CSE 601 Data Mining and Bioinformatics

Homework 3

Clustering Analysis for Complex Networks (Markov Clustering Algorithm and Pajek Visualization)

Submitted By
Mayur Tale
UB IT Name: mayurvin
Person # 50169256

About Markov Clustering Algorithm

MCL Algorithm is an unsupervised algorithm for *hard clustering* datasets that can be typically represented in graph format. It is considered to be a form of **hard clustering** because each data point can be categorized in only one of the clusters.

MCL is best suited for clustering of datasets where the data can be represented in the form of a network or graph with edges between set of vertices because it is based on probabilistic reachability of nodes from a particular node in the unidirectional-edged network. The edges may be unidirectional or bidirectional (unidirectional). The edges may further be weighted or un-weighted. The algorithm feeds on the dissimilarity of clusters and amplifies the difference to a significance value.

Advantages:

I. We do not have to know how many clusters to categorize the data into.
II. MCL scales well for large datasets – in linear proportion with number of vertices in graph
III. Not sensitive to outliers

Disadvantages:

IV. Hard to find optimal parameters for clustering – changing the power and inflation parameters yields different-sized clusters but no way to assure quality of clustering unless we ***visualize*** results, which needs tools like Pajek

MCL Algorithm Implementation details

- Psudeo code and working of algorithm
  - Load data files and store unique number of vertices from file in an ArrayList<Integer>.
  - Create an adjacency matrix initialized with all 0 values and insert value 1 in cells which has an edge between them (edges are read from file).
  - Set all diagonal cells as 1 for adding the Identity matrix to the adjacency matrix.
  - Loop while new Adjacency matrix and old are not equal
    a. New Adjacency matrix equal to old matrix
    b. Normalize the new matrix
    c. Expand the new Matrix with power coefficient (2)
    d. Inflate the new matrix with inflation coefficient
    e. Prune the matrix by changing values greater than 0.99 to 1 and less than 0.01 to 0
  - Generating ArrayList of clusters:
    a. For each row create an ArrayList of columns with values greater than 0
    b. Check if this ArrayList is already in the cluster list and add if not
  - For each list in lists of clusters write the cluster number on each new line of the .clu file
- To run the program, user has to do is run the program for each file separately (Power coefficient and Inflation coefficient can be changed in the program directly).
- Output will be, Number of clusters generated for the dataset and .clu file generated message
- I have extensively used ArrayList data structure throughout the implementation.

Visualization Results using Pajek:

Files: **1) attweb_net.txt**

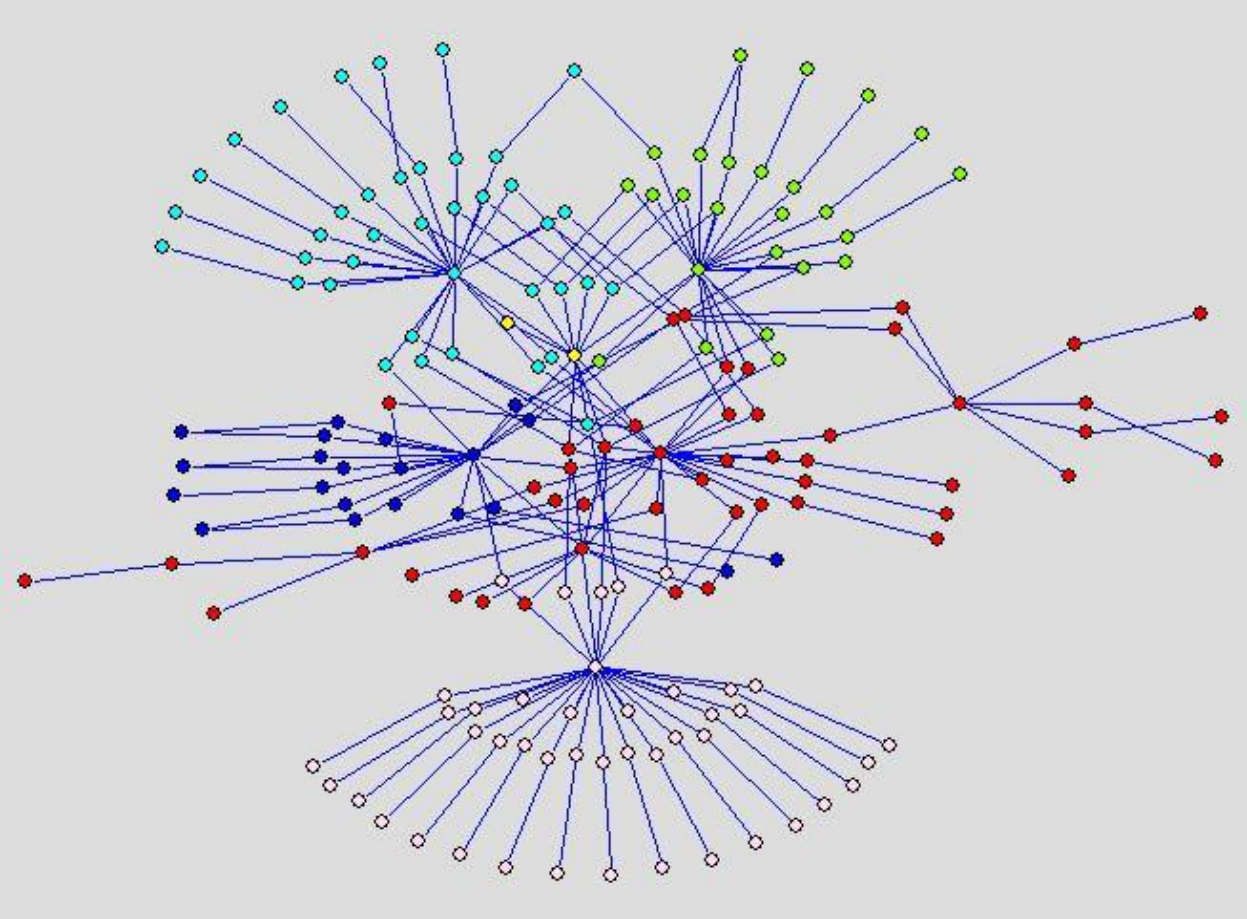
2) attweb_net.net

3) PajekFile_attweb_NEW.clu

Parameters:

Power Coefficient: 2

Inflation Coefficient: 1.32



Files: **1) physics_collaboration_net.txt**

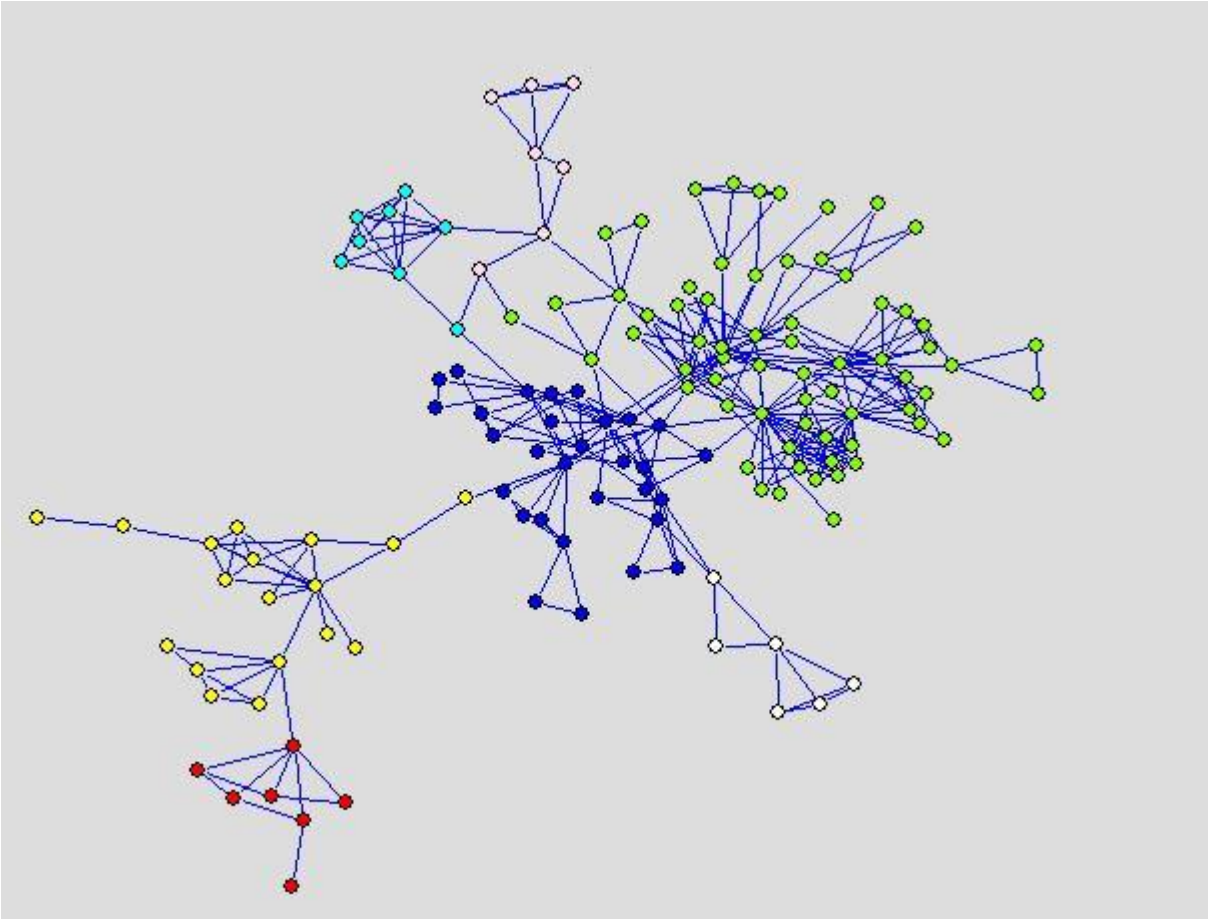
2) physics_collaboration_net.net

3) PajekFile_physics_collab_NEW.clu

Parameters:

Power Coefficient: 2

Inflation Coefficient: 1.20



Files: **1) yeast_undirected_metabolic.txt**

2) yeast_undirected_metabolic.net

3) PajekFile_yeast_metabolic_NEW.clu

Parameters:

Power Coefficient: 2

Inflation Coefficient: 1.22