CS-3404: Minor Project

A comparative study of clustering algorithms for Blog tag networks

Done by,
Abhijith V Mohan
10400EN001
CSE IDD-Part 3

Under the guidance of Prof. Bhaskar Biswas

Objective

- To implement the following 2 clustering algorithms:
  - Markov Clustering(MCL)
  - K Clique Maximal percolation
- To cluster the given data of blogs based upon their tags using both the algorithms and compare the results in terms of performance, and the quality of clusters generated

Input Dataset

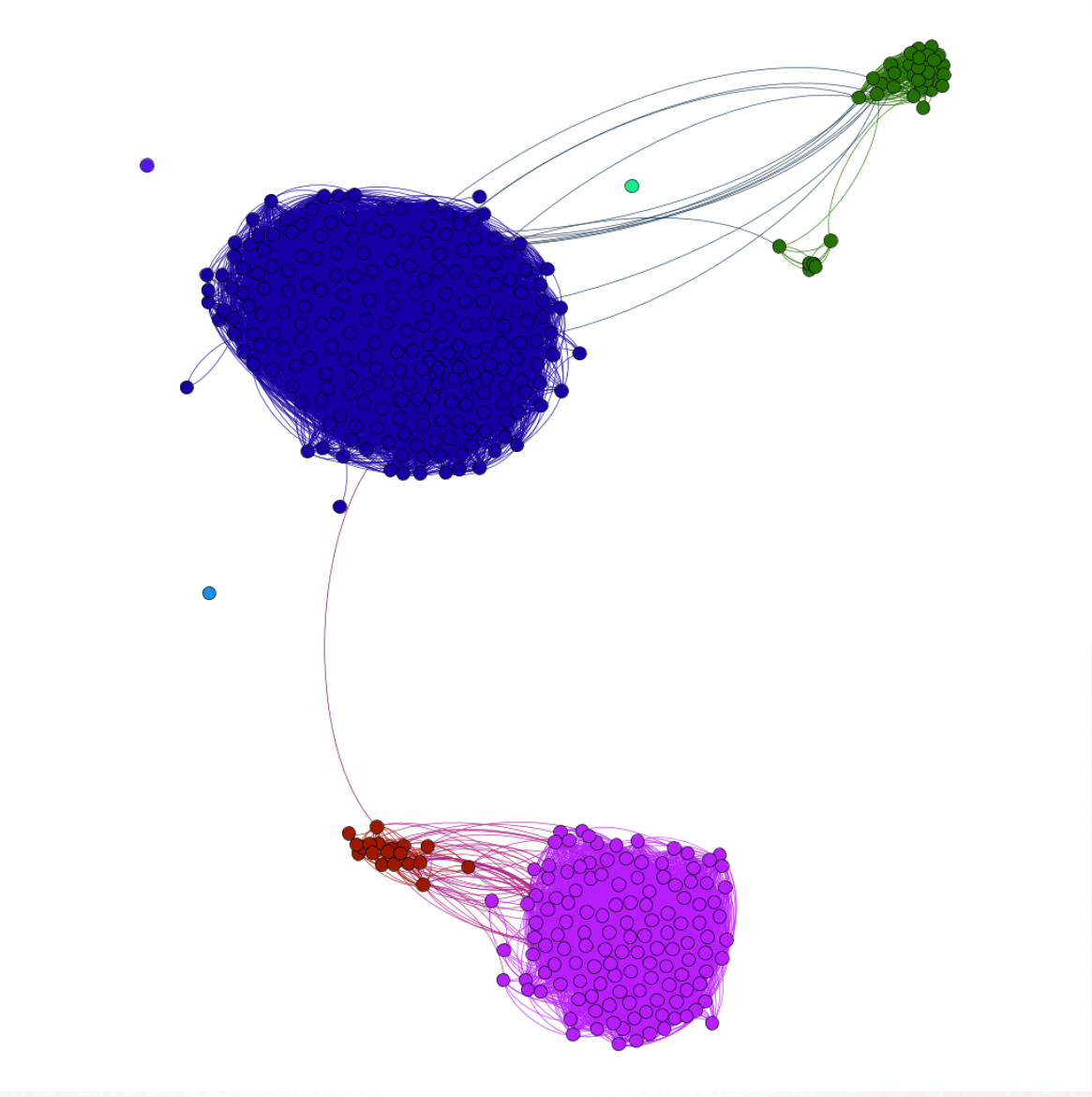
- The input dataset was a list of blogs with their corresponding tags from [source]
- A tag graph was constructed from whose nodes where the blogs and there where edges between two nodes who shared any tag
- If the tag set of the blog corresponding to node i is denotes as T_i, then the weight of edge connecting nodes i & j expressed as percentage is given by the formula:

$$W_{ij} = \frac{|T_i \cap T_j|}{|T_i \cup T_j|} \times 100$$

Clustering

- A cluster/community is a subgraph of a graph which has an intra cluster link density much higher and inter cluster link density much lower than the average link density of the graph.
- Intuitively, a node in a cluster has more edges to nodes within the clusters than to other nodes.
- Vertices with a central position in the clusters may have an important function of control and stability within the graph. For example, a central node in a social network cluster might be instrumental in the opinion formation and might exert a lot of influence on the cluster.

Clustering applied to an FB network



Why clustering?

- Recognising patterns and relations within the data
- Clustering Web clients who have similar interests and are geographically near to each other
- Identifying customer clusters with similar interests allows online retailers (such as flipkart.com) to set up efficient recommendation systems
- Ad-hoc wireless networks usually have no centrally maintained routing tables. Clustering the nodes enables one to route through efficient paths and at the same time have compact routing tables

Markov Clustering

Algorithm

- Column normalize the matrix
- While the steady state has not been reached:
  - Expand(W,e)
  - Inflate(W,i)
- Interpret the resulting matrix as the adjacency matrix of a graph whose connected components are the clusters in the original graph

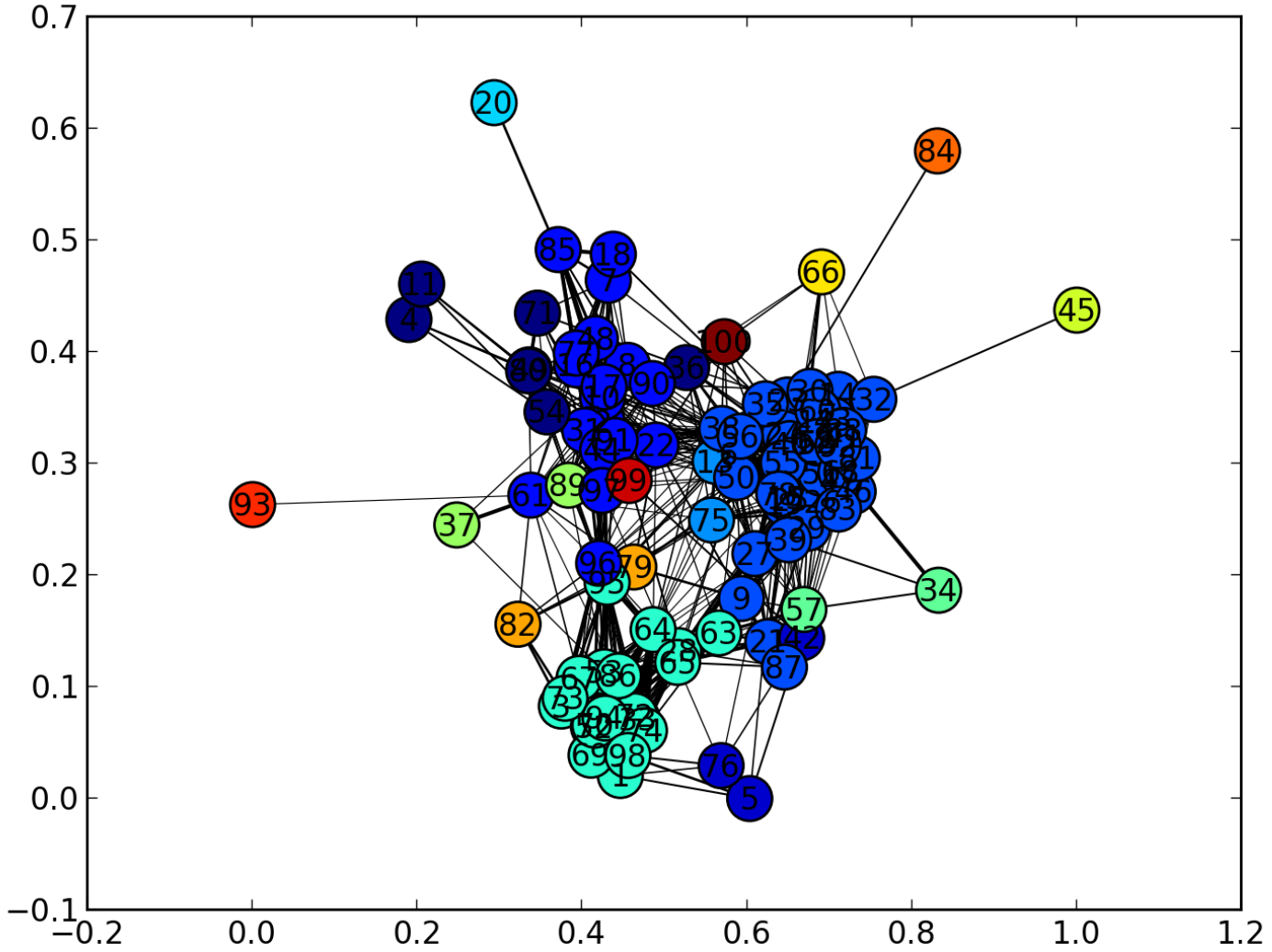
Expansion parameter

- The expansion parameter is the length of each random walk before the inflation process is applied to the matrix to deteriorate the weak probabilities further
- It does not have a very strong effect on the result compared to the inflation parameter

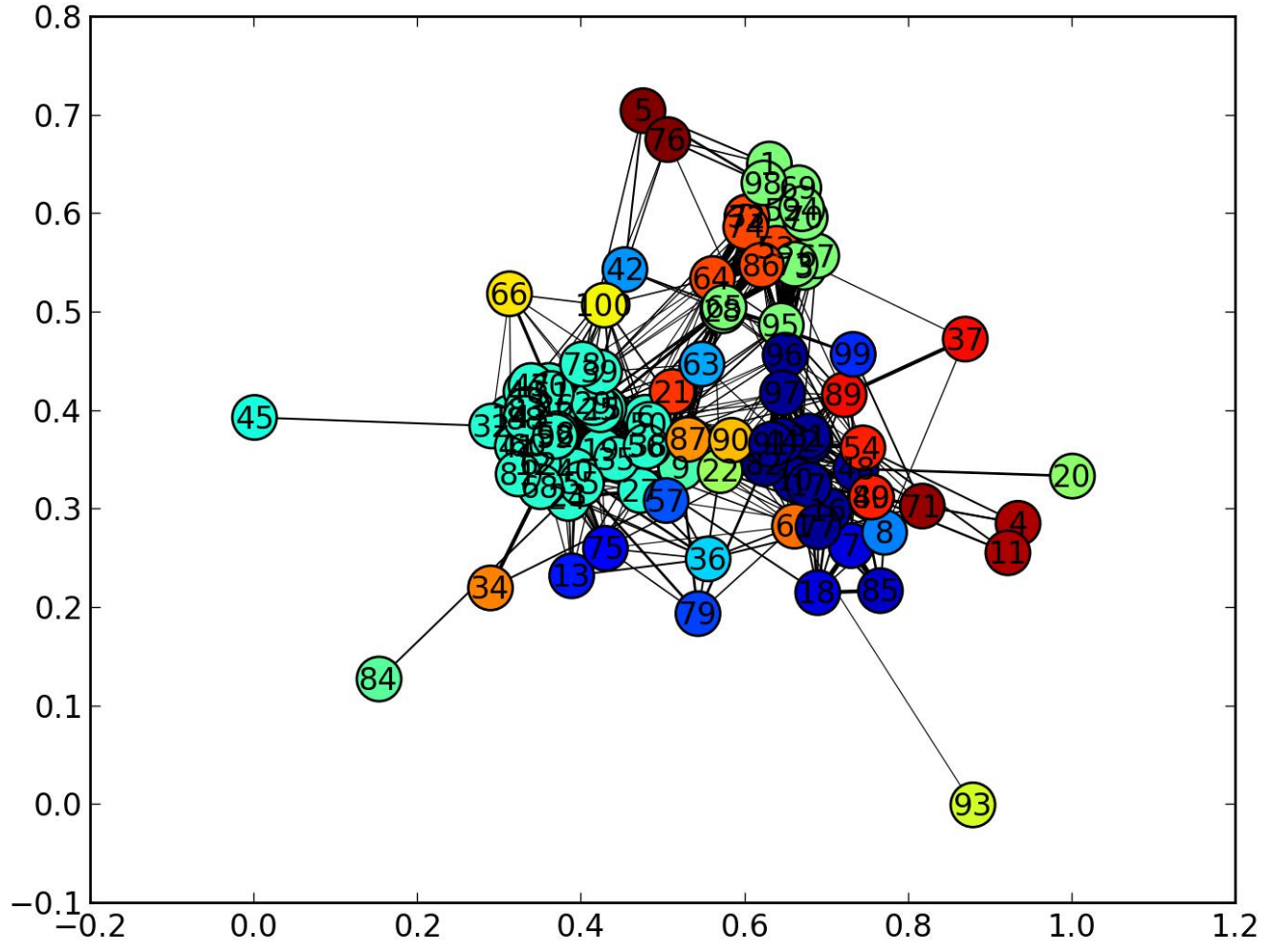
Inflation parameter

- The inflation parameter, i has a strong effect on the granularity of the clusters
- The weaker transition probabilites drop to zero with higher value of inflation parameter
- With higher i value, the members of any cluster are more strongly related.
- It has no physical meaning unlike the expansion parameter

n=100, e=2, i=2
16 clusters, 4 with min 5 nodes



n=100, e=2, i=5
34 clusters, 4 with min 5 nodes



Features of MCL

- $O(N^3)$ time complexity (expansion is $O(N^3)$ and inflation is $O(N^2)$)
- It has a natural parameter to influence cluster granularity
- Easily scalable and parallelizable

K Clique Clustering

Algorithm

- Given K and the graph G.
- Find C = the set of all K-cliques in G.
- Find the set of all maximal unions of adjacent K-cliques. 2 K-cliques are adjacent if they have K-1 nodes in common
- Each such union corresponds to a cluster in the graph.

Some considerations

- Some of the nodes may not be a member of any K-cliques. In that case, that node will not be a part of any cluster
- Some nodes may be part of more than one maximal union of adjacent K-cliques. In that case, the node will be a part of more than one cluster
- So the K-Clique clustering generates a cover rather than a partitioning.

N=100, K=10

- {2 6 10 77 16 17 22 89 90 91 80 31 96 97 38 44 48 49 50 54 56 61 95}
- {6 12 13 14 15 19 23 24 25 26 27 29 30 32 35 38 39 40 41 43 46 47 50 51 55 56 58 59 60 62 68 75 78 81 83 88 92}
- 2 clusters were formed. It can be seen that some nodes are present in both and some are in neither of them.

The parameter K

- If K=1, then every node on the graph is a part of the same cluster and the whole graph becomes a single graph
- If K=2, then every node other than the single nodes are part of a cluster. Each cluster corresponds to a connected component on the graph
- For large values of K, K-cliques may not be found especially in sparse graphs. Many nodes in such a case may not be a part of either cluster

Conclusion

- The 9614 node blogger dataset has 5 clusters with at least 100 nodes in them
- Of these 5, 2 clusters contain more than 1000 nodes in each (3108 & 2011 nodes respectively)
- So there are 2 major clusters in the dataset.

Implementation details

- Both algorithms were implemented in python with the aid of the following scientific computing libraries
  - Networkx, a graph library
  - Scipy, a scientific computing library
  - Matplotlib, a plotting library

End