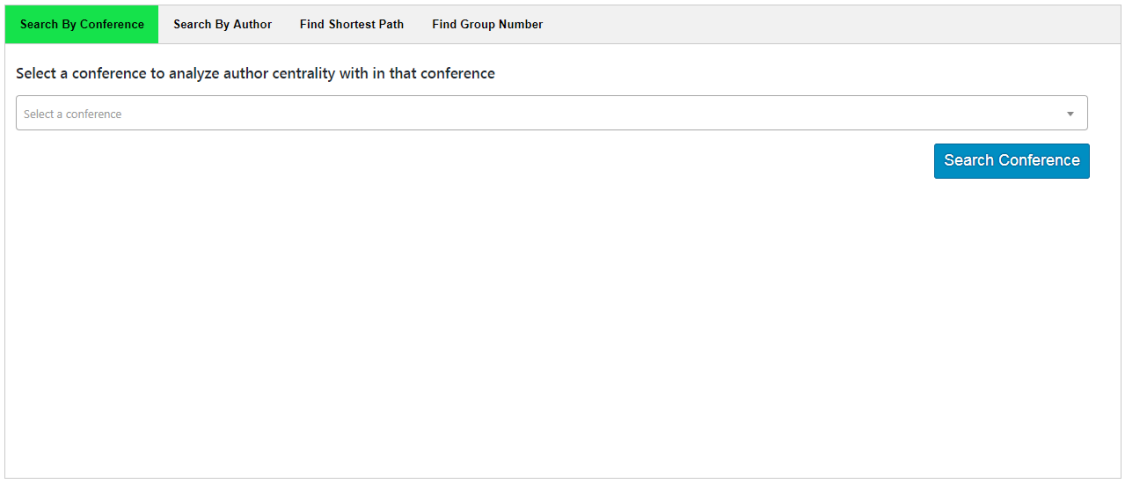
Algorithmic Methods of Data Mining
Homework 4

Group 3 – Vamsi Krishna Varma Gunturi, Matteo Manzari, Andrea Ferrara

To better visualize our hard work ☺ done by our group, we decided to create a small web application (an idea of Vamsi). The web application run well with reduce file in input but we haven't time to optimize for the full file. Below a screenshot:

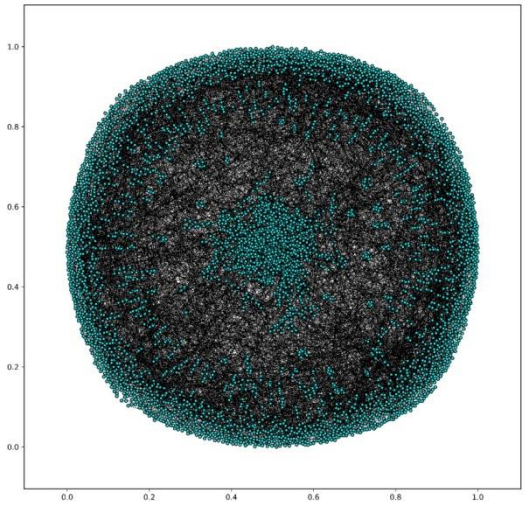


There are 4 sections that refer (in order) to points 2.a, 2.b, 3.a and 3.b and that points will be explained below.

Part 1 – By processing Json file create a graph G

A graph G= (V,E) consist of a set of nodes V, and a set of edges E. In our case we define as nodes the authors and the edges as weighted that is compute by 1- J(p1,p2) or 1- Jaccard similarity between two sets of pubblications thats "are contained" in the authors.

In our case we considered that graph G undirected and we assume that (u,v) and (v,u) are the same edge.The representation of the graph (reduced file json in input) is shown below:



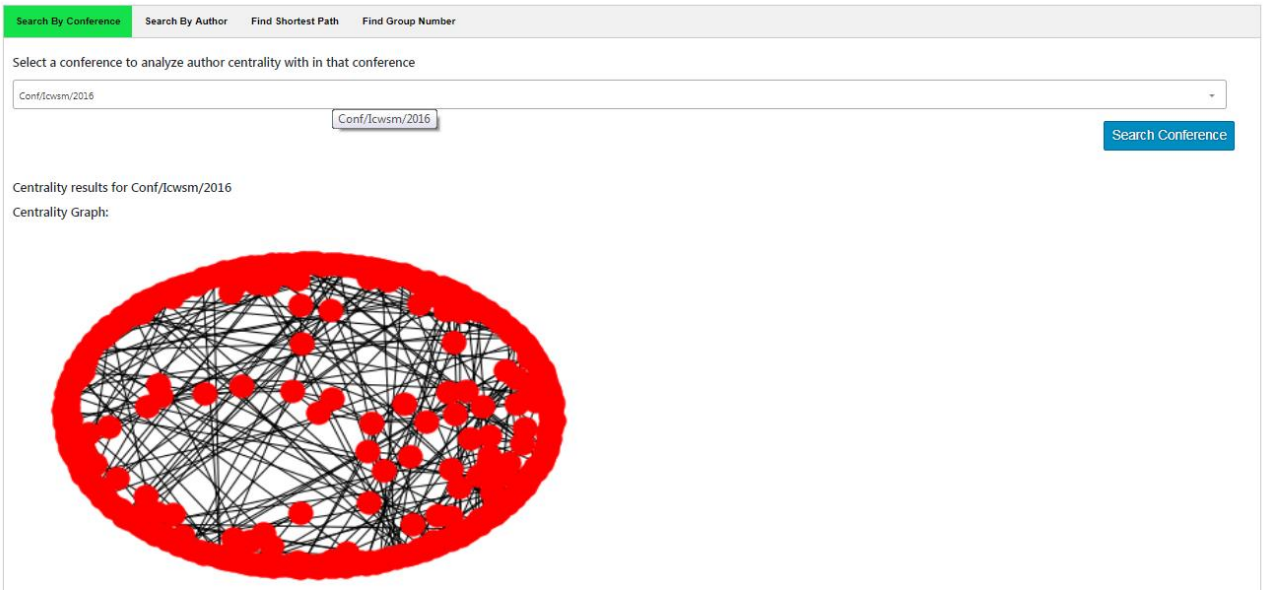
Part 2 – Compute some statistics and visualizations

2.a

Centrality measures how central is the node in the graph. Depending on what we mean by "central", there are versions of centrality measure.

1) Degree centrality:
   is a simple centrality measure that counts how many neighbors a node has.
2) Closeness centrality:
   measures the mean distance from a vertex to other vertices. Intuitively a node is central if its distance from the other nodes is small.
3) Betweeness centrality:
   measures the extent to which a vertex lies on paths between other vertices.

In the web application "Search by Conference", below is an example for the conference "Conf/Icwsm/2016"



and its table with the results of centrality measures:

Centrality Table:

Show 10 entries

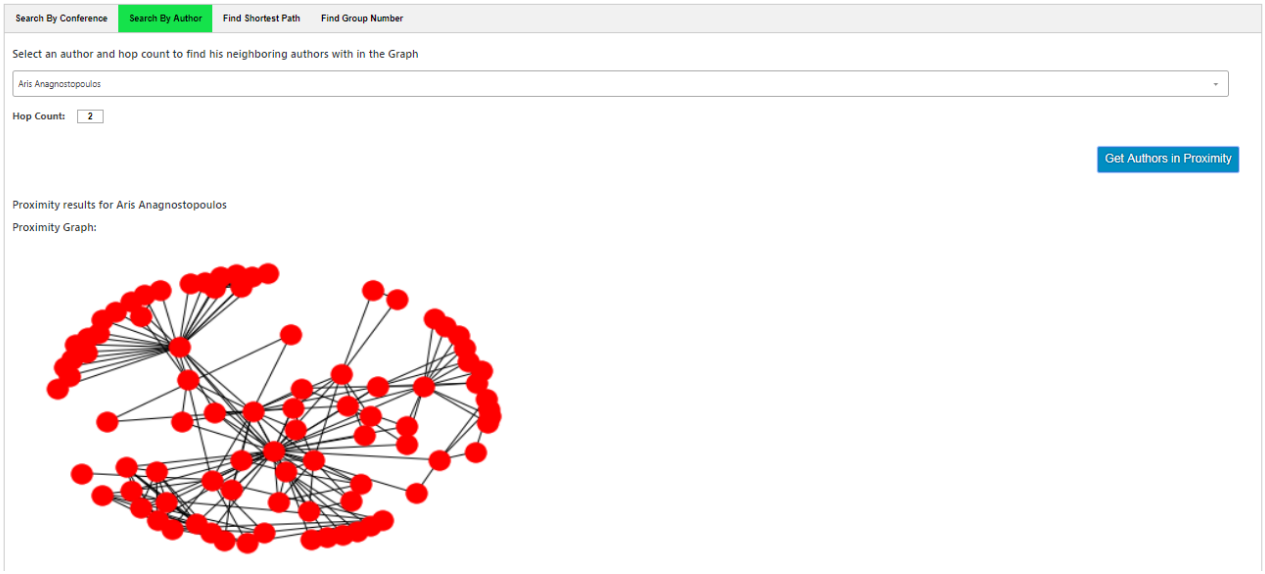
Search:

Author ID	Author Name	Degree Centrality	Closeness Centrality	Betweeness Centrality
322	Yizhou Sun	0.014164305949008499	0.030823221005454315	0.0014808138037599795
8289	VicenÇ GÓMez	0.0056657223796034	0.012086874409820586	0
8513	Huan Liu	0.0311614730878187	0.03623080363799016	0.0025860588891750366
8718	Lyle H. Ungar	0.0113314447592068	0.0113314447592068	0
8988	Jiawei Han 0001	0.014164305949008499	0.04214603688500896	0.0037181303116147315
10591	Gareth Tyson	0.0084985835694051	0.0084985835694051	0
11044	Dah Ming Chiu	0.0056657223796034	0.0056657223796034	0
11073	Benjamin D. Horne	0.0113314447592068	0.0113314447592068	0
11074	Sibel Adali	0.0113314447592068	0.0113314447592068	0
11379	Yipeng Zhou	0.0056657223796034	0.0056657223796034	0

Showing 1 to 10 of 351 entries

Previous 1 2 3 4 5 … 36 Next

2.b In the web application if you select an author and the number of the hop count you can visualize this (in this example we have selected our fantastic professor ☺):



And its table whith the measure of the distances between Aris and another authors:

Proximity Table:

Show 10 entries

Search:

Author ID	Author Name	Distance From Aris Anagnostopoulos
287	Ricardo A. Baeza-Yates	1.778181818181818
15628	Luca Maria Aiello	1.835
15924	Rakesh Agrawal 0001	1.7857142857142856
16323	Salvatore Orlando 0001	1.8171428571428572
16617	Matteo Riondato	0.9473684210526316
16618	Eli Upfal	0.9473684210526316
17528	Gerhard Weikum	1.7833333333333332
18263	Aristides Gionis	0.8
18910	Michael Mathioudakis	1.6333333333333333
19210	Gabriele Tolomei	1.8171428571428572

Showing 1 to 10 of 80 entries

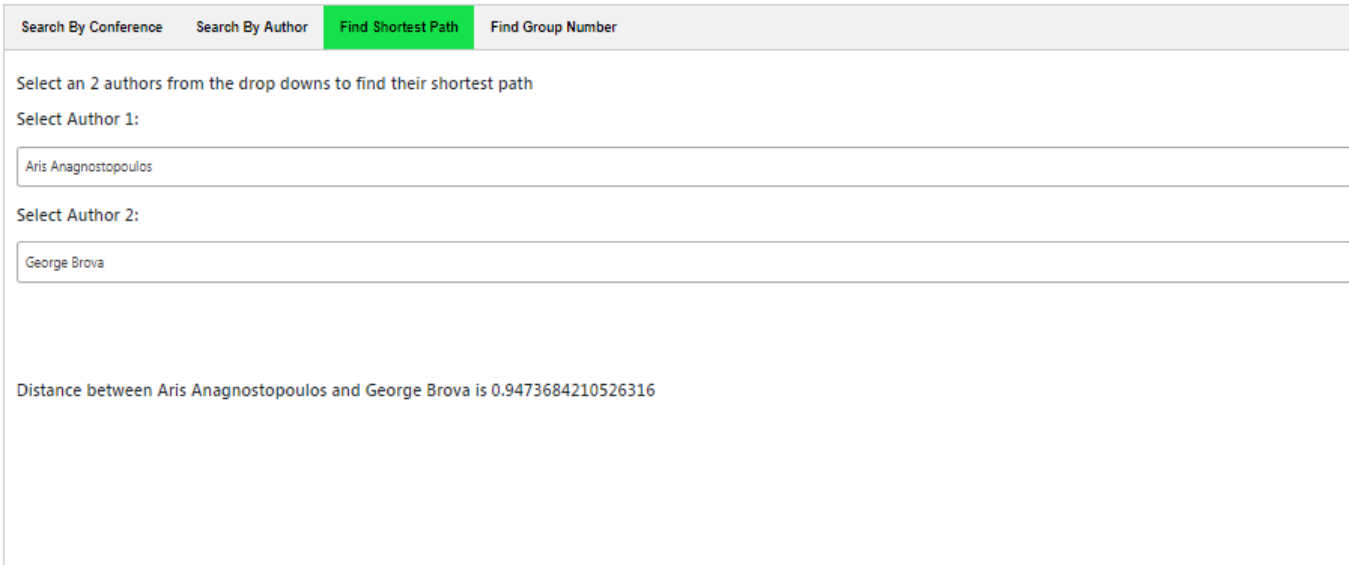
Previous 1 2 3 4 5 ... 8 Next

Part 3 – Compute some generalized version of the Erdos number.

a) To compute "Aris number" we have implemented from scratch the algorithm to compute the shortest path. So we have followed the logic of the Dijkstra's Algorithm.

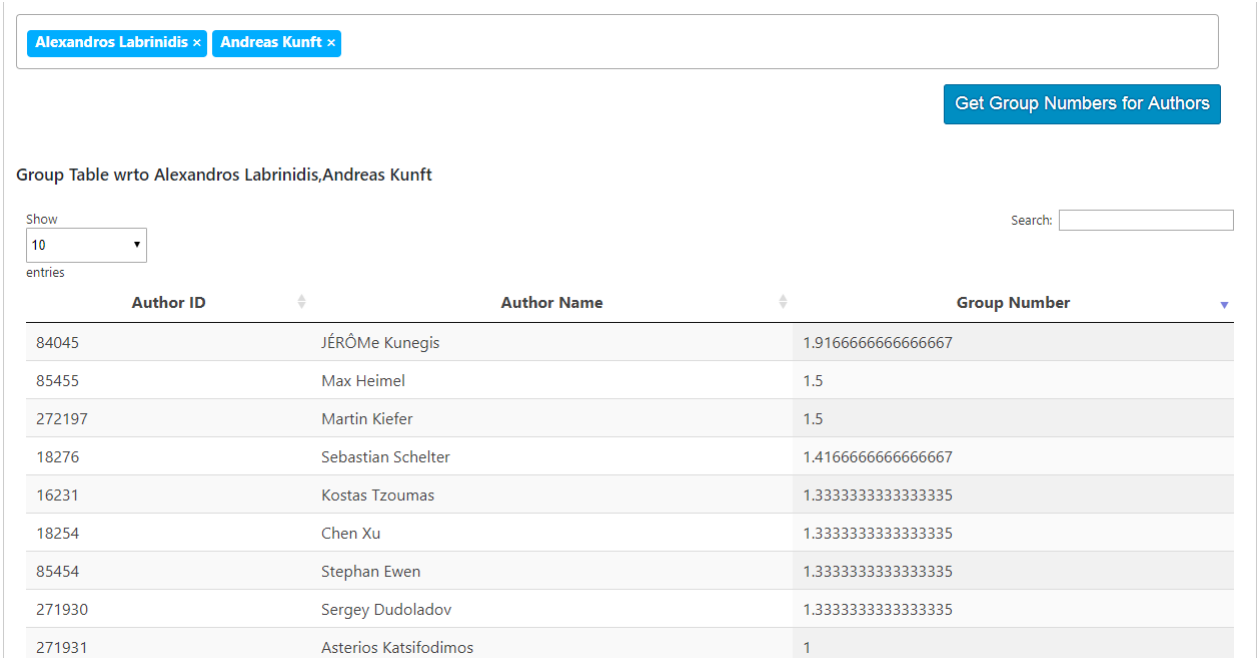
   In the web application you can see this in the "Find Shortest Path".

   For instance we compute the weight of the shortest path that connects "George Brova" and Aris and the result is:



b) To compute group number with respective to set of authors selected,
In the web application you can see this in the "Find Group Number" tab. User is provided a multi select drop down where in he can select multiple authors based on which group number of every node is calculated,

For instance we compute the group number of every author present in the graph based on author subset, Alexandros Labrinidis, Andres Kunft and the result is:



Author ID	Author Name	Group Number
84045	JÉRÔMe Kunegis	1.9166666666666667
85455	Max Heimel	1.5
272197	Martin Kiefer	1.5
18276	Sebastian Schelter	1.4166666666666667
16231	Kostas Tzoumas	1.3333333333333335
18254	Chen Xu	1.3333333333333335
85454	Stephan Ewen	1.3333333333333335
271930	Sergey Dudoladov	1.3333333333333335
271931	Asterios Katsifodimos	1