Algorithmic Methods of Data Mining
Homework 4

Davide Aureli, Michela Bragagnolo, Dastagiri Dudekula

December 22, 2017

PART 1

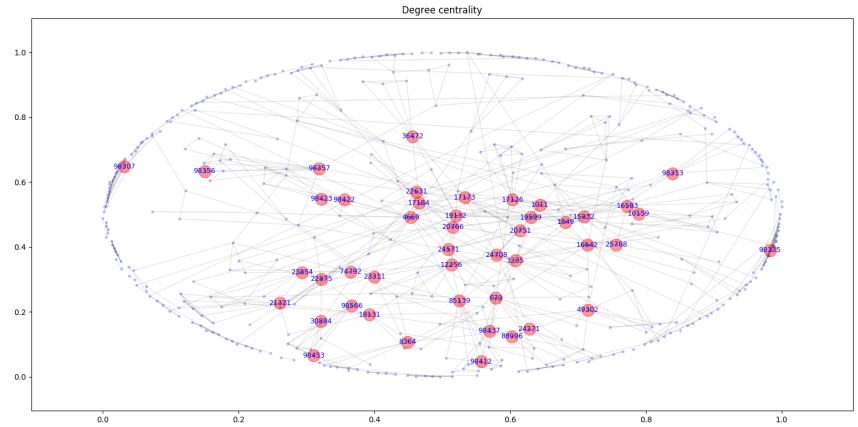
In this point we have to create a graph G, whose nodes are the authors and they are connected with weighted edges if they share at least one publication. On the full dataset we obtained a disconnected graph with 904664 nodes and 3679242 edges. The weight of the edges is higher if the nodes are different, which means that the authors don't share a lot of publications. On the other hand, if they contributed to the same publications the similarity is higher and the weight of the edge is lower.

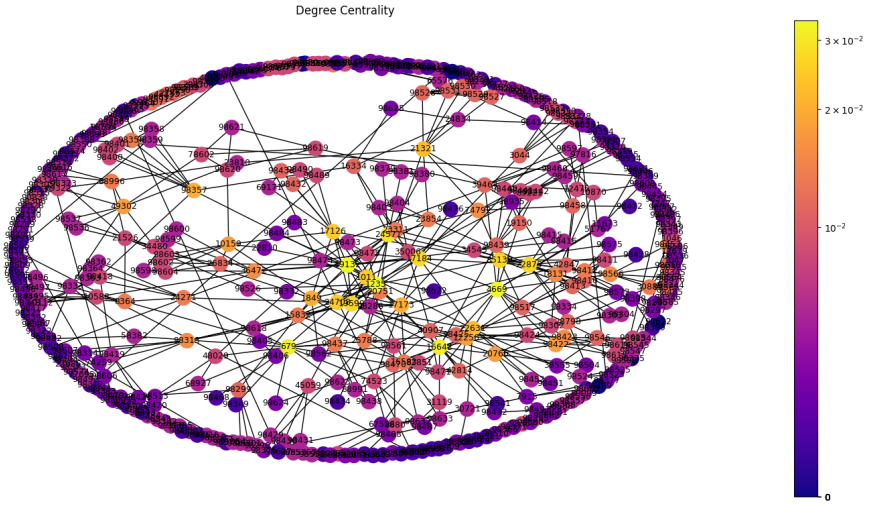
PART 2

2.A

Given a conference in input, we need to return the subgraph induced by the set of authors who published at the input conference at least once. After obtaining the subgraph we can calculate some centrality measures, which give us relative measures of importance in the network.
To show the results of these measures we used as example in input the conference 'conf/iitsi/2010'.
Associated to this conference we obtained a disconnected subgraph with 447 nodes and 726 edges.

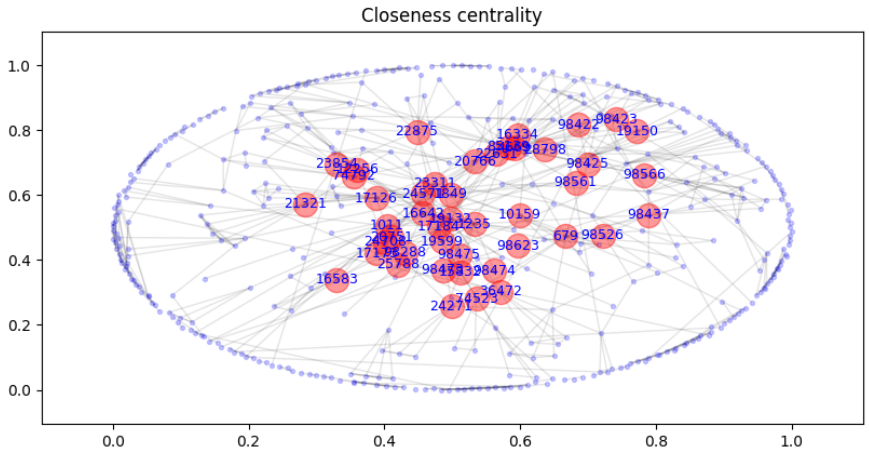
The first measure we consider is the *degree centrality*; the intuition is that nodes with more connections are more influential and important in a network. The highlighted nodes in the next graph corresponds to the most relevant authors.

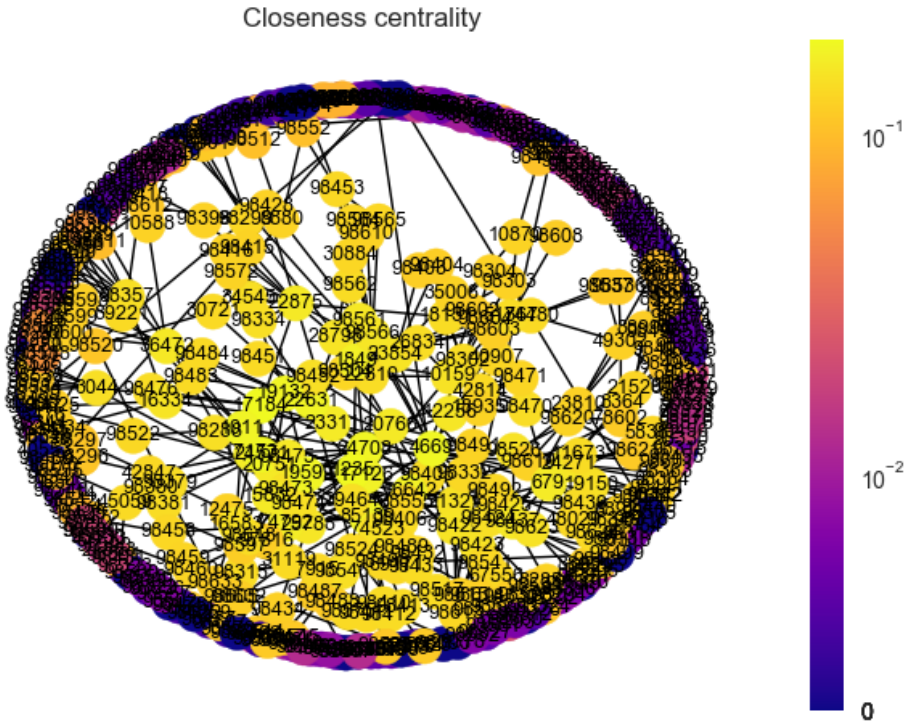




On the second figure we can identify better the main authors, characterized by yellow nodes, such as: Ying Wang (id:1235), Wei Li (id:19132) and Hui Li (id:4669).

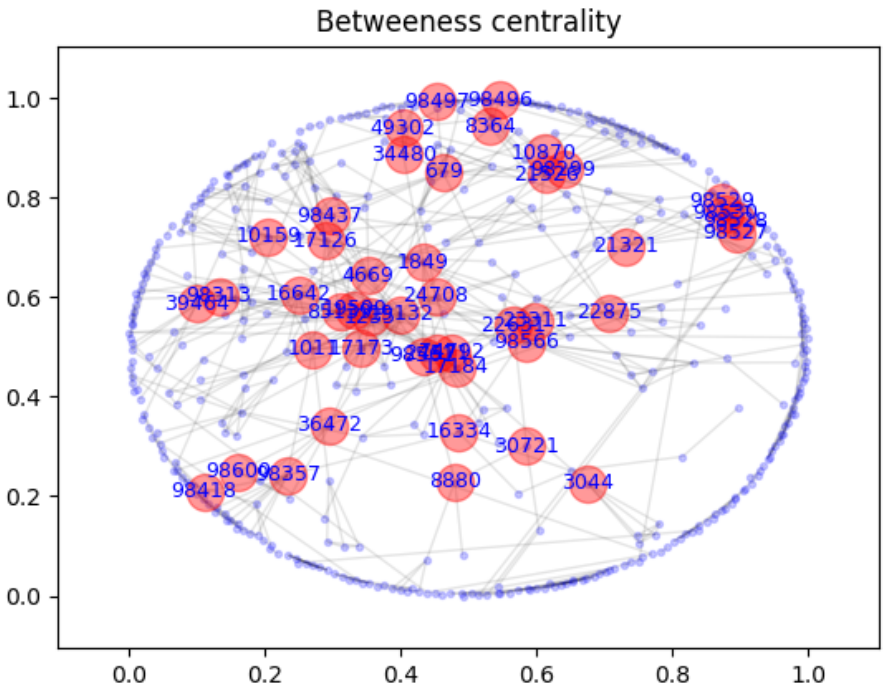
The *closeness centrality* is a measure where each node's importance is determined by closeness to all other nodes.

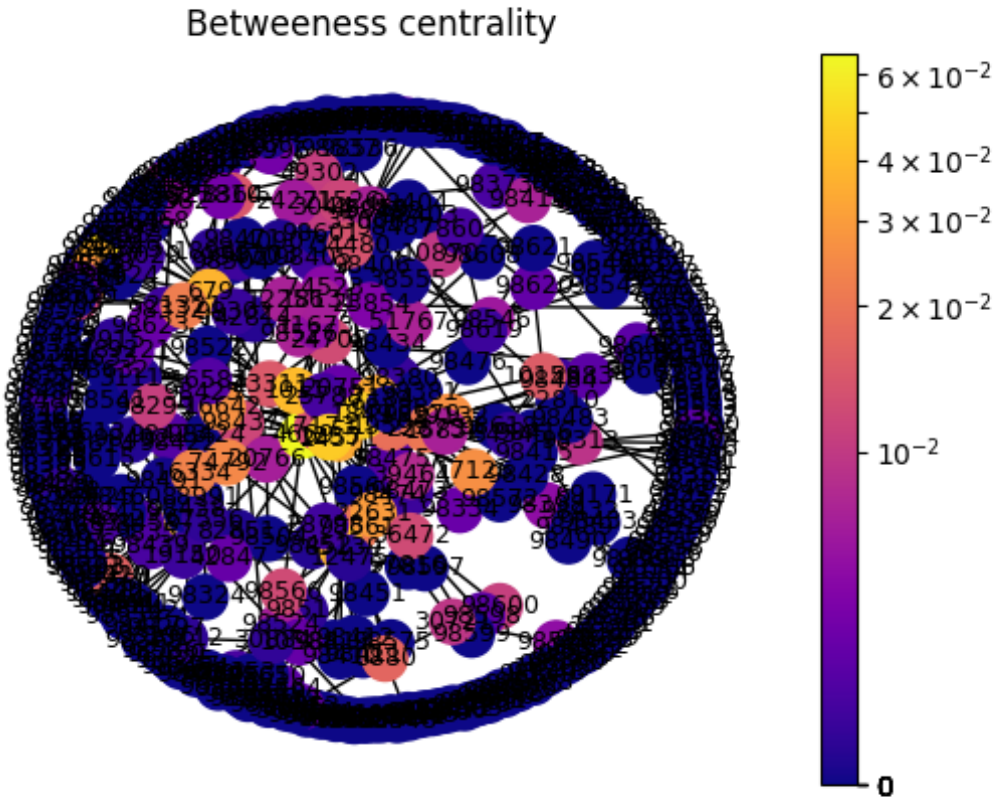




As we can see on the figures, all the most important nodes, with high closeness, are centered in the middle of the graph.

The last measure is the *betweeness centrality*, where vertices with high betweenness may have considerable influence within a network by virtue of their control over information passing between others. They are also the ones whose removal from the network will most disrupt communications between other vertices because they lie on the largest number of paths.

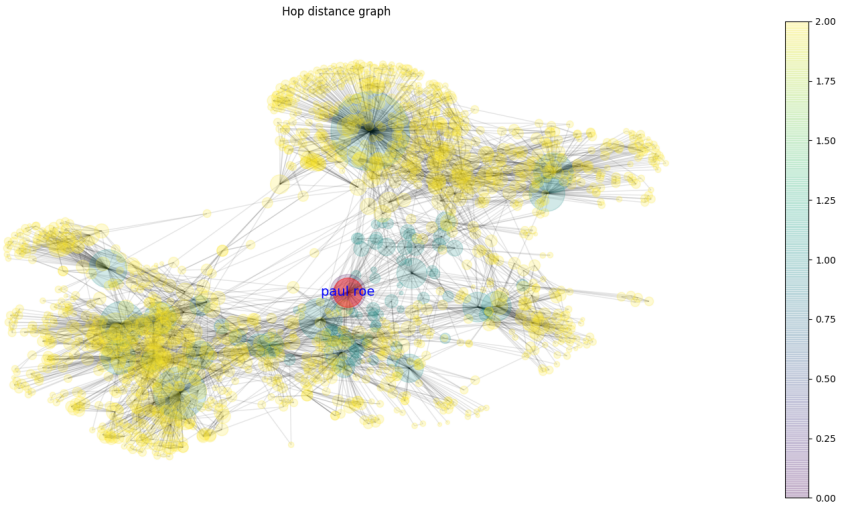




The most influent nodes are situated approximately in the middle, where there are more connections, as we can notice with the huge contrast in the color of the nodes. The most important author in that case seems to be Hui Li with id 4669.

2.B

Given an author and a certain distance we want to obtain a subgraph, from the main graph obtained before, where we consider just the nodes connected to the node of the artist given in input with a certain distance. For example, if we want to see the connection of the author Paul Roe until level 2, we obtained the following plot.



The red node highlighted is our author and the other nodes are colored based on their distance, at distance 1 we have green nodes and at distance 2 we have yellow nodes. Also the size of the nodes reflects the number of connections of each node, more connections it has more bigger is the representation.

PART 3

3.A

This is a kind of generalization of Erdos number, describing the collaborative distance between authors.
We need to take in input an author (id) and returns the weight of the shortest path that connects that input author with Aris. We provide here an example considering the same author as before: Paul Roe.
The weight of the shortest path between Aris an this author is equal to 3,729 but if we consider an author, such as Nicola Ricci, the weight become 7,401, so this author is more distant. We found also that the $12,425\%$ of nodes cannot be connect with Aris.

3.B

The aim of this point is to calculate the GroupNumber:

$$GroupNumber(v) = min_{u \in I}\{ShortestPath(v, u)\}$$

for each node v in the graph, given a set I of input nodes u.
The first step to obtain the GroupNumber is to find the shortest paths between each node v of the graph and each node u in the given set. Once found them, we can get the desidered number. For example, if we use as list of input nodes (176994, 24151, 9741, 21462) and for simplicity we take 6 random nodes from the graph, the result we should obtain is the following:

This node 895419 cannot reach any node of the list.
The shortest path for node 895420 is destination : 9741 with the cost: 6.415319865319866
The shortest path for node 895421 is destination : 9741 with the cost: 7.190516811461665
The shortest path for node 895416 is destination : 176994 with the cost: 6.086420629639118
The shortest path for node 895417 is destination : 176994 with the cost: 6.560370215041854
The shortest path for node 895425 is destination : 9741 with the cost: 6.751091129527354