DATA MINING

ASSIGNMENT-3: CLUSTERING ANALYSIS FOR COMPLEX NETWORKS

TEAM MEMBERS:

Srikanth Madduri Venkata	Arjun Sundaresh	Sridhar Vadlamani
Ssmaddur	arjunsun	sridharv
50134851	50169917	50168092

Markov Clustering Algorithm

The Markov Cluster Algorithm is a fast and scalable unsupervised cluster algorithm for graphs or networks. The MCL algorithm simulates flow within a graph and promotes flow in a highly connected region and demotes otherwise, thus revealing natural groups within the graph.

In some cases, when the hub has a low curvature, many of the vertices adjacent to the hub are removed from the cluster that it represents. However, in order to make up for that, the hub cluster will absorb many other vertices—some of which are not directly connected to the hub itself—to form a large-sized core cluster. The same can be observed in ATT network as well.

Markov Clustering algorithm takes O(n3) time while finding the power or while expanding it. And O(n2) time while inflating. This is a lot of time and it can be reduced by pruning appropriately.

Implementation

Creating the adjacency matrix:

1. Initially we have given an index for each node between 0 and n-1 where n is the number of distinct nodes in the network.
2. As we parse through the dataset the interaction between two nodes has been marked as 1 with the help of their indices.
   Eg: if nodes “a” and “b” are interacting with each other, then if the indices of the nodes are 2 and 6 respectively, then the cells [2][6] and [6][2] of the adjacency matrix has been marked as 1.

Adding Self loops:

3. We have also added self-loops for every node in the network in the adjacency matrix.

Normalizing the adjacency matrix:

4. Once the self-loops are added, we have normalized the matrix. By taking the sum of the data in every column and then dividing every element in it with the sum.

Expand the Matrix:

5. After normalizing the matrix, we have expanded the matrix by calculating it’s power of ‘n’. In our datasets, the ‘n’ value ranged from 30 to 150.

Inflating the Matrix:

6. Once the matrix is expanded, it is inflated by calculating the power of ‘x’ for each and every element in the matrix.

Normalizing the Matrix:

7. Again, after the inflation of the matrix, the matrix is normalized as described above.

Iterating till Convergence

8. The steps 5 to 7 are repeated till the matrix is converged.

Injection of results into pajek

9. Once the matrix is converged, the clusters are formed from the resultant matrix.
10. A cluster is actually a row in the matrix whose members are non-zero elements in the matrix.

Observations:

From the visualization it can be observed that as “expand” and “inflate” values are increased the number of clusters are decreased and vice-versa.

Visualization Results for Datasets:

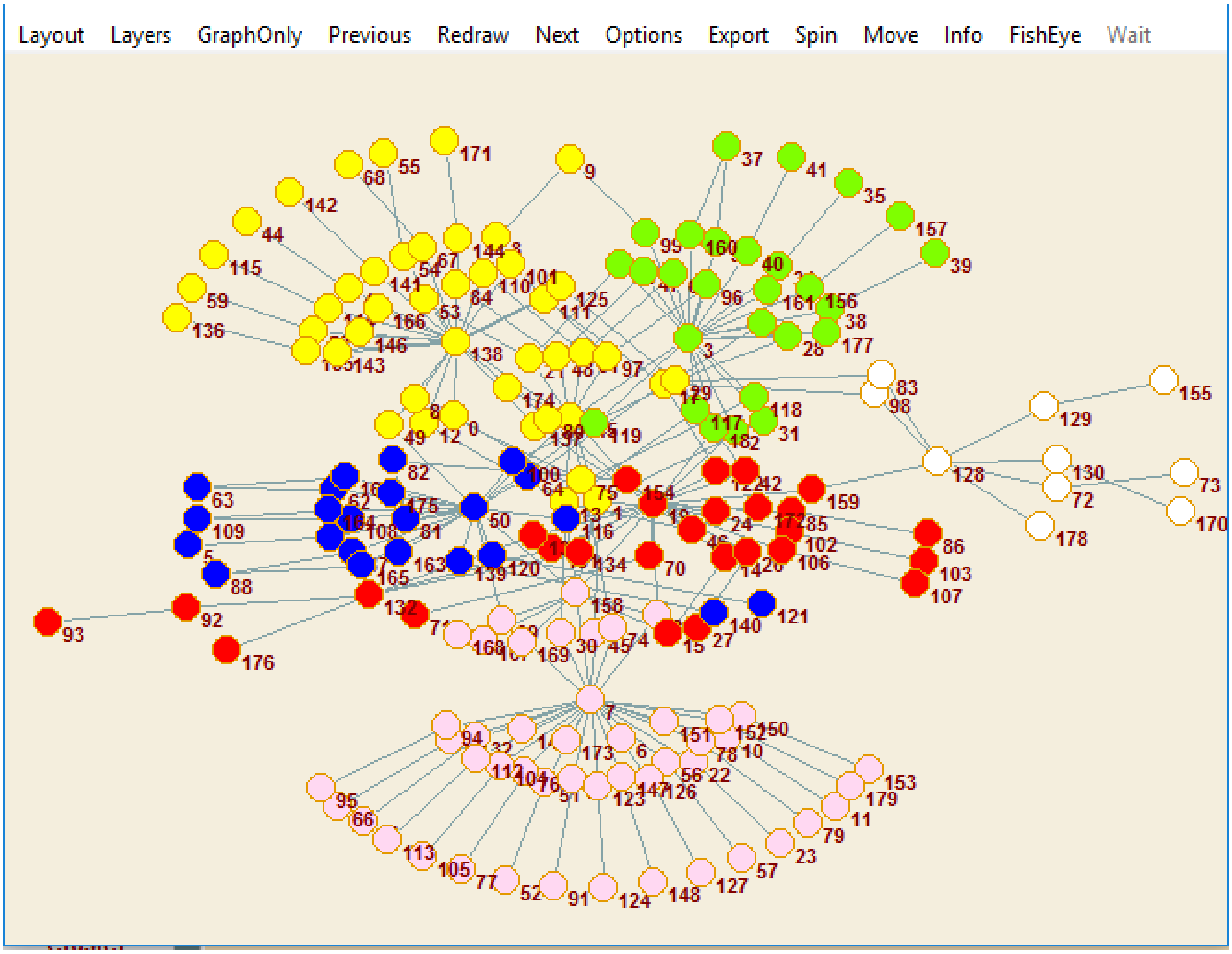
1. AT&T Web Network

Total nodes – 180.

Total edges – 228.

Total clusters - 6

Power – 30 Inflation - 30



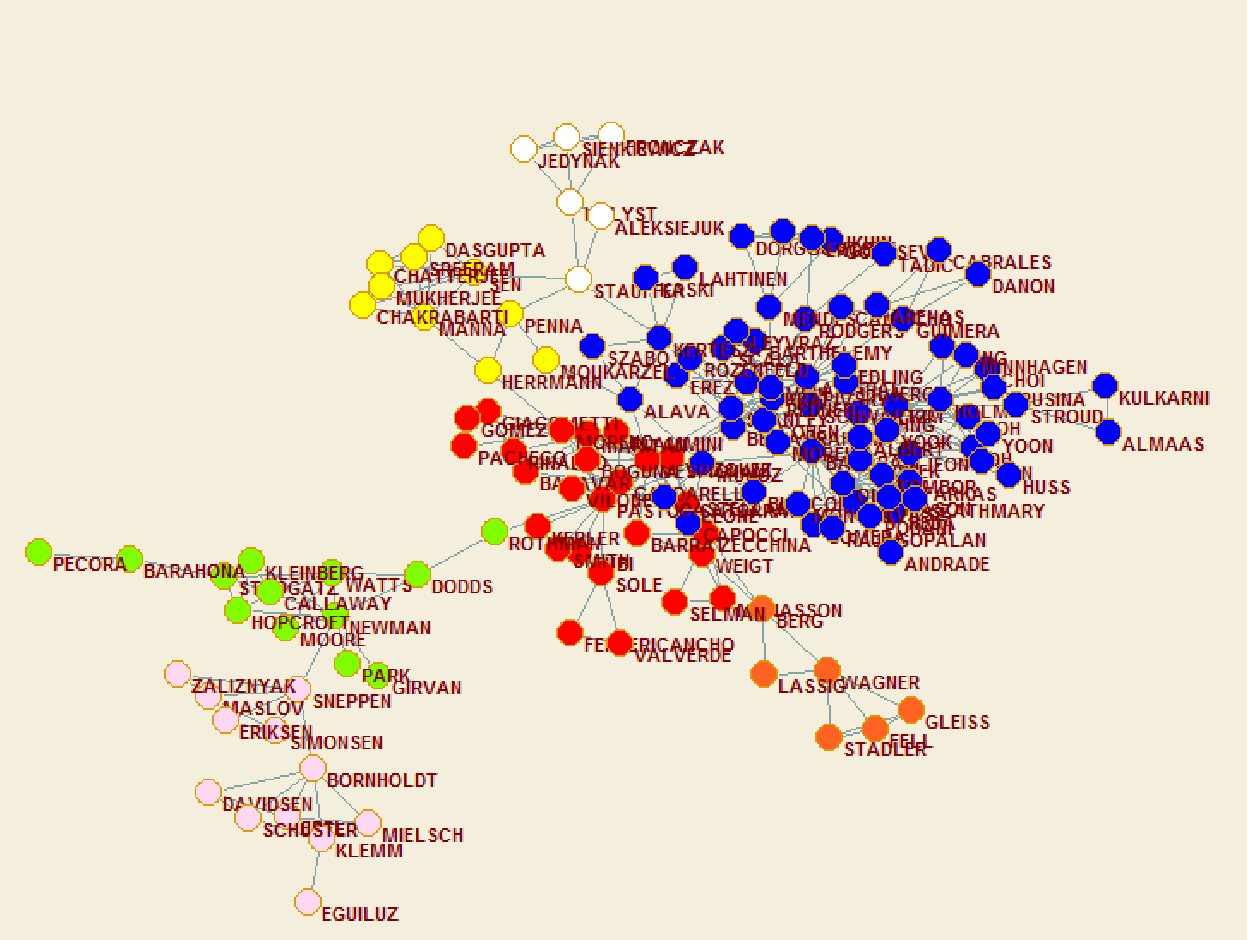
2. Physics Collaboration Network

Total number of nodes – 142

Total Edges – 340

Total Clusters – 7.

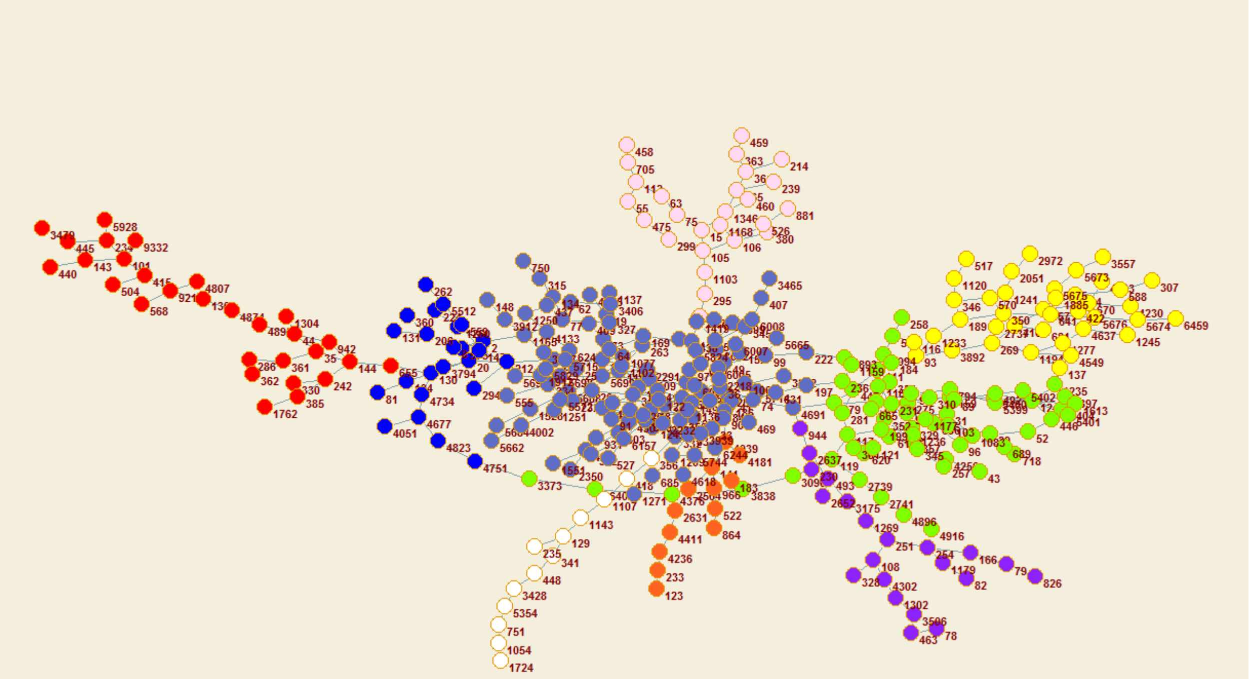
Power: 30 Inflation: 50



3. Yeast Undirected Metabolic

Total number of nodes – 359
Total number of edges – 435
Total number of cluster – 9

Power - 140   Inflation - 140



459
458
705
363
214
36
239
63
55
460
881
475
75
1346
1168
526
380
299
15
106
105
1103
295
5928
3470
445
9332
143
440
101
504
415
4807
568
921
4874
1304
489
942
286
361
35
242
144
362
330
385
1762
750
315
262
5512
360
131
148
3912
1250
4137
3406
3465
407
6008
5665
222
517
2972
3557
1120
2051
5673
307
346
570
1241
1885
588
1230
6459
189
350
5676
5674
4637
1245
1233
3892
269
277
4549
137
258
116
93
184
130
3794
20
81
4734
4677
4051
4823
4751
555
4002
5662
3373
1551
2350
527
540
1107
1143
129
235
341
448
3428
5354
751
1054
1724
418
685
1271
4376
2631
522
864
4411
4236
233
123
4618
5744
4181
3838
469
197
4691
944
2637
119
493
3098
2739
2741
3175
1269
4896
4916
251
254
166
79
826
1179
82
108
328
4302
1302
3506
78
463
5402
5399
446
5401
397
1083
689
718
52
96
257
43
281
620
352