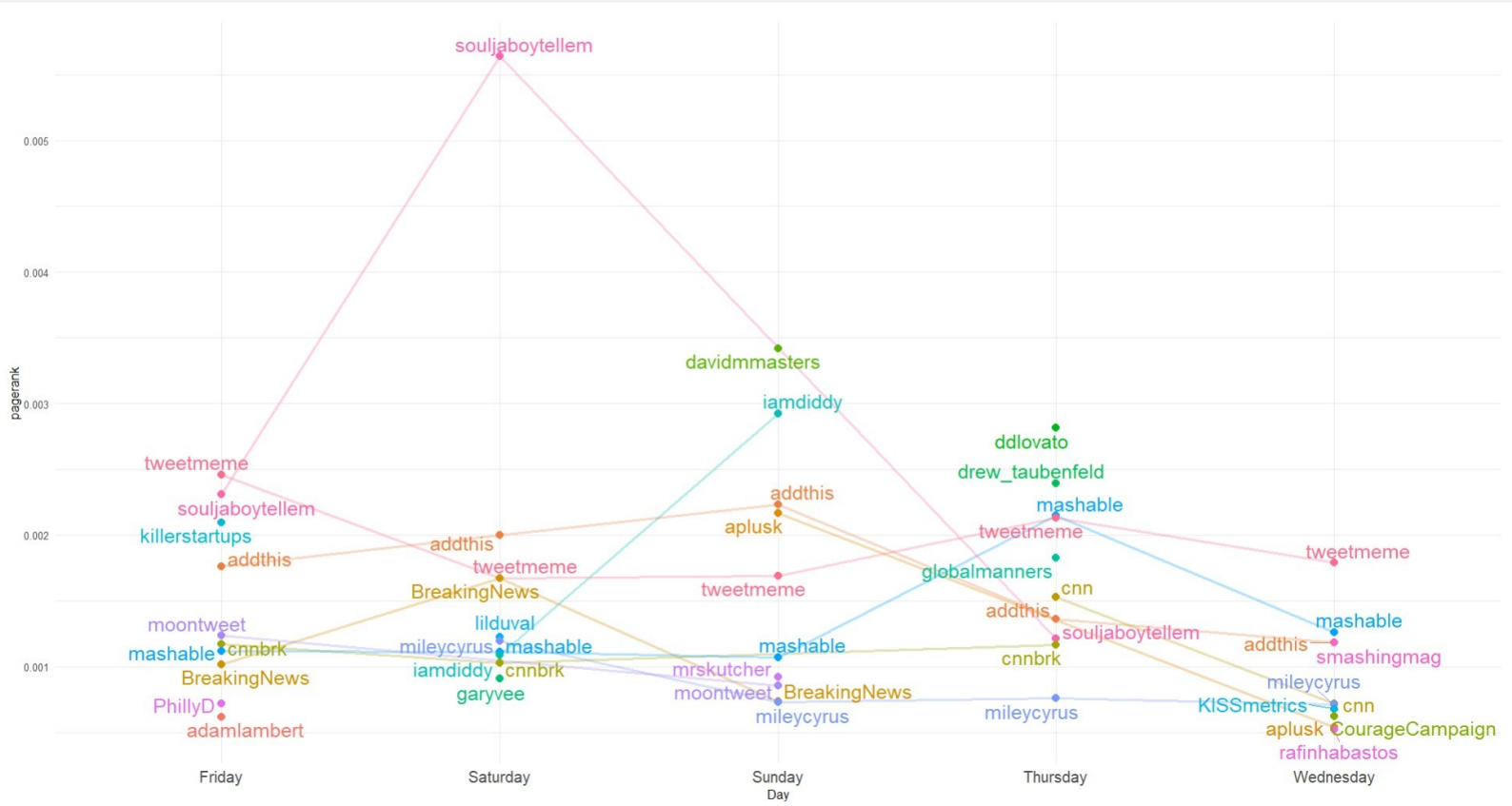
Out Degree

The following plot is created to have a better understanding of how the out degree for the top 10 is changing thought time, for each day. The plot has on the x axis the days of the week that we have in our disposal and on the y axis the out-degree. As can be seen with the use of out degree we have many nodes that are not in the top 10 continuously. Such nodes are, deana1981, drejone71, medic_ray and many others. But we also have some nodes that appear to be on the top 10 list continuously. Nodes such as, swbot, dudebrochill, wootboot. For the trend analysis we find mostly not a big variation for those nodes that are continuously in the top 10.



PageRank

The following plot is created to have a better understanding of how the PageRank for the top 10 is changing thought time, for each day. The plot has on the x axis the days of the week that we have in our disposal and on the y axis the PageRank. As can be seen, the souljaboytem, the addthis, breaking news, iamdiddy, tweetmeme mileycyrus, mashable and other nodes are found in the top 10 for most of the weeks. Nodes such as, davimmasters, ddlvato, PhillyD and others are being found in the top 10 but only in particular days.



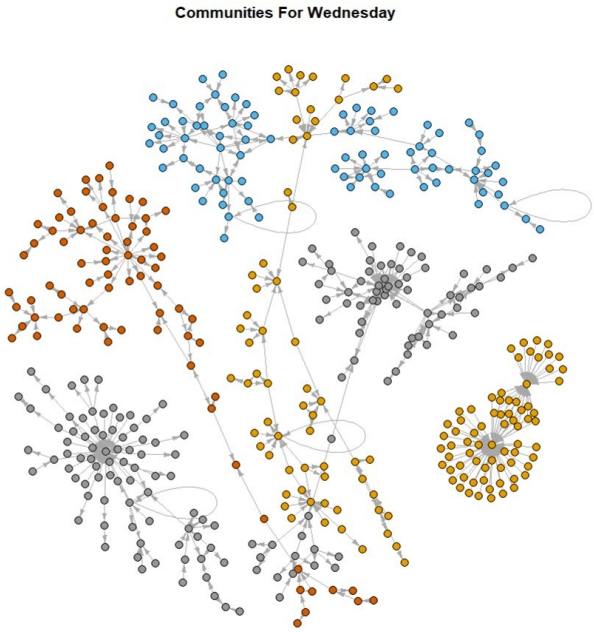
Communities

After running the 3 algorithms for community detection we found the following. The fast greedy algorithm took, for each graph, approximately 20 minutes to run. The infomap clustering had to be abandoned since it took more than 1 hour to run. Finally, the Louvain clustering which is known for scalling good with large graphs run extremely fast. It took only 10 seconds to run for each graph.

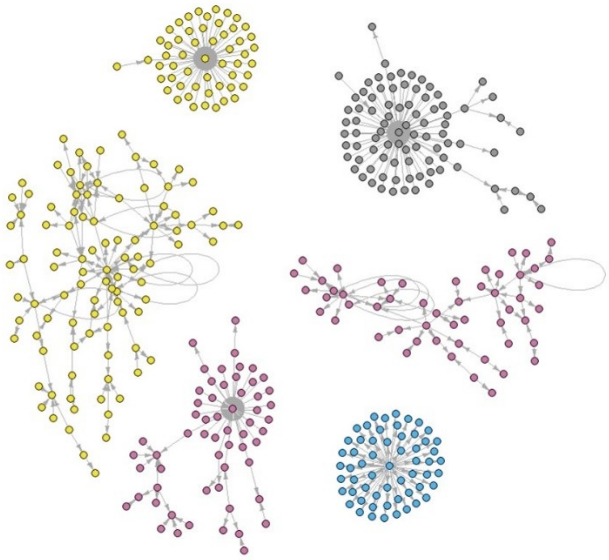
For a random user we picked ‘aaronkaiser’ and tracked his evolution with the use of Louvain algorithm and found that in these 5 different days he was found in different communities. More specifically, at Wednesday he was found in the community with id 34, at Thursday in the community with ID 573, at Friday in the community with ID 20, at Saturday in the community with ID 571 and at at Sunday in the community with ID 83.

To find similarities between the communities of the user with name aaronkaiser we calculated the average weight for the verteces that appear in each of the communities(that aaronkaiser appear) for each graph. The overall average weight is 1.20 and the following average weights for the subgraphs were found. The first average weight is 1.15, the second 1.16, the third 1.06, the fourth 1 and the fifth 1.16. We also found the topic for each of the communities. Based on the topics we found that there is no similiraty, between the communities to which aaronkaiser appear.

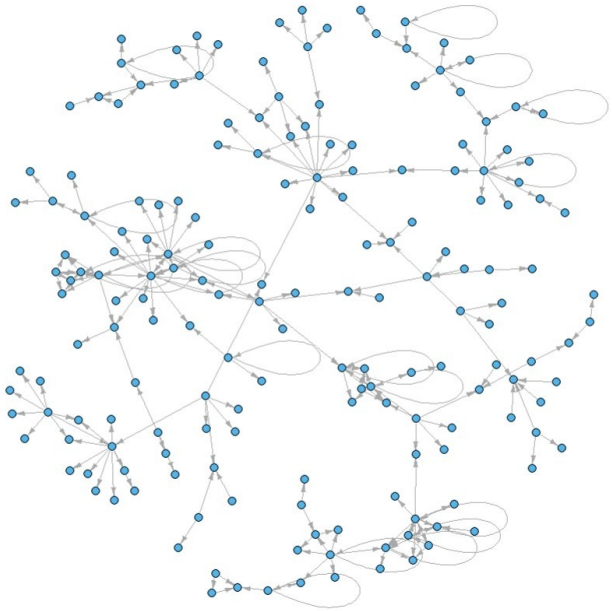
To plot our graphs we took a sample from each graph and plotted them.



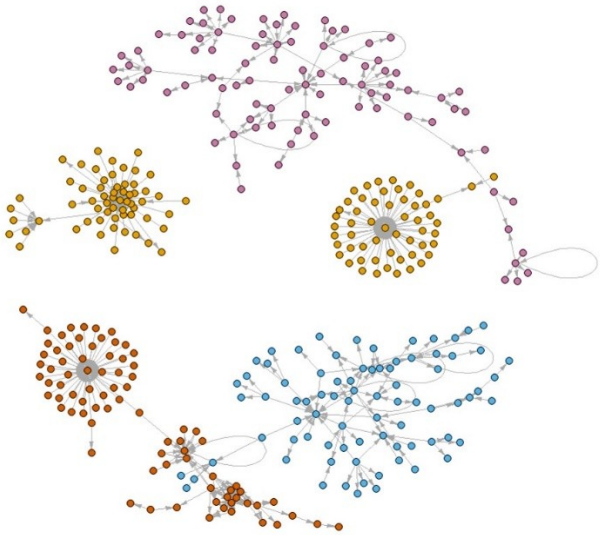
Communities For Thursday



Communities For Friday



Communities For Saturday



Communities For Sunday