

CNT 5805 Final Project Preliminary Report

I. Project Team number - 9

Project Title - Collaboration Insights Between Network Scientists

II. Project Member Names:

- a. Jainam Shah
- b. Jalique Francis
- c. Dylan D' Andrea
- d. Matthew Olajide

III. Briefly describe the data you want to analyze:

The dataset contains a network of collaboration between scientists who are working and providing research on networks around the globe. This network was compiled from bibliography of multiple articles by M. Newman and by S. Boccaletti. The edges/vertices of the network contain the names of the scientists who are authors of research papers, and the edge represents the names of authors who appear on the same research paper. The dataset has a network which consists of 1589 nodes, 2742 edges and is an undirected network. Over the course of project multiple entities like scientists can be added in the dataset in order to build clusters and make the network more vast.

Data Source - <http://vlado.fmf.uni-lj.si/pub/networks/data/collab/netscience.htm>

IV. Explain your motivation for analyzing this data: Why would you want to analyze this data? Do you have experience in this area of study?

The primary motivation to analyze this data is to see the structure and get insights how the group of scientists collaborate with each other and perform research in areas related to networks and their formations. The subject of network is a vast area which comprises a lot of multiple fields. Numerous scientists and scholars work on this field and provide information to get a clearer picture of networks. How they collaborate with other network scientists as well to research on topics related to network, its integrations, formations, and varied areas within the field and what insights can be gained from this data set, how network scientists interact, which countries do scientists belong to, their gender, their knowledge based on how much they have contributed to the network field etc. can be known through visualization of this network. The social network analysis of co-citation analysis on network science presented during class solely explores citations and has far smaller network than the dataset that will be analyzed for this project.

Note:- The dataset does contain metadata about some authors but not all, it would need a little cleansing for research part and addition of one or two columns within the dataset to answer research questions.

V. Describe what specific research questions your analysis will address.

- Are there preferred scientists or outlier fewer collaborative scientists? The number of scientists forming a group or collaborating with multiple scientists for research, as the network when visualized through “Gephi” presents some disconnected nodes which indicates certain scientists collaborating individually.

(For example: - in the network, there are around 30% scientists contributing individually, they are not connected with other scientists and are totally isolated.)

- Are there any distinct social circles or communities that exist within the network? For instance, Are there any communities that are closely grouped and has maximum output for networking field.? Which country of origin for the scientist’s provides maximum insight into the field of network?

(For example: - there is a specific community of scientists which has collaborated with maximum scientists, or there is a community with maximum contribution in the network field, approximately there are 40% of scientists from United States contributing to this field or a percentage of scientists from a specific country contributing maximum etc.)

- Are there any social hierarchies within the scientists' collaboration? Are there any Systems of social organization in which some individuals enjoy a higher social status than others? Can we find if there’s a repeating theme with one individual within some groups, then that person is higher on the social hierarchy?

(For example: - there is specific scientists who has maximum connections in the network, or a community of scientists having maximum connections in the network)

- Which is the predominant gender of the scientists that contributes the most to the networking field?

(For example: - around 70% of scientists contributing to the network field are females or males)

- What are the hubs of scientists who have contributed the most in the field of network science?

(Scientists forming larger communities in the network, the one with common connections, for example: - around 40% of the scientist in this network have connections with other communities in the network.)

- Are there any scientists who don’t collaborate with anyone and provide their own research in the field of network?

(Scientists contributing individually, and the scientists who are collaborating individually are either male or female, for example- out of scientists contributing individually 80% are females)

VI. Describe any data cleansing you plan to do.

None at the moment (Should this change we will edit this answer) but We might add multiple columns related to countries, gender of the scientists etc. to provide more insights within the dataset.

VII. It is okay, but I want to know if you working with any entities outside (e.g., professors, businesses, universities, labs, governments) our classroom on this project? Please explain.

At this moment in time, we are not working with any entities outside of the classroom, however should we potentially decide to ask for insight from various scientists on their perception of other data scientists from previous collaboration, we can do so.