Visualization of Social Networks

By: **(CSE - C)**

Shashanka Venkatesh	-	18 5001 145
Vaibhav Sankaran	-	18 5001 186
Venkataraman Nagarajan	-	18 5001 192
Vishakan Subramanian	-	18 5001 196
Vishnu K Krishnan	-	18 5001 200

Dataset & Viz Tool

Dataset

The dataset that we visualize here is the famous Citation Network presented by the researcher **Stanley Milgram**, a psychologist at Yale University.

The dataset is modelled with other **researchers as nodes** and the edge weight between a node $\mathbf{N}_i$ to $\mathbf{N}_j$ represents the **number of citations** that author i has made to author j.

The dataset has a total of **231 nodes** & **1329 edges**.

The data is present as 2 CSV files: **nodes.csv** & **edges.csv**

Visualization Tool

The visualization tool that we selected for performing social network analysis was **Gephi**, which is an open graph visualization platform. Gephi provides a **highly-interactive UI** with powerful SNA tools to properly visualize large social networks and supports a wide variety of dataset formats like **GML**, **GraphML**, **NET** etc.

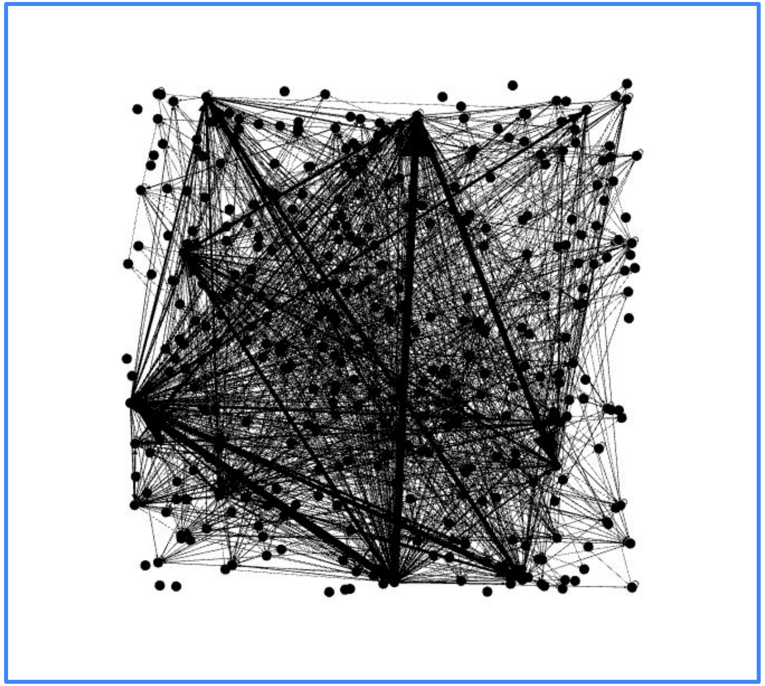
The tool was written in Java using NetBeans. It is very popular among social network researchers.

Initial Visualization

Initial Loading

The social network upon plainly importing the dataset files looked like this:

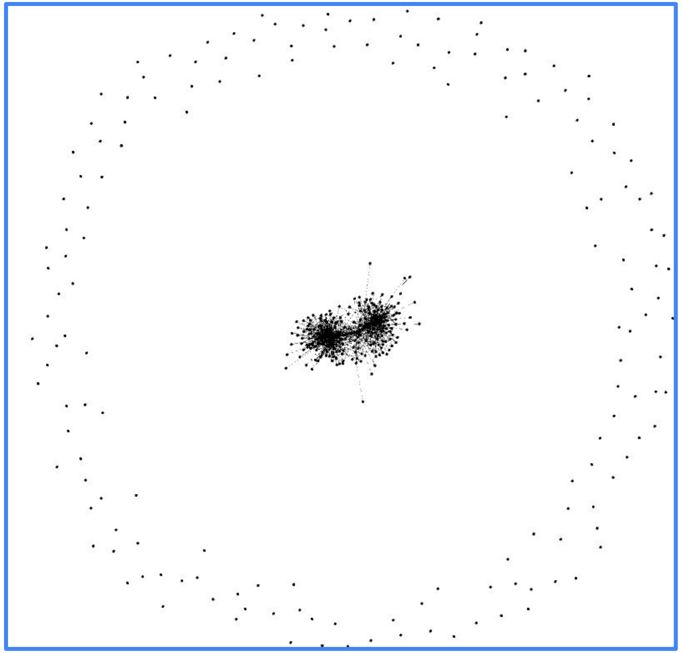
The structure is very cluttered and the layout needed to be changed in order to view the network in a better manner.



Yifan-Hu Layout

The Yifan-Hu layout was used to transform the social network. It is a very fast algorithm that has good quality on large graphs. The algorithm uses a combination of a multi-level approach and a force-directed approach.

The repulsive forces on one node from a cluster of distant nodes are approximated by a Barnes-Hut calculation, which treats them as one super-node. It stops automatically.

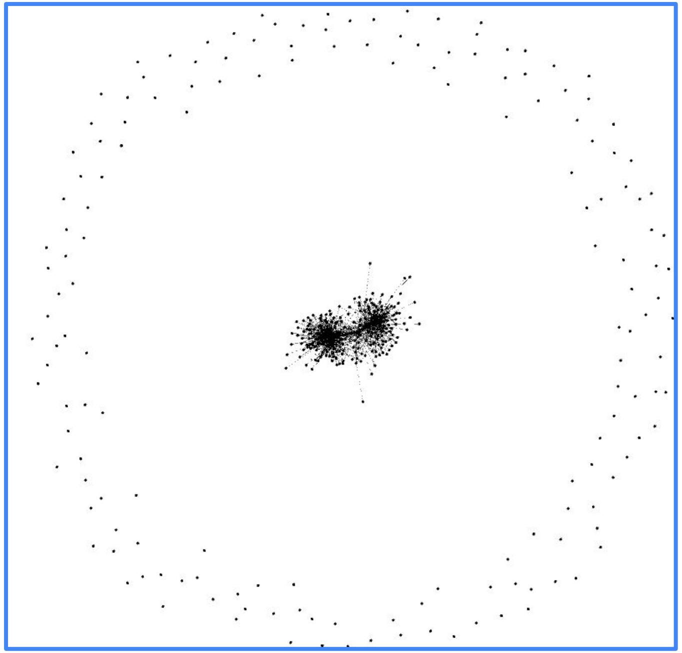


Yifan-Hu Layout - Implication

Here, the algorithm effectively separates out a core network from its periphery.

The periphery nodes are authors that have cited another author's paper but possess no other interesting attributes apart from them.

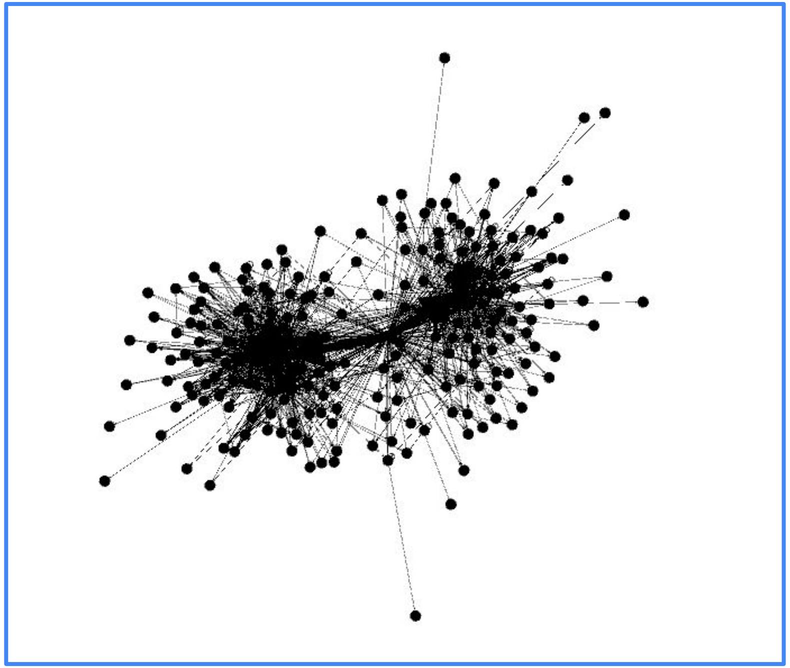
We can remove them from analysis as they're not a major part of this network.



Yifan-Hu Layout - Core Structure

The core network structure was obtained by filtering off the periphery nodes.

This was accomplished by adding a filter: **Degree Range**, with a **minimum degree = 1**.



Filtering

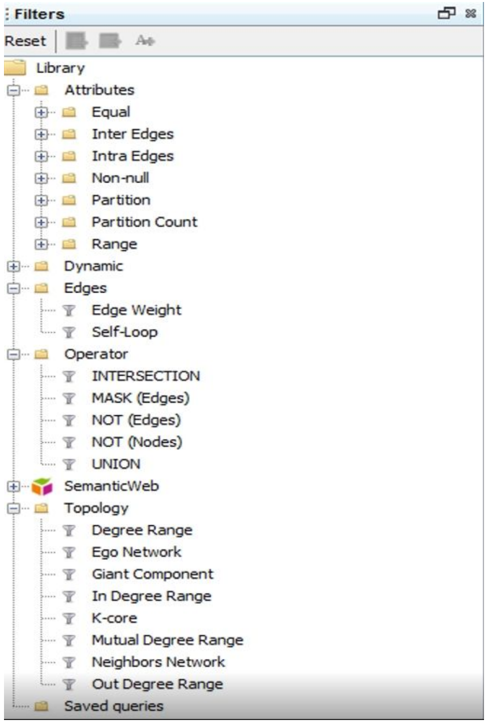
Filtering functions in Gephi

Gephi filters are categorized into several groups, shown as individual folders in the Filters tab.

Within each of these folders are multiple filtering selections that can be used on their own as **simple filters**, or combined to create **complex filters**. Primary filter categories include:

1. Attributes: enable filtering on nodes, edges, partitions, clusters, and various graph measures
2. Edges: strictly for the connections within the network
3. Operator: executes a few functions on the graph.
4. Topology: range of graph measures like degree ranges that can alter the topology of the network.

We will focus on Topological filtering for our use-case.



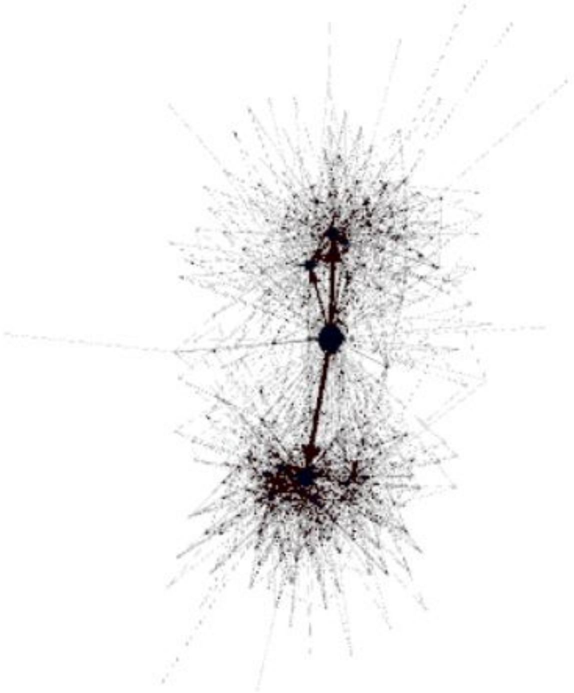
- Library
  - Attributes
    - Equal
    - Inter Edges
    - Intra Edges
    - Non-null
    - Partition
    - Partition Count
    - Range
  - Dynamic
  - Edges
    - Edge Weight
    - Self-Loop
  - Operator
    - INTERSECTION
    - MASK (Edges)
    - NOT (Edges)
    - NOT (Nodes)
    - UNION
  - SemanticWeb
  - Topology
    - Degree Range
    - Ego Network
    - Giant Component
    - In Degree Range
    - K-core
    - Mutual Degree Range
    - Neighbors Network
    - Out Degree Range
  - Saved queries

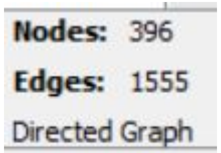
Filters (Topology) - Giant Component

A giant component is a connected component of a network that contains a significant proportion of the entire nodes in the network.

Typically as the network expands the giant component will continue to have a significant fraction of the nodes.

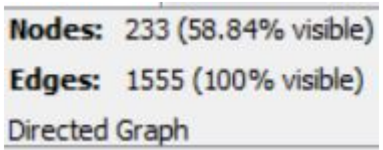
The graph obtained is the result of filtering by the giant component filter.





Nodes: 396
Edges: 1555
Directed Graph

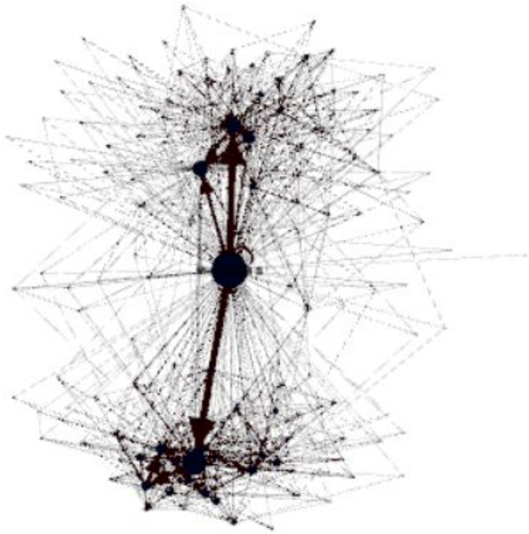
→



Nodes: 233 (58.84% visible)
Edges: 1555 (100% visible)
Directed Graph

Filters (Topology) - Ego Network

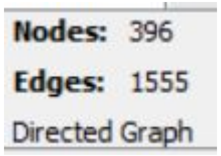
Egocentric networks are local networks with one central node, known as the “ego” which is surrounded by “alters” (other nodes directly connected to the ego).



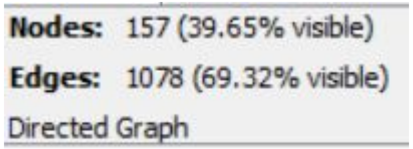
The network shown has been derived by applying the **Ego Network** filter with the parameters:

Ego: Milgram, Depth: 1, self: true

The Ego was chosen in such a way that the peripheral nod were filtered off.



Nodes: 396
Edges: 1555
Directed Graph



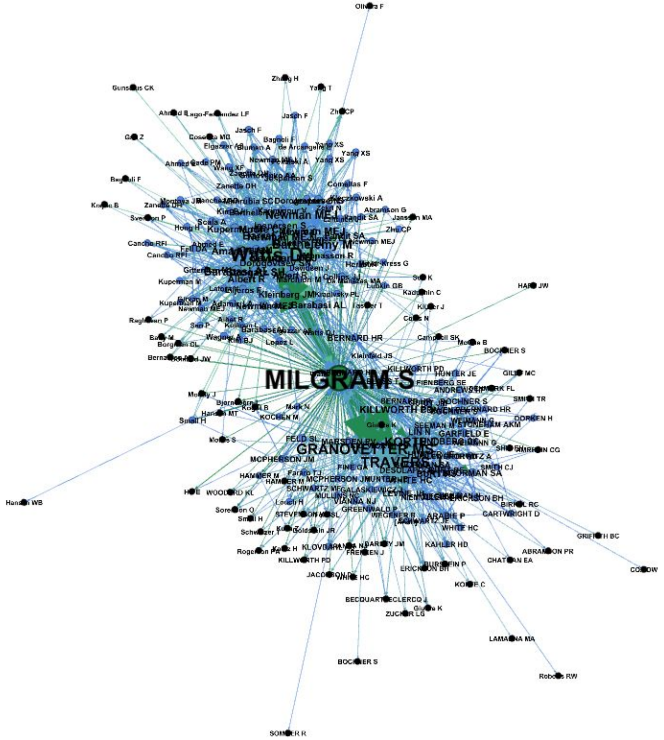
Nodes: 157 (39.65% visible)
Edges: 1078 (69.32% visible)
Directed Graph

Filters (Topology) - K-core

The K-Core of a graph is the maximal subgraph, such that each node in that subgraph has a degree of **at least K.**

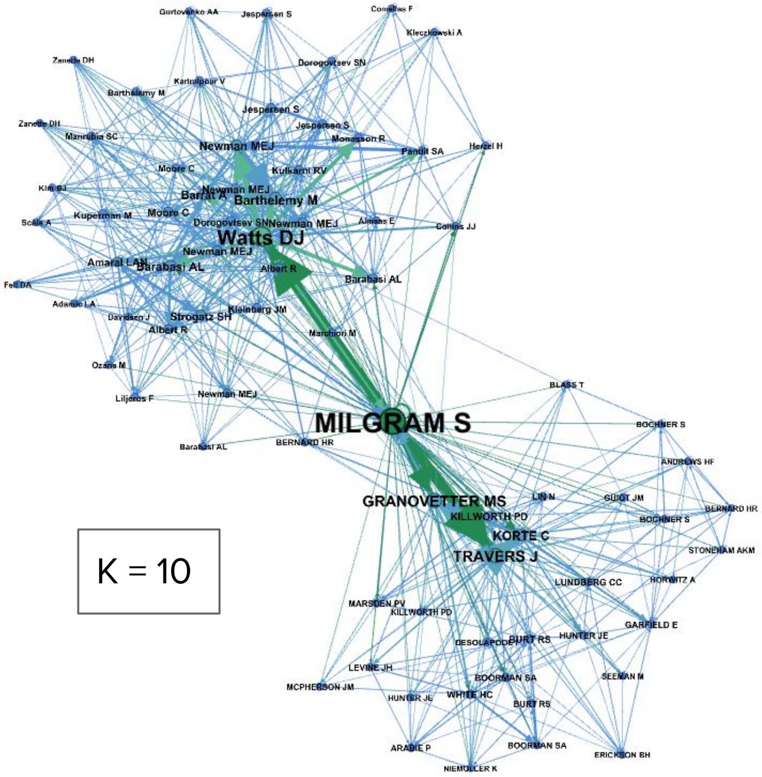
A K-Core filtering of K=1 simply removes all nodes with a degree less than 1. Beyond that, the process of removing nodes becomes recursive, in order to ensure that the resulting subgraph has all nodes with at least degree = K.

This kind of filtering allows us to understand the core nodes of the network, the nodes that are highly connected to other nodes. This allows us to understand which nodes are the most important in the network.



Setting K = 1 removes nodes with degree = 0
Giving us this core network of nodes.

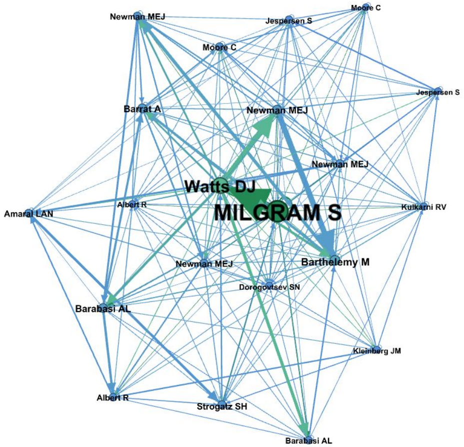
Trying out Higher values for K



K = 10

Setting **K = 10** reveals a highly interconnected core of the network. Each researcher's citations + the number of times they have cited others in this core is at least 10. They are of greater importance to the study than the others.

We decided to try finding the maximum possible value before all nodes are filtered out, and that turned out to be **15.** That is the maximal interconnected subgraph of this network, with the highest degree count.



K = 15

Analysis

Network Diameter

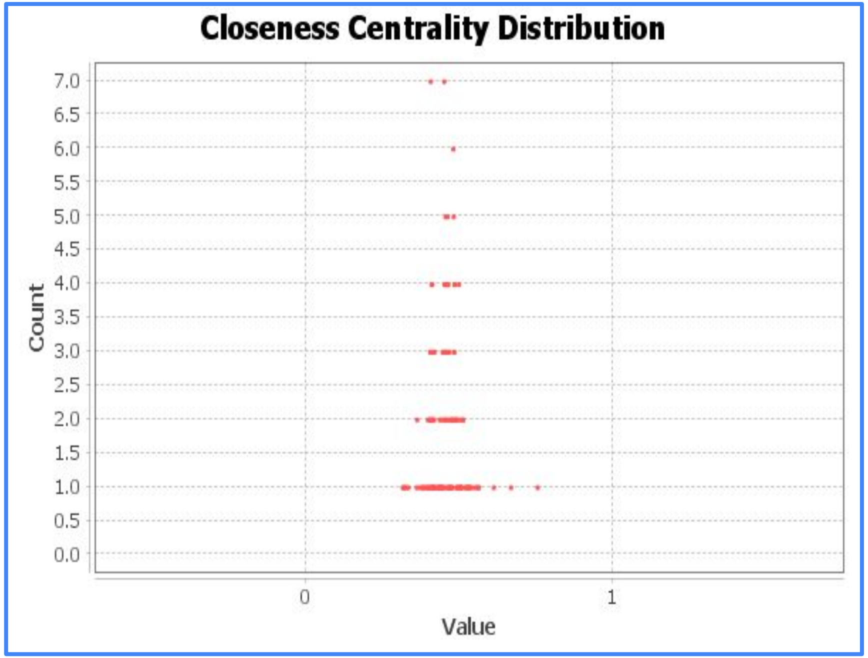
Considering the previous core structure for further analysis, we performed an analysis of the network's diameter properties.

The edges were interpreted as undirected connections.

Diameter = **4**

Radius = **2**

Average Path Length = **2.23**

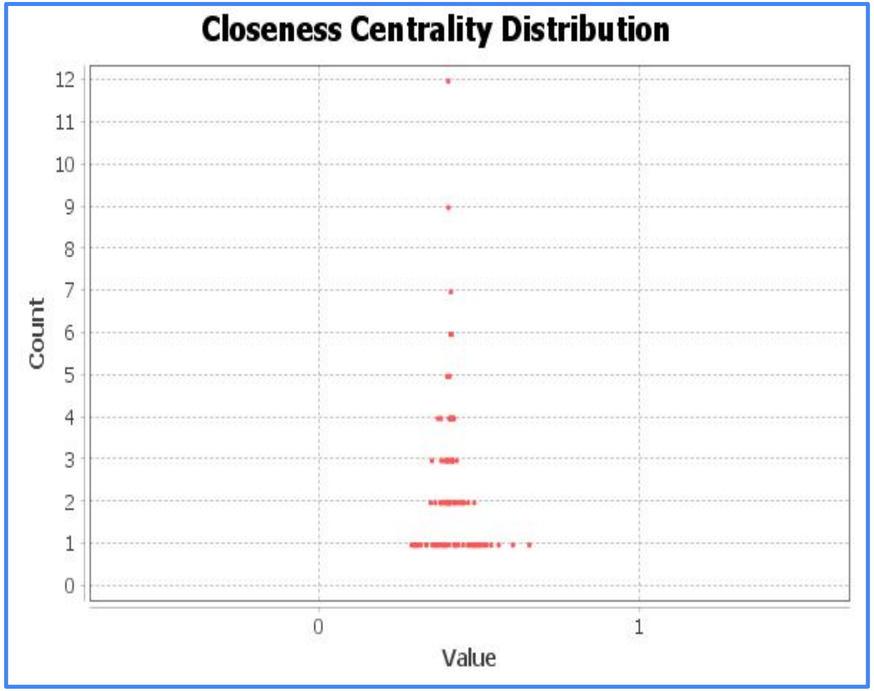


Weighing the Nodes - I

The node sizes were redrawn proportionally with respect to each node's **betweenness centrality measure**.

The picture on the right represents the obtained visualization.

As we can see, there are clearly **two separate clusters** with **one giant central node**.

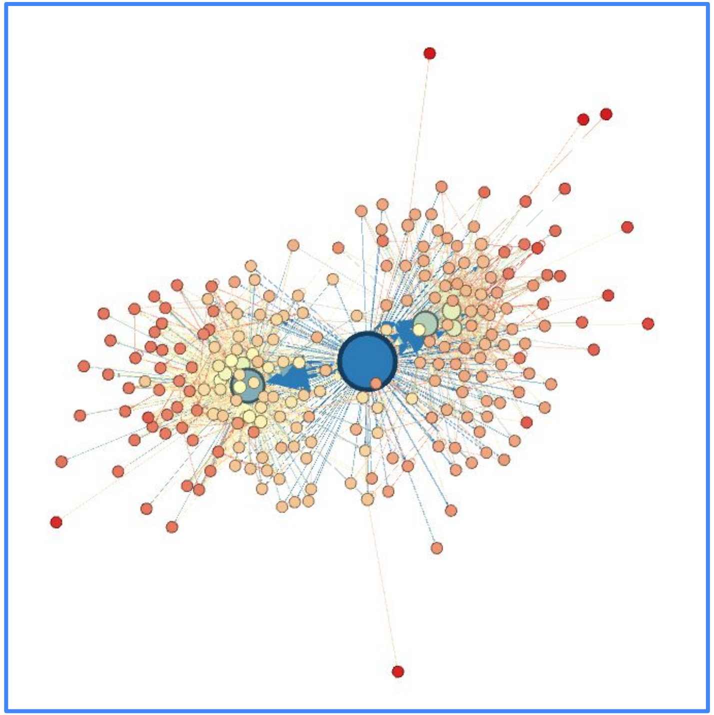


Weighing the Nodes - II

Furthermore, each node was assigned a color based on their **closeness centrality measure.**

The color palette moves from red to yellow to blue. (lowest to highest)

Now, we can see that there are 2 central nodes in the two separate clusters and a giant blue central node that has a lot of connections to other nodes.

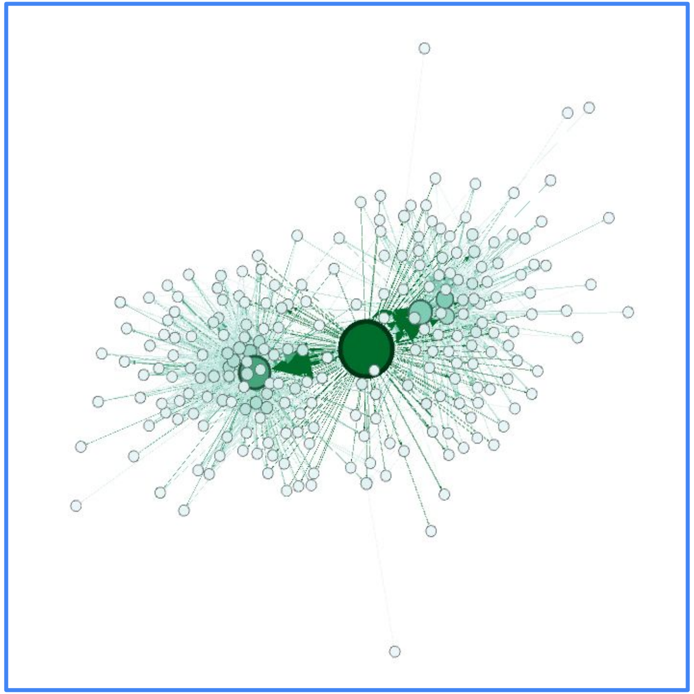


Weighing the Nodes - III

Now, each node was assigned a color based on the number of edges that the node possessed.

The color palette moves from white to green. (lowest to highest)

Once again, we can clearly see a strong central node connected to all other nodes - implying that this author was cited by most other authors in this domain.

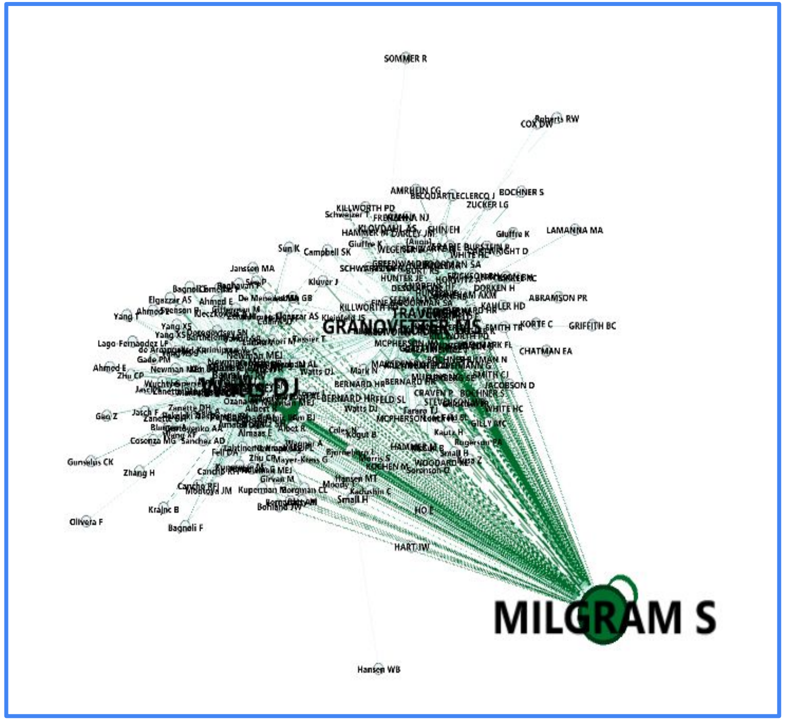


Labelling the Nodes

We turned on labels for all the nodes to better understand the social network.

The central big node was also dragged out to the side to understand the other structure properly.

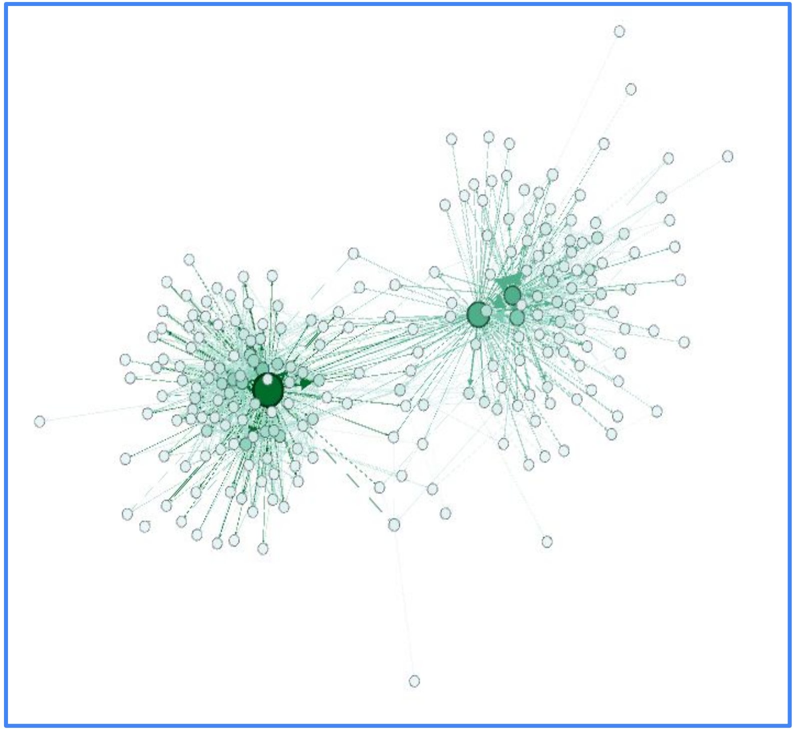
This central node was **Stanley Milgram** himself, and it made sense why he was a central point to this citation network - the dataset was constructed by him.



Removing a Node

Stanley Milgram's node is an uninteresting link in this social network as we already know that this network is about his citation connections.

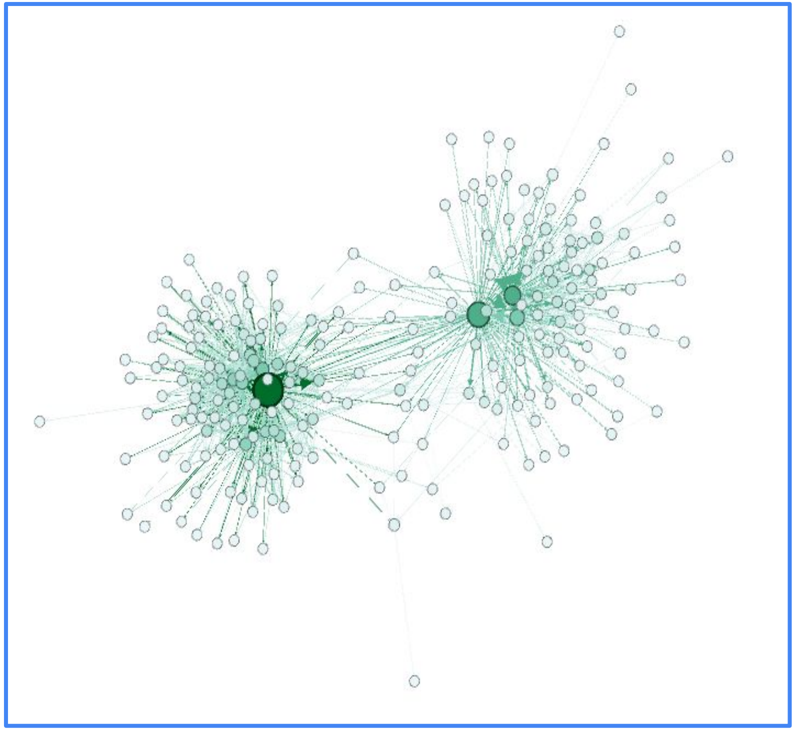
Thus to understand the network better, we removed Milgram's node from the picture and then redrew the node layout with the Yifan-Hu procedure and the Expansion procedure to zoom in.



Analysis - Without Milgram

Recomputing the Metrics

The Network Diameter, Node Size, Node Colors need to be recomputed and redrawn because of the removal of Milgram's node.



Network Diameter

Considering the previous core structure without Milgram's node for further analysis, we performed an analysis of the network's diameter properties.

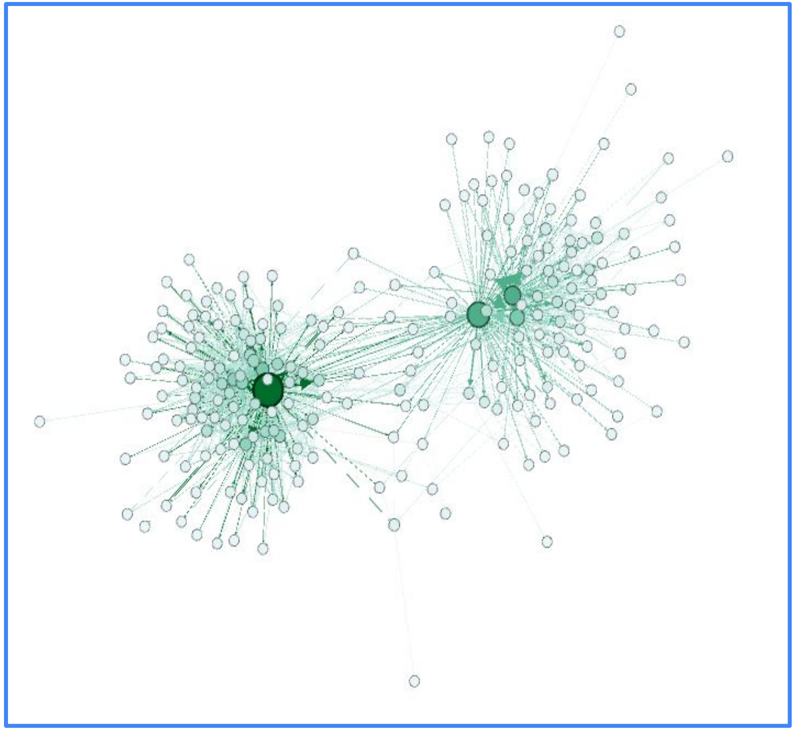
The edges were interpreted as undirected connections.

Diameter = **5**

Radius = **3**

Average Path Length = **2.49**

Clearly, we can see that the absence of Milgram's node has affected the structure a lot, since the metrics have increased significantly.

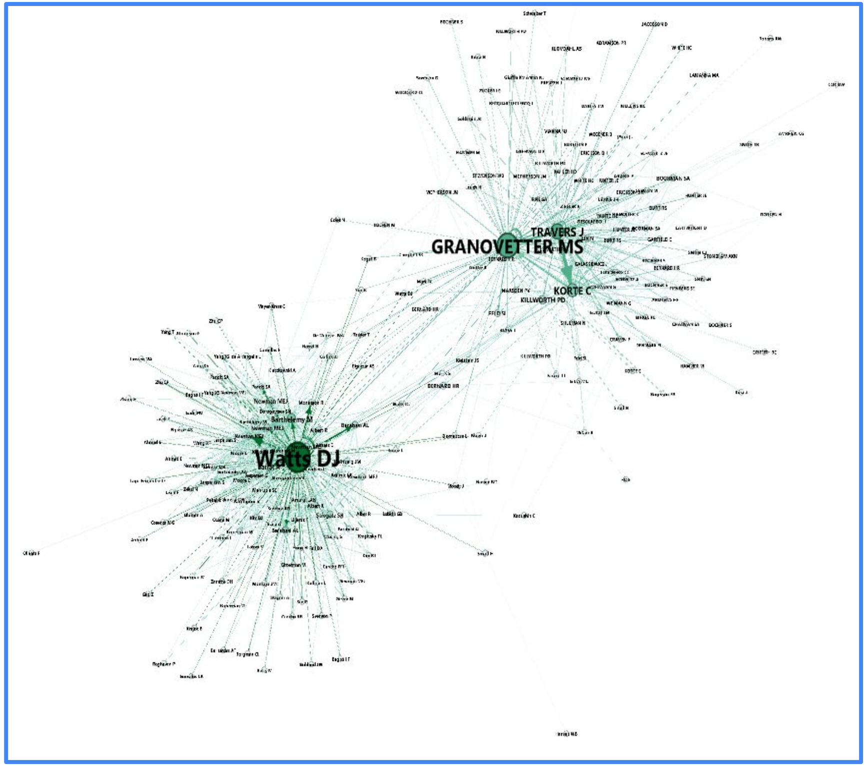


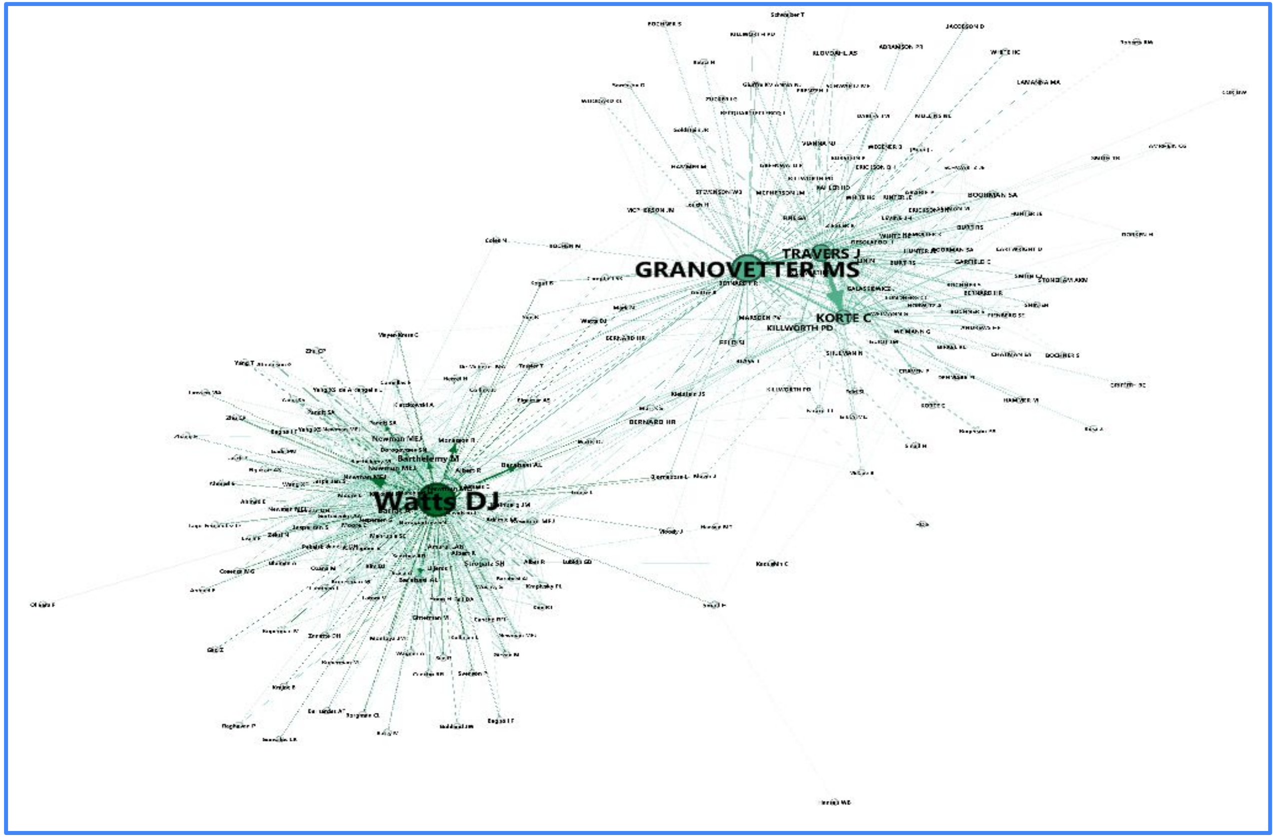
Redrawn Layout

The network was redrawn using the **Yifan-Hu** layout and then zoomed in using the **Expansion** layout.

Node sizes were proportionately assigned based on **betweenness centrality** measure & **node colors** were assigned based on their **number of degrees**.

Node labels were also drawn in the layout.



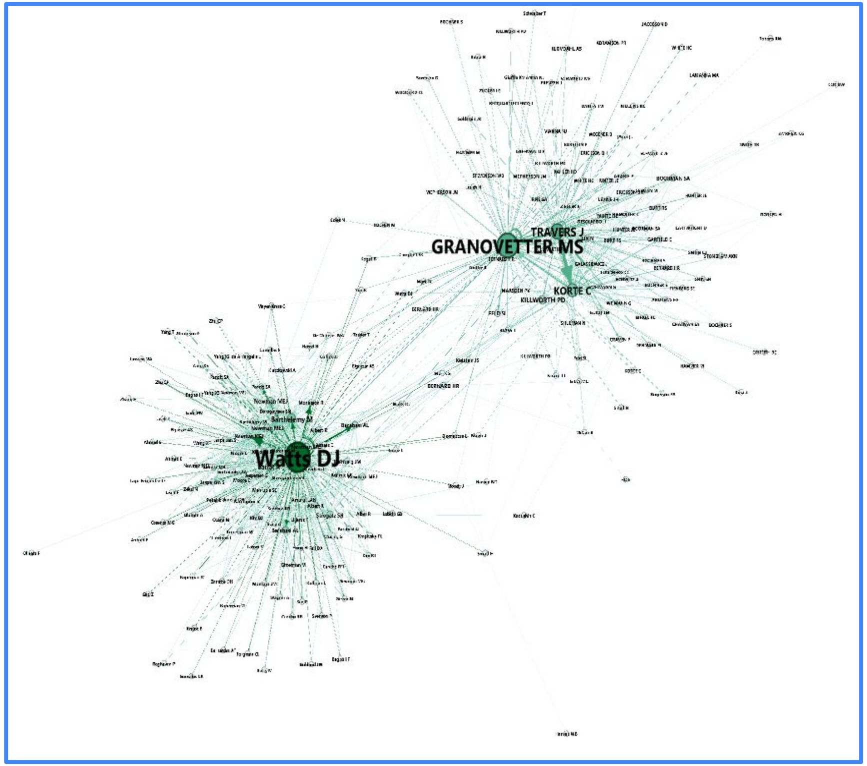


The two separate clusters can now be clearly seen, with a central node for each cluster - **Watts DJ** & **Granovetter MS**

Inference - I

From the two clusters seen in the graph, by hovering over each node within a cluster, we can see that the nodes have more connections towards their own cluster than towards the other cluster.

This simply means that people within one cluster mostly cite people who are also present in the same cluster.

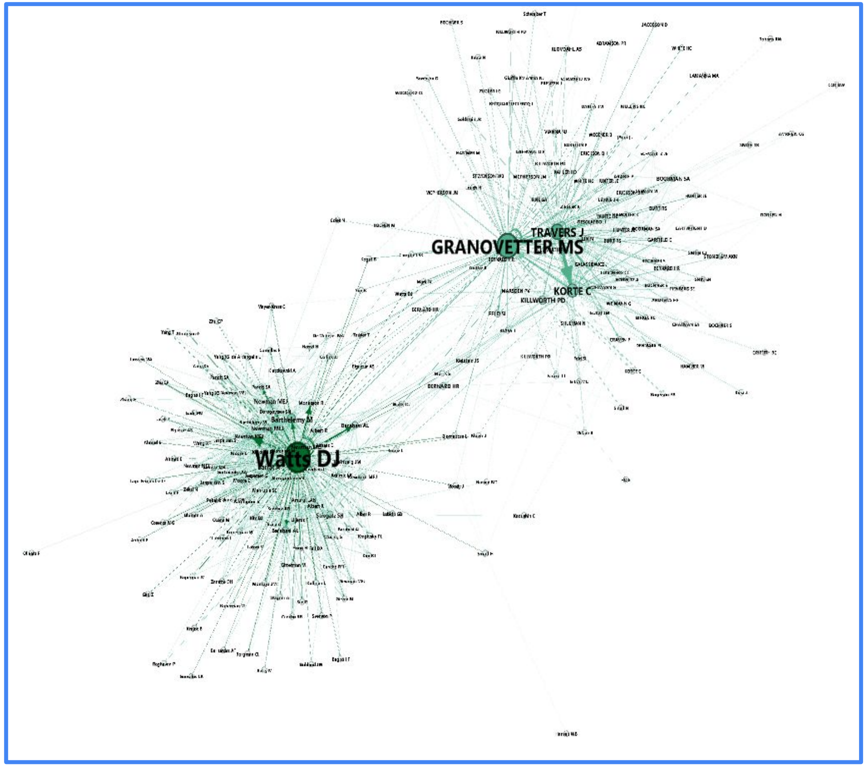


Inference - II

From further domain knowledge & research, we can infer that the two separate clusters are communities of theoretical sociologists and mathematical sociologists.

Watts DJ is a famously known sociologist who worked on the mathematics of the **small-world model** - he is present in the mathematical circle.

Granovetter MS is also a famous sociologist who worked on the theory behind the **strength of weak ties** - he is present in the theoretical circle.

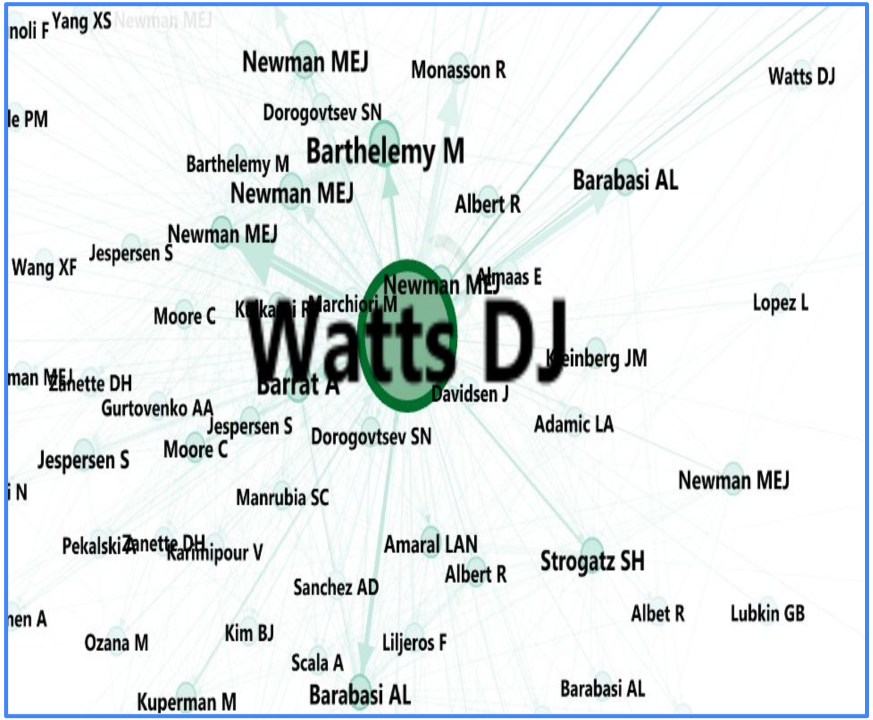


Inference - III

Watts DJ is surrounded by people like **Strogatz SH** and **Newman MEJ**, who are also researchers famously involved with the study of SNA.

Strogatz SH is also the doctoral advisor of Watts DJ and hence the relationship is reflected in this graph.

There are also some **mistakes** with this dataset, as we can see some names are duplicated - Newman MEJ appears multiple times. These entries need to be merged.



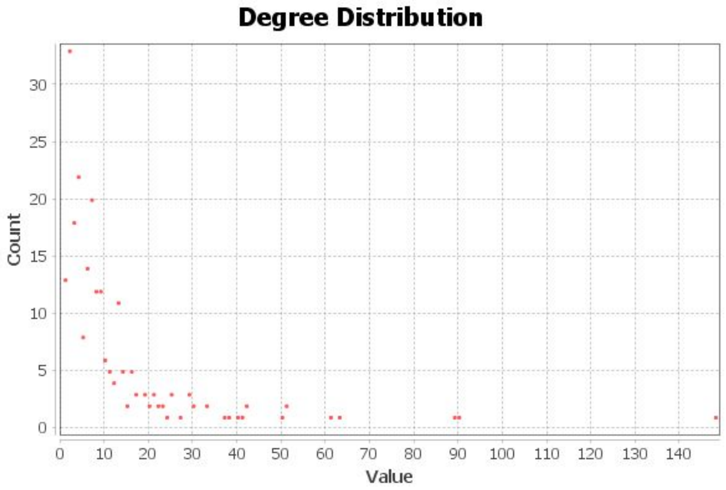
Average Degree

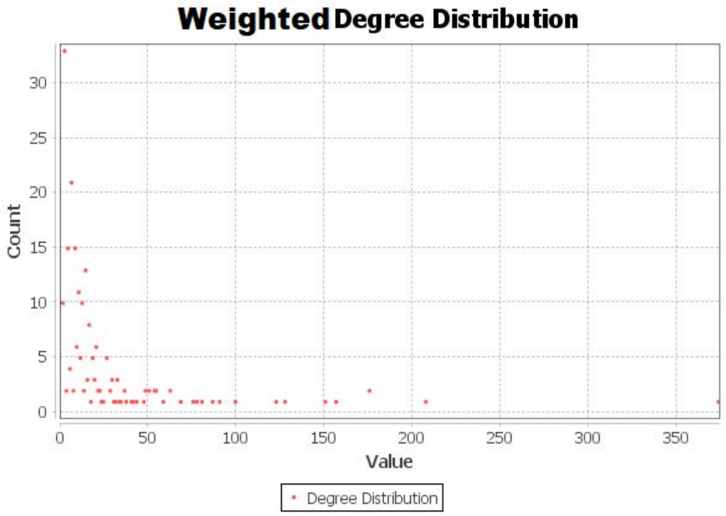
The average degree of the graph after removing Milgram was computed to be **5.753.**

The weighted average was found to be **10.797.**

The **in-degrees** were mostly concentrated at the **lower values** while the **out-degrees** were more **spread out**, indicating that **most people cited a few prominent researchers a lot.**

The network density was found to be **0.025.**





Modularity Analysis

Changes to Dataset and Algorithm used

From Inference - III we could see that there are duplicates in the dataset. So the first thing we did was, we removed the duplicates and merged them. This was done using the inbuilt option provided by Gephi tool - Detect and Merge node duplicates.

Algorithm used to find modularity is:

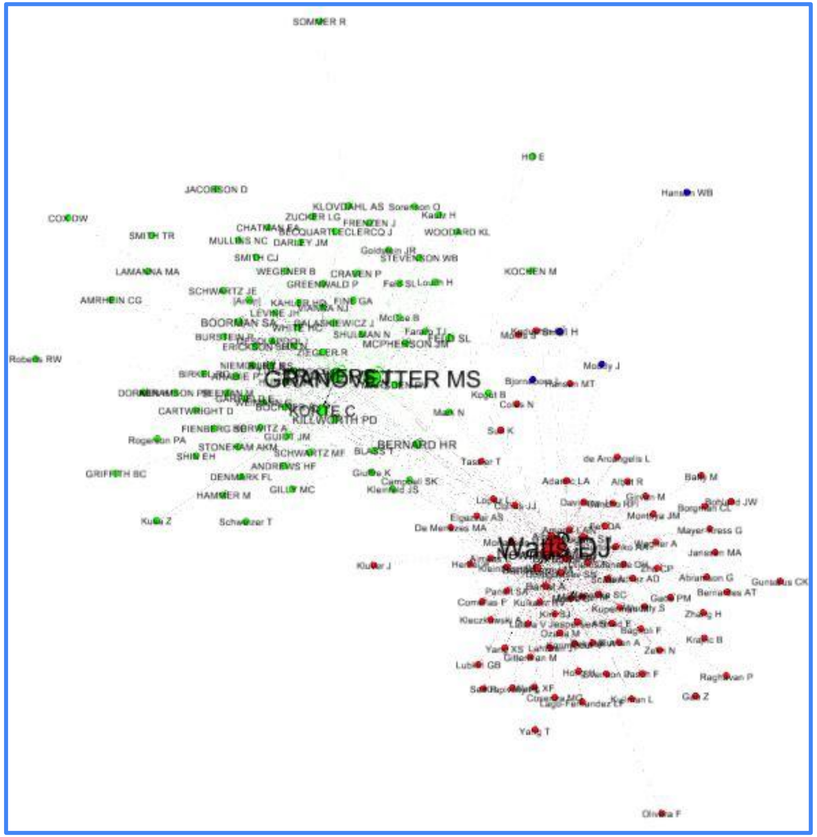
Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000

Resolution:

R. Lambiotte, J.-C. Delvenne, M. Barahona Laplacian Dynamics and Multiscale Modular Structure in Networks 2009

Modularity for Inference = 1

- By default, Gephi tool sets the inference to 1
- On running the Algorithm mentioned earlier, we ended up with a **modularity value of 0.370**
- Also we had just 3 modularity classes. This makes complete sense according to Inference - II, i.e. one class is the **people in the mathematical circle(red)**, the other class is the **people in the theoretical circle(green)** and the other class is the **people who are in-between, who cite papers from both classes(blue)**.



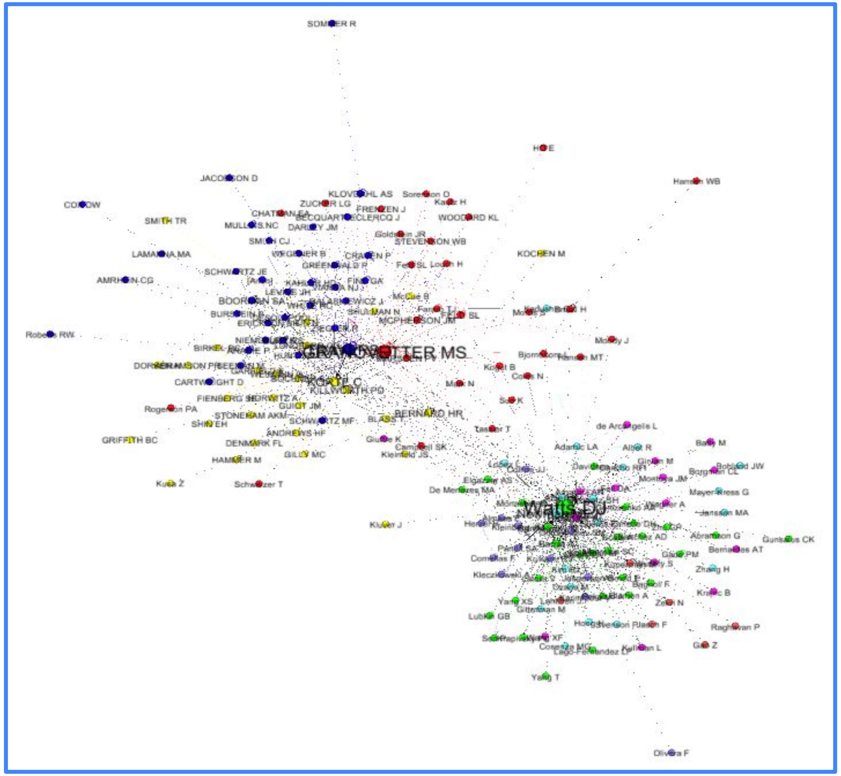
Modularity for Inference = 0.5

- Curious to know more, we decided to reduce the resolution to get more modularity classes. This led to us reducing the inference to **0.5**
- Now we ended up with modularity values as mentioned below:

Modularity: 0.276

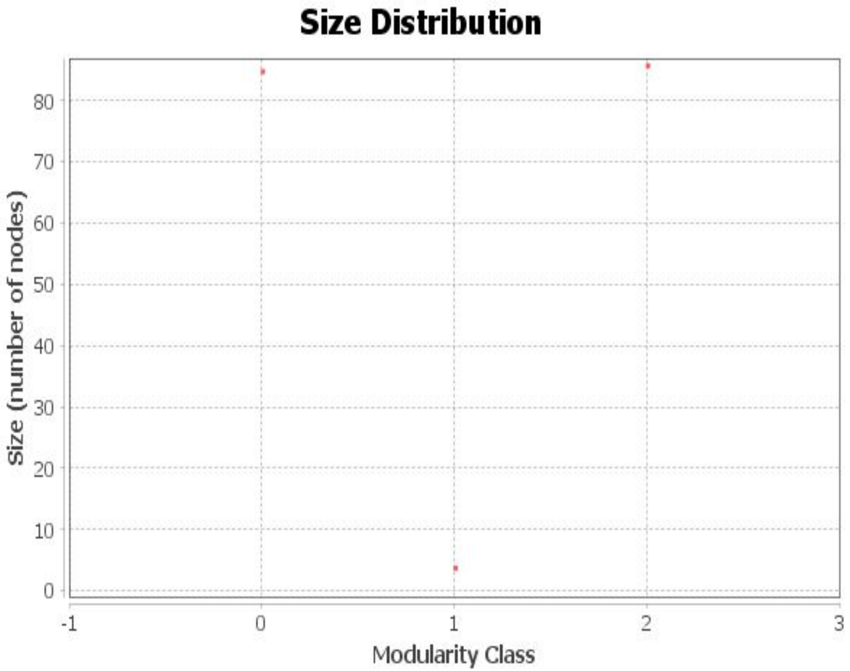
Modularity with resolution: 0.065

- Also, we obtained **8 modularity classes** this time as seen in the diagram.

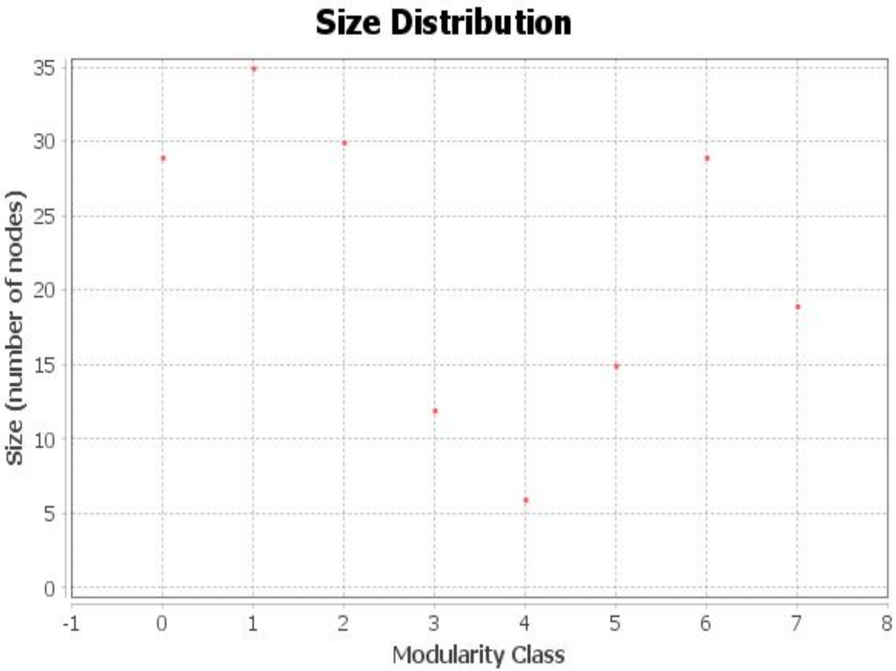


Size Distribution Comparison

Inference = 1, Classes = 3

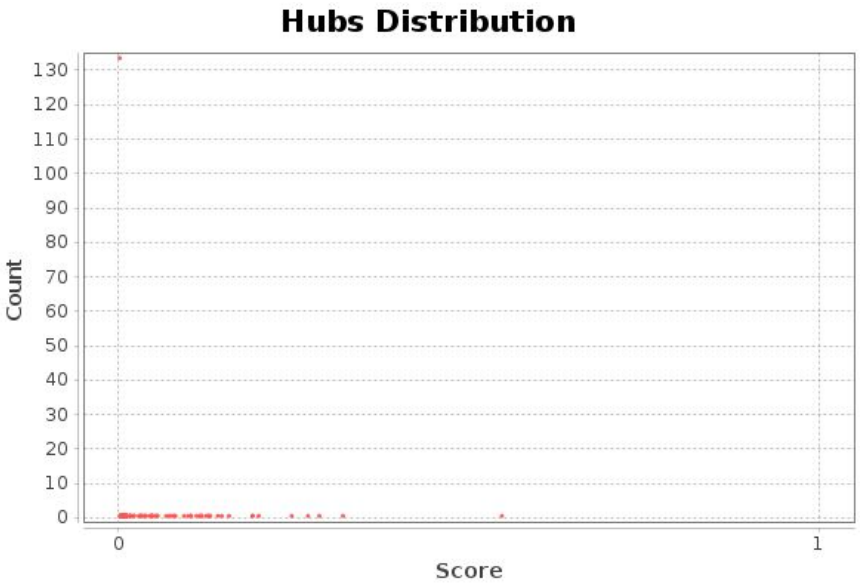


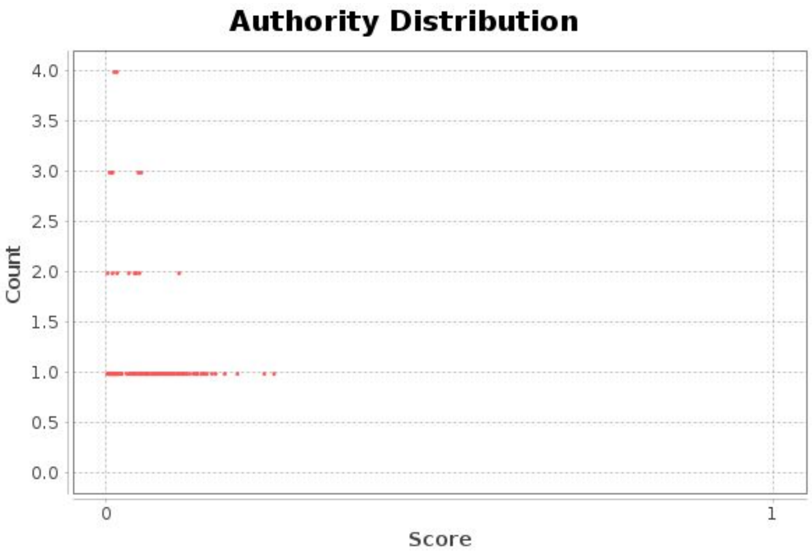
Inference = 0.5, Classes = 8



The two graphs clearly shows us the size of each class. On the left we have a graph with 3 classes (Inference = 1.0) and on the right we have a graph with 8 classes (Inference = 0.5) which is plotted on x-axis and size on y-axis.

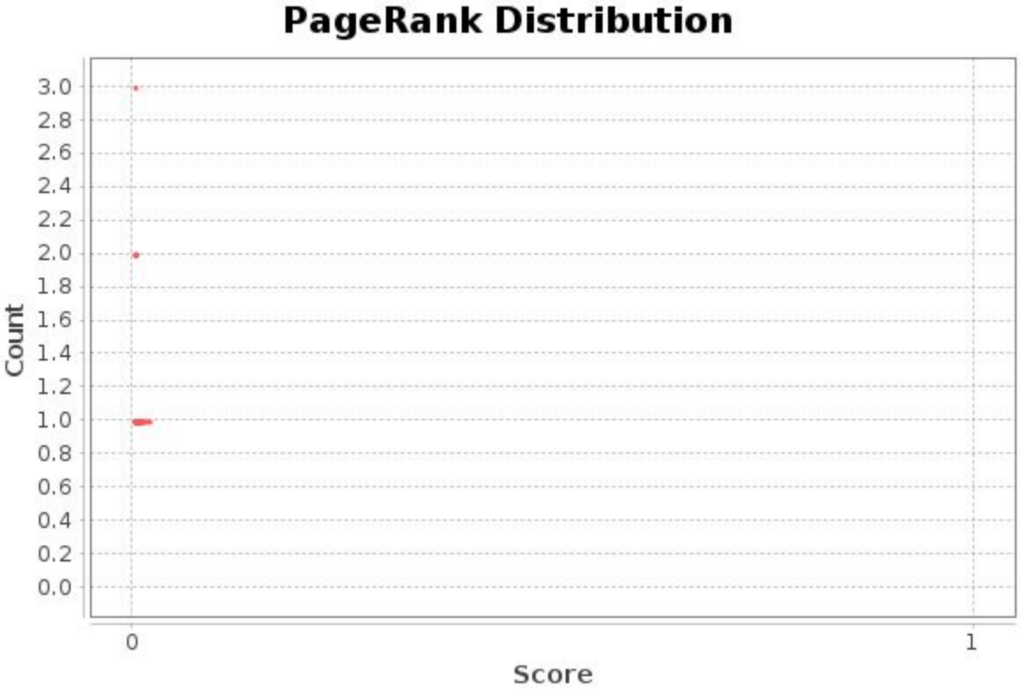
HITS Analysis





An authority value is **computed as the sum of the scaled hub values that point to that page**. A hub value is the **sum of the scaled authority values** of the pages it points to.

PageRank



PageRank unlike HITS does not yield any meaningful results.

Appendix

Definitions

- Diameter - A graph's diameter is the largest number of vertices that must be traversed in order to travel from one vertex to another.
- Closeness Centrality - A metric that describes how close a node is to other nodes in the network.
- Betweenness Centrality - A metric that describes how much a node acts as a bridge between two other nodes. It can identify boundary spanners & bottlenecks.
- Modularity - It is a measure of the structure of networks or graphs which measures the strength of division of a network into modules.