Detection of communities in social networks.

Rahul Bajaj
Manu Bansal

Guide :
Prof. Anil Vullikanti
NDSSL , Virginia Tech.

Introduction

- “*community structures*”

- Tendency of vertices to divide into groups with dense connections within groups and only sparser connections between them .

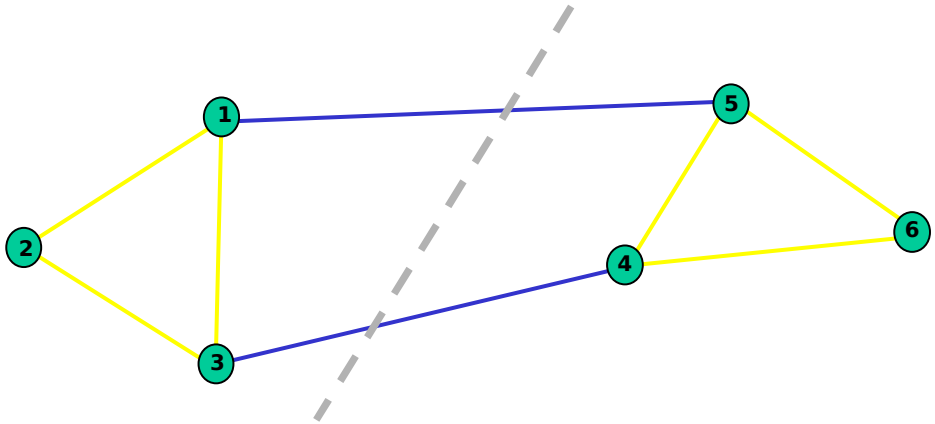
- *Social networks , biochemical networks , information networks* such as the web have all been shown to posses strong community structures

- **Clustering**

Clustering Objectives

- Traditional definition of a “good” clustering:
  1. Points assigned to same cluster should be highly similar.
  2. Points assigned to different clusters should be highly dissimilar.

- Apply these objectives to our graph representation

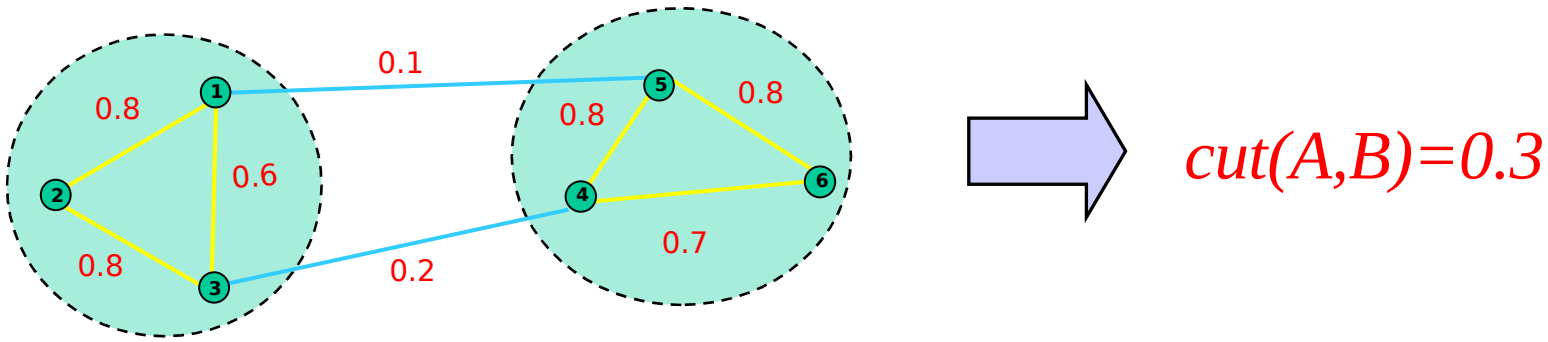


1. Maximize weight of within-group connections
2. Minimize weight of between-group connections

Graph Cuts

- Partitioning objectives as a function of the “edge cut” of the partition.
- *Cut:* Set of edges with only one vertex in a group.

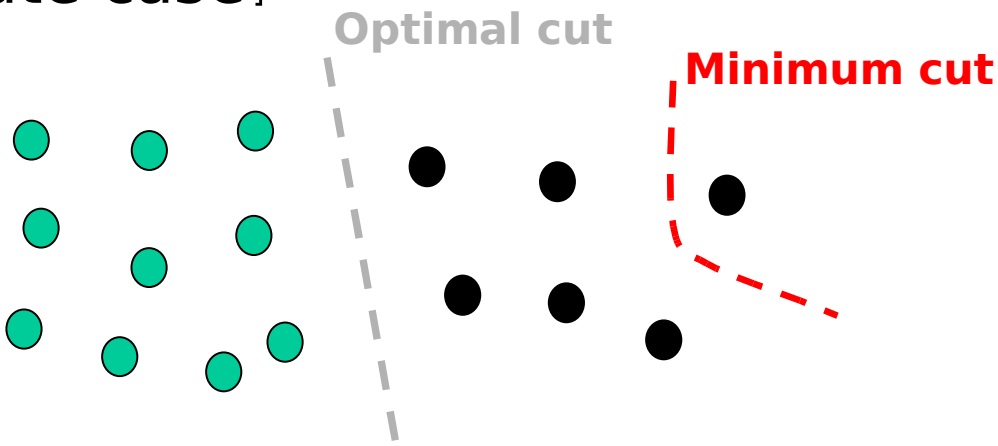
$$cut(A,B) = \sum_{i \in A,\, j \in B} w_{ij}$$



Graph Cut Criteria

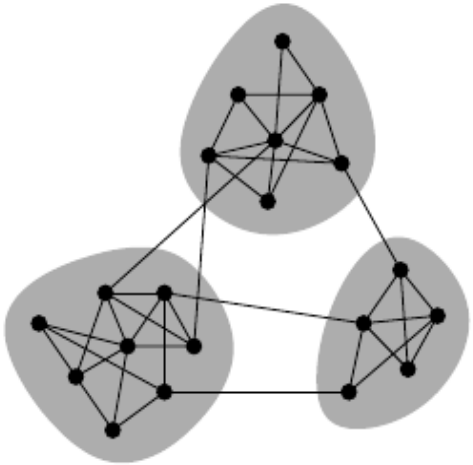
- **Criterion: Minimum-cut**
  - Minimise weight of connections between groups $\min cut(A,B)$

- Degenerate case:



- Problem:
  - Only considers external cluster connections
  - Does not consider internal cluster density

Problem Statement



Courtesy:M.E.J. Newman,2006.

- We had to come up with an algorithm that could divide a social network into clusters.
- The problem was that the algorithm should scale up for graphs containing millions of nodes.
- The quality measure should be such that that it helps analysis of disease spread in the network.

The Quest

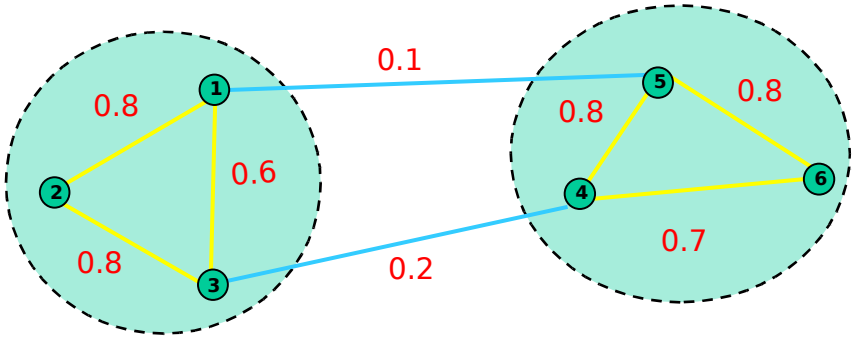
- In our quest for the algorithm we encountered the work of computer scientists , mathematicians ,physicists working on the problem.
- Spread of Epidemics through a network is related to the **spectral radius** of the graph[A. Ganesh]
- The **spectral radius** is related to the **conductance** .
- Kannan, Vetta and Vempala were using **conductance** as their quality measure.
- We also looked into the modularity based algorithms proposed by Mark Newman.

Conductance

The conductance of a cut $(S, \bar{S})$ in G is denoted by:

$$\phi(S) = \frac{\sum_{i \in S, j \notin S} a_{ij}}{\min(a(S), a(\bar{S}))}.$$

Here $a(S) = a(S, V) = \sum_{i \in S} \sum_{j \in V} a_{ij}$.



Numerator =0.3

Denominator=4.7

The Algorithm

- Step 1 :Normalize the Adjacency matrix

  $\tilde{A}=I - L$

  $L=(D - A)/d$
- Step 2 :Find the second Eigen vector of $\tilde{A}$.
- Step 3 :Sort the components of the eigenvector.
- Step 4 :Min-conductance cut based on the sorted eigenvector
- Step 5. If the conductance of the cut is less than the specified measure $\acute{\alpha}$, recurse on the cut.
- Step 6 : local refining on the clusters
- Running time : O(kMlog(n))

Differences between this algorithm and the original one proposed by Kannan ,et al.

- Kannan et al had used their algorithm mainly on document-term matrices , searching web documents ,etc but not in detection of communities in social networks.

  (As far as their paper is concerned)

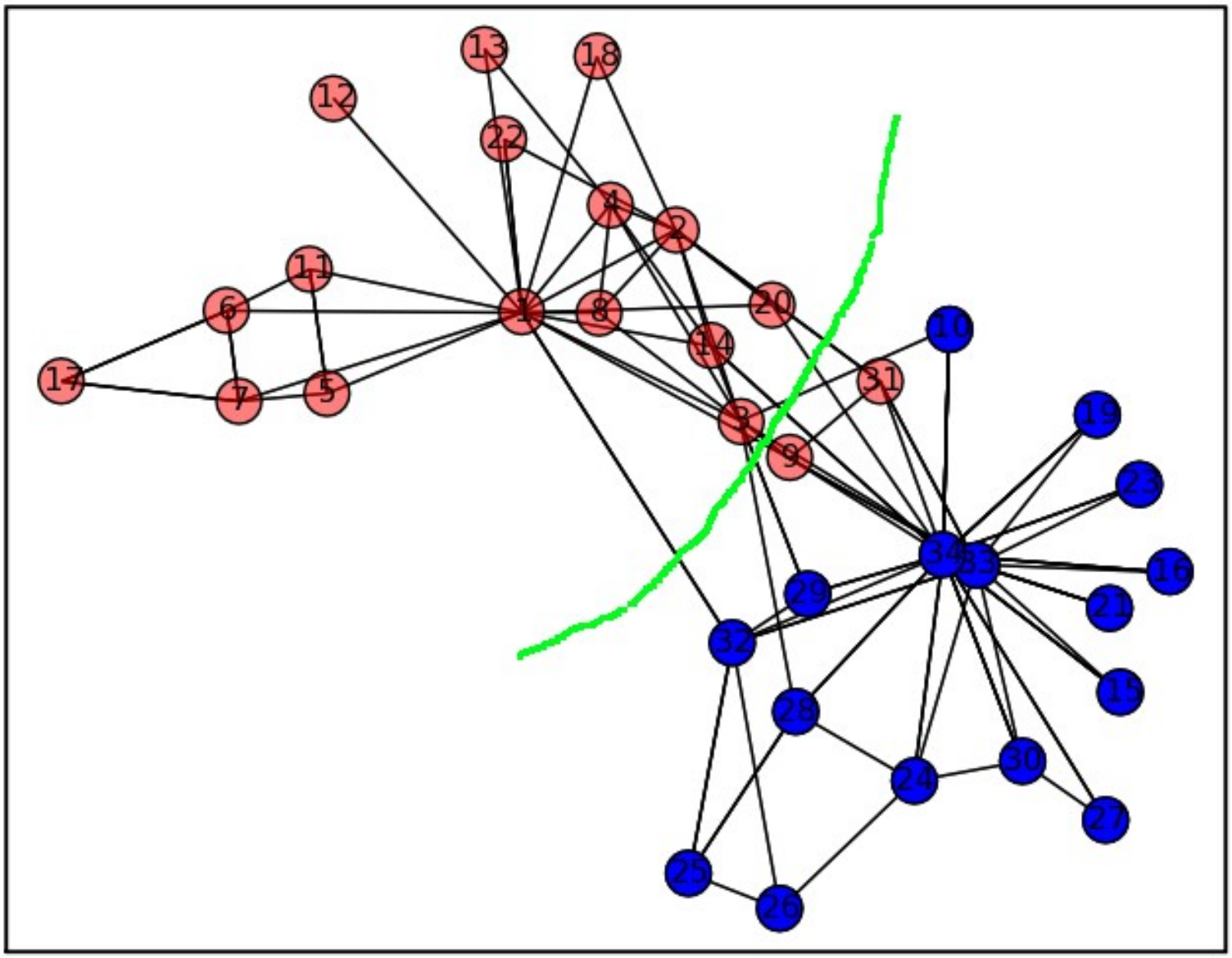
- We are using a different form of normalization . Though its not producing the graph Laplacian the eigenvectors of our normalized matrix is same as that of the graph Laplacian .

Finding the second eigenvector: Power Method

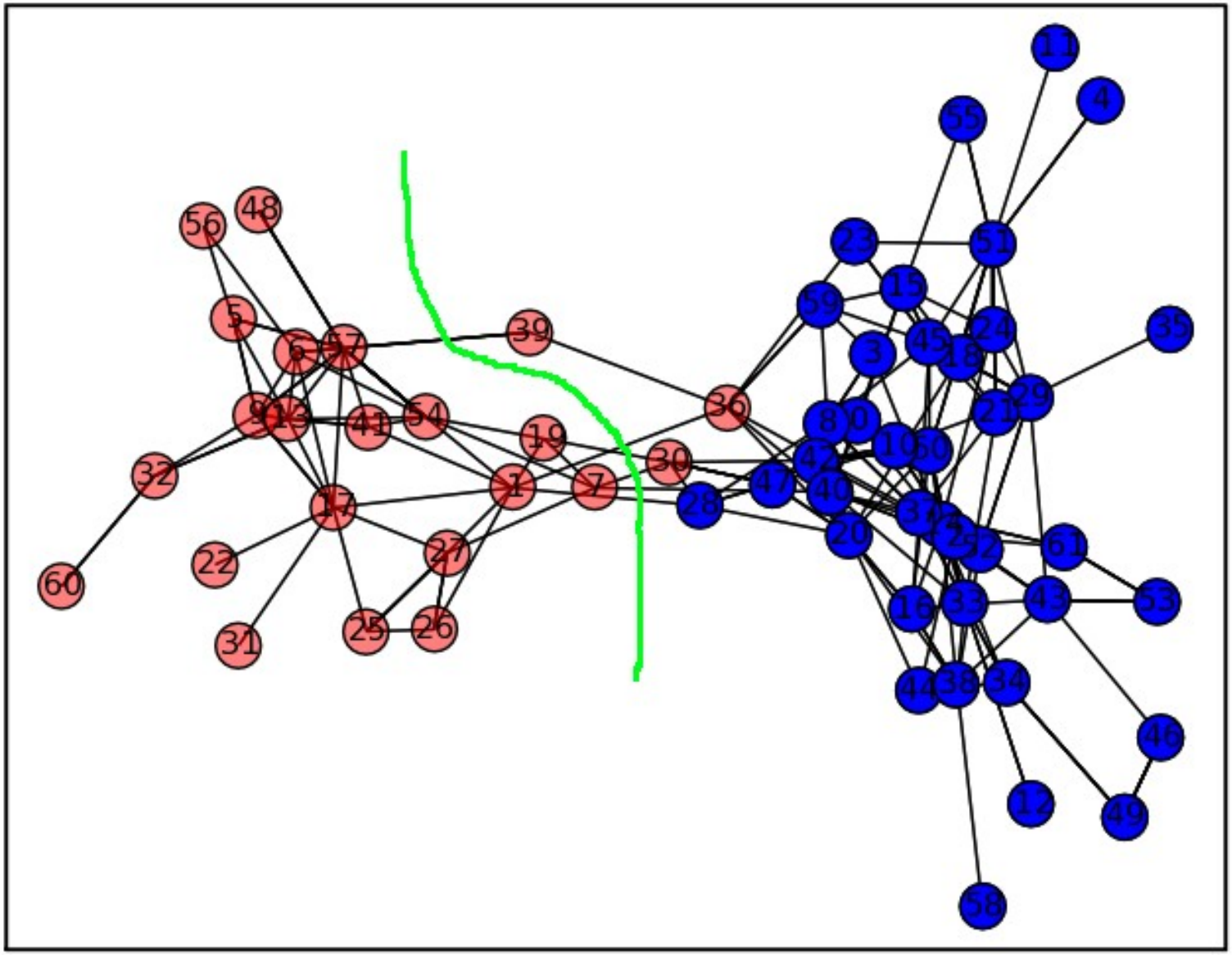
- Used for computing the leading Eigen vector of the matrix.
- One matrix-vector multiplication in each iteration.
- The rate of convergence depends on the ratio between the corresponding eigenvalues.
- Steps:
  - Pick up a random vector **υ** orthogonal to the first Eigen vector.(In our case because of the normalization ,the leading eigenvector is unity.)
  - Normalize **υ**
  - Set **υ= Aυ**
  - Repeat above two steps O(log n) times.

Experiments

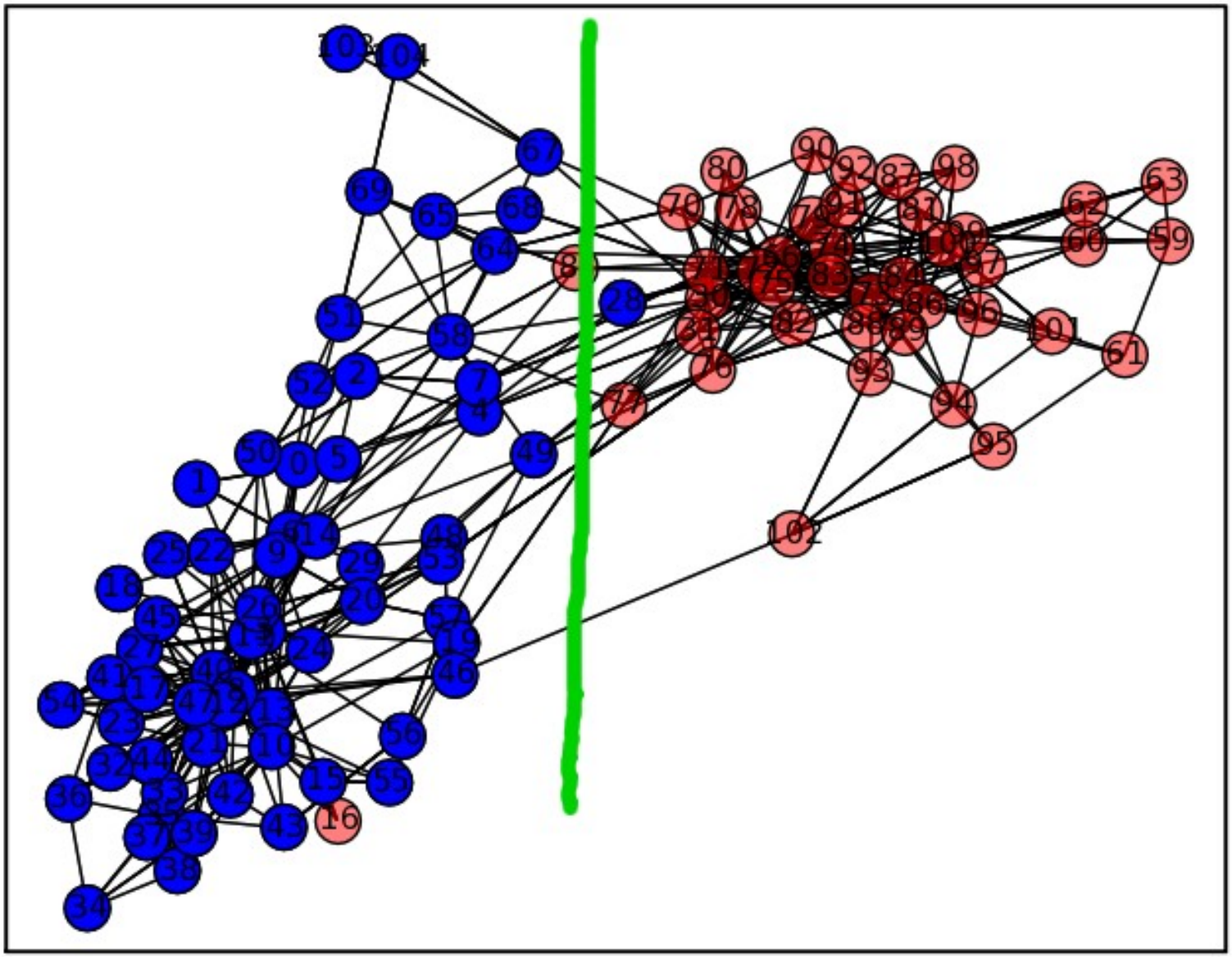
- **Network:** **Nodes** **Time(seconds)**
  - Zachary karate club network 34 0.992
  - Dolphin social network 62 1.532
  - Political books network 104 3.023
  - Political blogs network 1490 61.139
  - Astro collaborator network 16000 ~2 hours



Zachary Karate Club Network.



Dolphin Social Network



Network of political books.

Alternative Approach To Find The Eigen Vector

- Deflation Method:

- Involves a Dense matrix-vector multiplication in second step.

- Was doubling the running time.

Other Clustering Approaches

- Various “modularity” based algorithms by Mark Newman.

- In 'Finding community structure in networks using the eigenvectors of matrices', Newman suggests a method which involves computing the leading eigenvector of Modularity Matrix.

- Modularity matrix is ***not sparse .*** As a result the sparse matrix-vector multiplication in power method becomes a dense matrix-vector multiplication.

- We also looked into the work of Wakita , Tsurumi from Tokyo Institute of Technology . They had implemented a faster version of an algorithm proposed by Clauset , Newman and Moore(2004)

Future work

- A strategy of **refining the clusters** obtained from the algorithm.

- Scaling up for the Episims Graph (which has 1.6 million nodes.)
  - Introducing some heuristics for the matrix multiplication
  - Some local optimization

- Analyze the disease spread on the clusters

- Relation between Modularity and Conductance.

References

1. On clustering : Good , Bad And Spectral .[Kannan ,et al]

2. On a Recursive Spectral Algorithm for Clustering from Pair wise Similarities.[Cheng , Kannan ,Vempala , Wang]

3. A divide And Merge Methodology for Clustering [Cheng , Kannan , Vempala , Wang]

4. Finding community structure in networks using eigenvectors of matrices.[M.E.J Newman]

5. Modularity And Community Structure In networks[M.E.J Newman]

6. The Effect of Network Topology On the Spread of Epidemics [A. Ganesh]

7. A Tutorial On spectral Clustering[Ulrike Von Luxburg]

Thank You.