

Clustering and Applications to Biodiversity

Presented by: Alassane Ngaide, Frederic Anglade
Mentors: Dr. Urmi Ghosh Dastidar, Dr. Gene Fiorini

Mathematic Department
NYCCT, CUNY DIMACS, Rutgers University
Date:07/19/2012

Problem

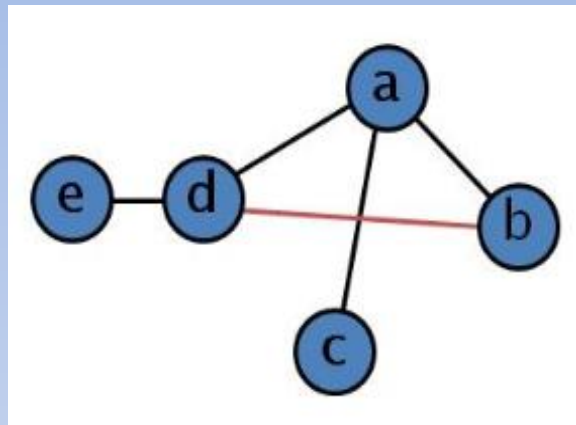
- *Given the competition graph $G = (V, E)$, is it possible to partition the competition graph G into two sub-graphs G_A and G_B using a combination of Fiedler order and linkage-based refinements to minimize $\text{cut}(A, B)$ while maximizing $W(A)$ and $W(B)$ at the same time?*
- *Note: The strength between two nodes is given by their edge weight W_{ij} and the strength between two clusters A and B is given by : $\text{cut}(A, B) = W(A, B)$ where:*

$$W(A, B) = \sum W_{ij} \quad i \in A \quad j \in B$$

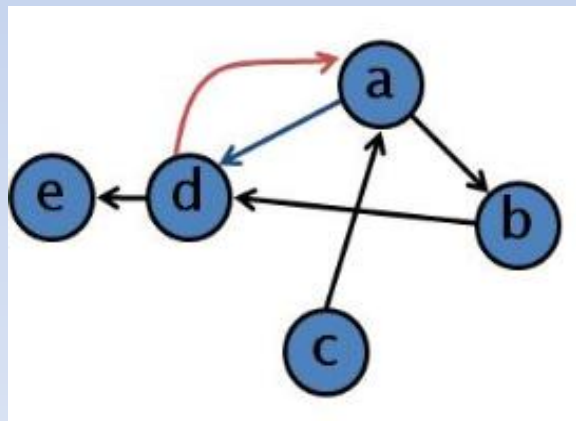
Adjacency Matrix

- An adjacency matrix is an $n \times n$ matrix where the ij -th entry takes the value of 1 if there is an edge between vertices i and j . Otherwise, the element is 0.

$$A = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{Otherwise} \end{cases}$$



0	1	1	1	0
1	0	0	1	0
1	0	0	0	0
1	1	0	0	1
0	0	0	1	0



0	1	0	1	0
0	0	0	1	0
1	0	0	0	0
1	0	0	0	1
0	0	0	0	0

Spectral Clustering

Uses information obtained from the Eigenvalues and eigenvectors of the Laplacian matrix for partitioning of graphs;

Laplacian

- Given A , the Laplacian matrix L is defined as:
- $L = D - A$

$$L = \begin{cases} \deg(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \sim v_j \\ 0 & \text{otherwise} \end{cases}$$

Normalized Laplacian

- Given A , the normalized Laplacian L is defined as:

$$L = \begin{cases} 1 & \text{if } i = j \text{ and } \deg(v_i) \neq 0 \\ -\frac{1}{\sqrt{\deg(v_i)\deg(v_j)}} & \text{if } i \neq j \text{ and } v_i \sim v_j \\ 0 & \text{otherwise} \end{cases}$$

Basic spectral bi-clustering algorithm

- Partitions a graph into two clusters
- Nodes within the same cluster, vertices are more connected to each other than with those in the other cluster.

Input: Weighted Laplacian Matrix

Find the eigenvector v corresponding to the second smallest eigenvalue for one of the following problems:

$Lv = \lambda v$ (L : Laplacian),

$L''v = \lambda v$ (L'' : Normalized Laplacian).

Output: Clusters $A = \{j; v_j \geq 0\}$ and $A' = \{j; v_j < 0\}$

Min-max cut

- Given a weighted graph $G=G(E,V)$ with node set V , edge set E and weight W partition G into two sub-graphs A, B using the min-max clustering principle.
- Minimize similarity between sub-graphs.
- Maximize similarity within sub-graph.
- Similarity between sub-graphs A, B is the cut size:
 $\text{cut}(A,B)=W(A,B)$ where:

$$w(A,B) = \sum_{u \in A, v \in B} w_{uv} \quad w(A) \equiv w(A, A) \quad w(B) \equiv w(B, B)$$

The Mcut Function

- Minimize $\text{cut}(A, B)$ while maximizing $W(A)$ and $W(B)$ at the same time.
- Objective function to be minimized:

$$Mcut = \frac{\text{cut}(A, B)}{W(A)} + \frac{\text{cut}(A, B)}{W(B)}$$

Fiedler vector

- Eigenvector associated with the second smallest eigen-value.
- Believed to provide the best linear search order for finding the optimum cut.
- However, it is possible to have nodes sharing higher linkage to the other cluster than the one they are currently assigned to by using only the information of Fiedler order.

$$cut(A, B) = \sum_{e_{uv} \in E} \frac{(x_u - x_v)^2}{4} w_{uv} = \frac{x^T (D - W) x}{2}$$

$$(D - W) x = \lambda x$$

Linkage-based refinement

- Identify the nodes near the cut.
- Define linkage l as a similarity measure between two sub-graphs:

$$l(A, B) = \frac{W(A, B)}{W(A)W(B)}$$

- Linkage between a node and a sub-graph:

$$l(A, u) = \frac{W(A, u)}{W(A)} \qquad l(B, u) = \frac{W(B, u)}{W(B)}$$

Linkage difference

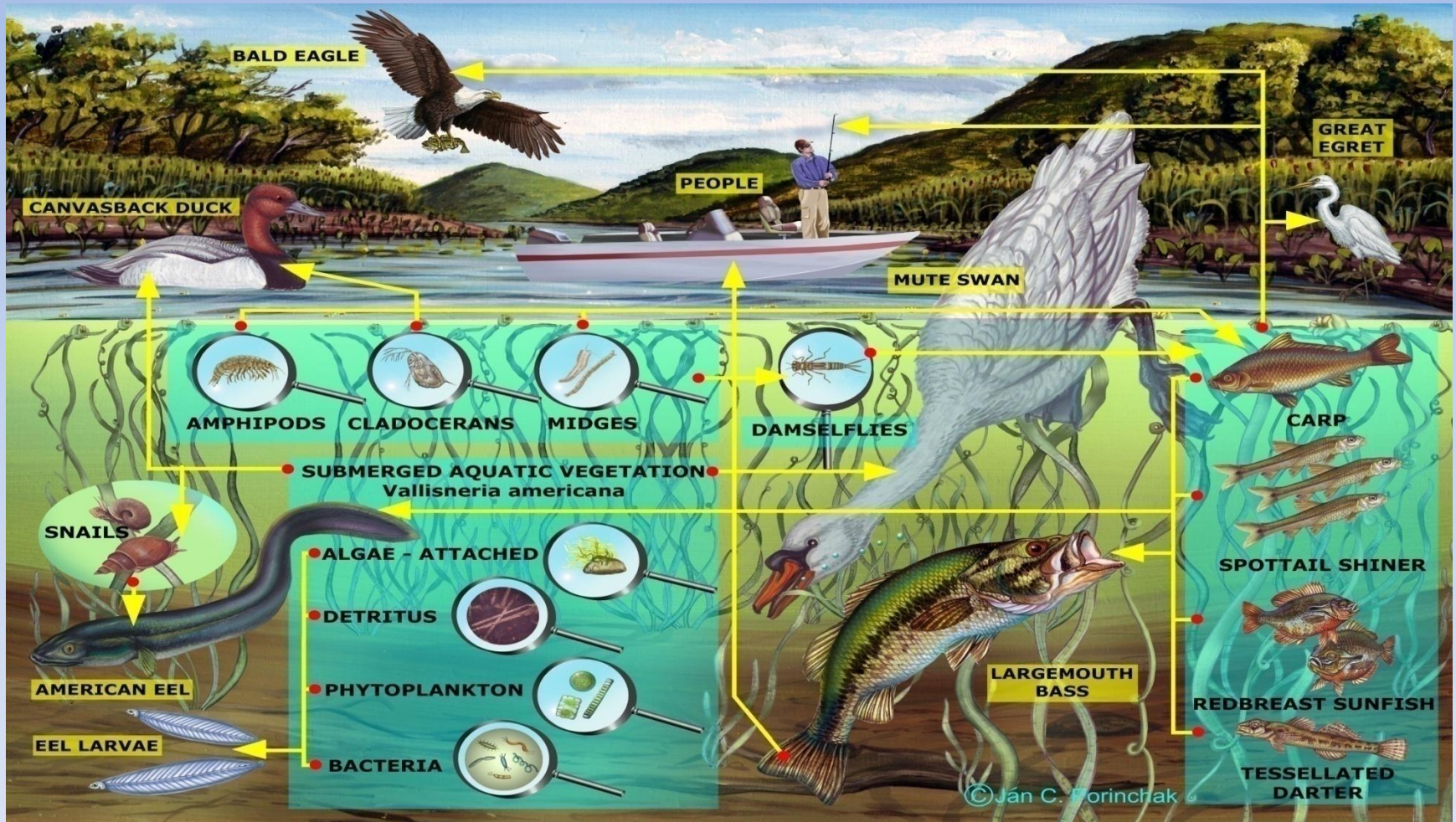
- The linkage difference: $\Delta l(u) = l(A, U) - l(B, U)$
- If u is well inside A we expect $\Delta l(u) \succ 0$
- If u is well inside B we expect $\Delta l(u) \prec 0$
- If u is near the cut we expect $\Delta l(u) \approx 0$
 - In this case it might be a possible candidate to move to a different cluster if the objective function M_{cut} is reduced

Combination of the two methods

It is observed that a *linkage differential order* provides a better ordination than the **Fiedler order** therefore we implement a combination of Fiedler order with linkage differential order to get a better result on minimizing the cut between cluster while maximizing $W(A)$ and $W(B)$ at the same time.

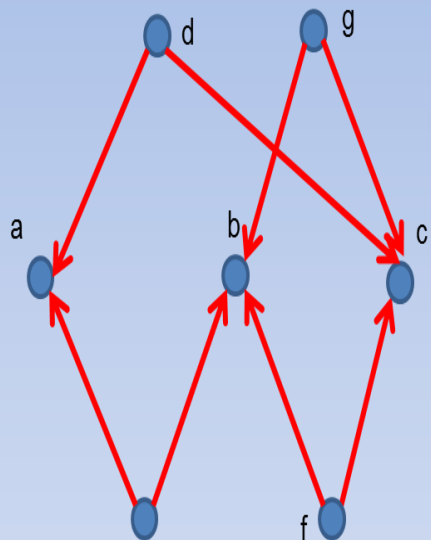
Application

Hudson River Food Web

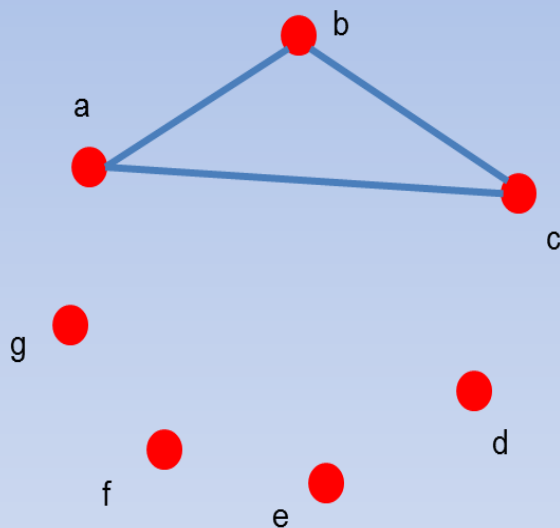


Competition graph for Hudson River Food Web

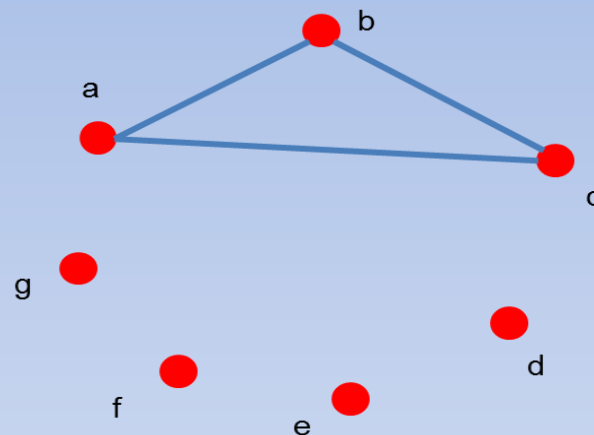
- A **competition graph** is a graph where the vertices are species in the ecosystem and there is an edge between two vertices if they have a common prey. If vertices are isolated, they either do not have any prey in common with the other species of the ecosystem or they are primary producers. For example, in our Hudson River Competition Graph, the isolated vertices are bacteria, phytoplankton, polychates, algae, water celery, eel larvae, spartina alterniflora, typha, phragmites, snapping turtle and great blue heron.



Let D be a Digraph
with node set $V(D) = 7$



$\{v_i, v_j\} = 1$, If $\exists v_k \mid (v_i, v_k) \cap (v_j, v_k)$
are **directed edges** in $E(D)$



Weighted
Competition
Graph

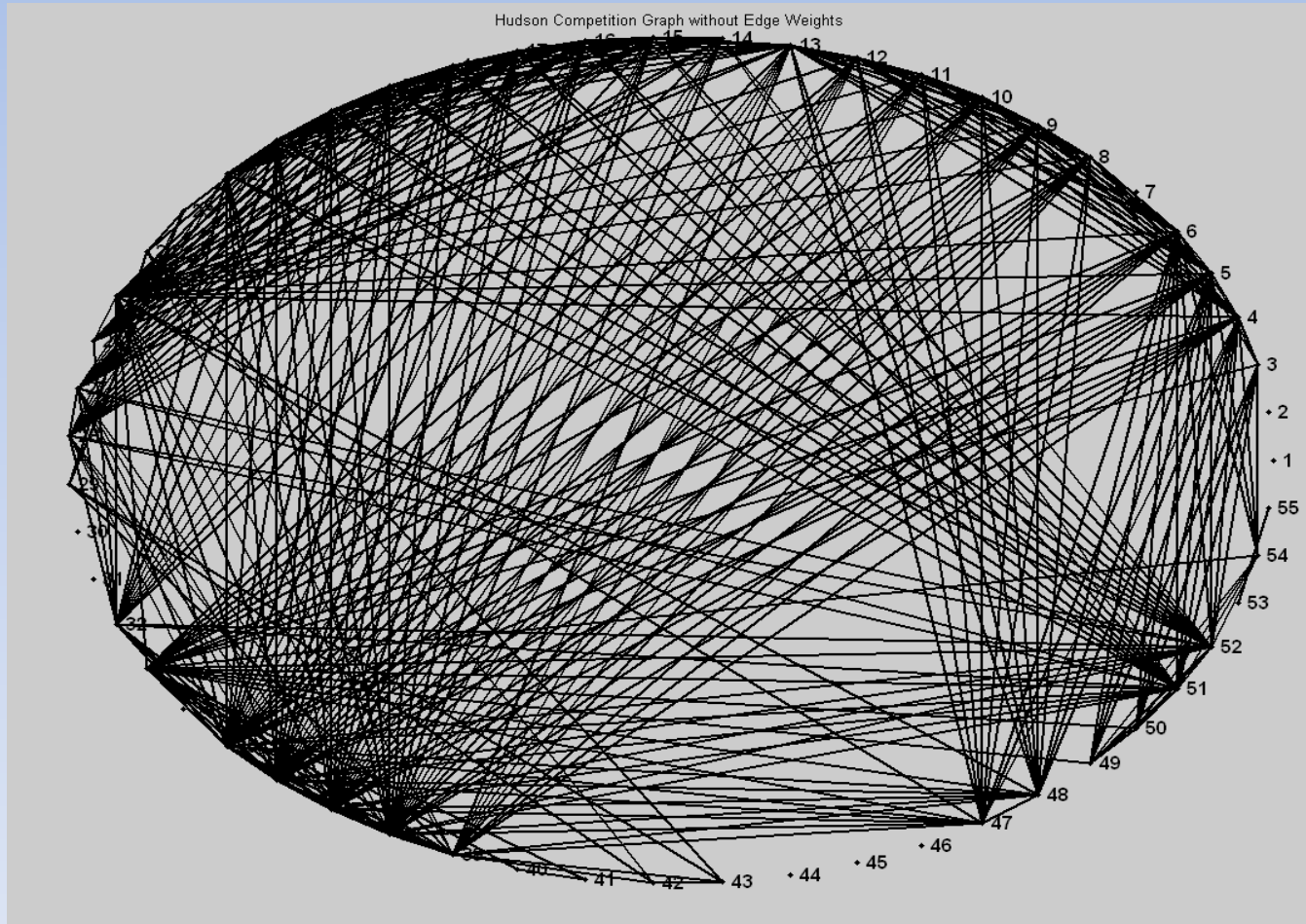
Adjacency matrix

Hudson Competition

[illegible]

Bipartition

Hudson Weighted Competition Graph



Fiedler Vector corresponding to the smallest non-zero Eigen Value for 44 nodes

- Eigen-value 1.0191
- Node 35 and 35 less connected

0.0291
0.0296
0.0296
0.0295
0.0305
0.0305
0.0305
0.0305
0.0305
0.0315
0.0336
0.0335
0.0335
0.0335
0.0335
0.0335
0.0335
0.0340
0.0336
0.0448
0.0439
0.0314
0.0348
0.0338
0.0340
0.0443
0.0339
0.0132
0.0316
0.0316
0.0316
0.0316
0.0334
-0.6904
-0.6904
0.0443
0.0443
0.0318
0.0318
0.0291
0.0291
0.0318
0.0318
0.0291

Fiedler Vector corresponding to the smallest non-zero Eigen Value For 42 nodes

- Eigen-value 3.7991

```
0.0874
0.0503
0.0503
0.0488
0.0463
0.0463
0.0463
0.0463
0.0463
0.0508
0.0631
0.0616
0.0616
0.0616
0.0616
0.0616
0.0432
0.0631
-0.3337
-0.4089
0.0506
0.0701
0.0648
0.0562
-0.4363
0.0658
0.0534
0.0515
0.0515
0.0515
0.0515
-0.0658
-0.4443
-0.4443
0.0554
0.0554
0.0874
0.0874
0.0679
0.0679
0.0874
```

Fiedler order

+

Linkage based refinements

- Algorithm
 1. Computer the Fiedler Vector. Sort the nodal values to obtain the Fiedler order
 2. Search for optional cut point corresponding to the lowest Mcut based on the Fiedler order.
 3. Do linkage-based refinements

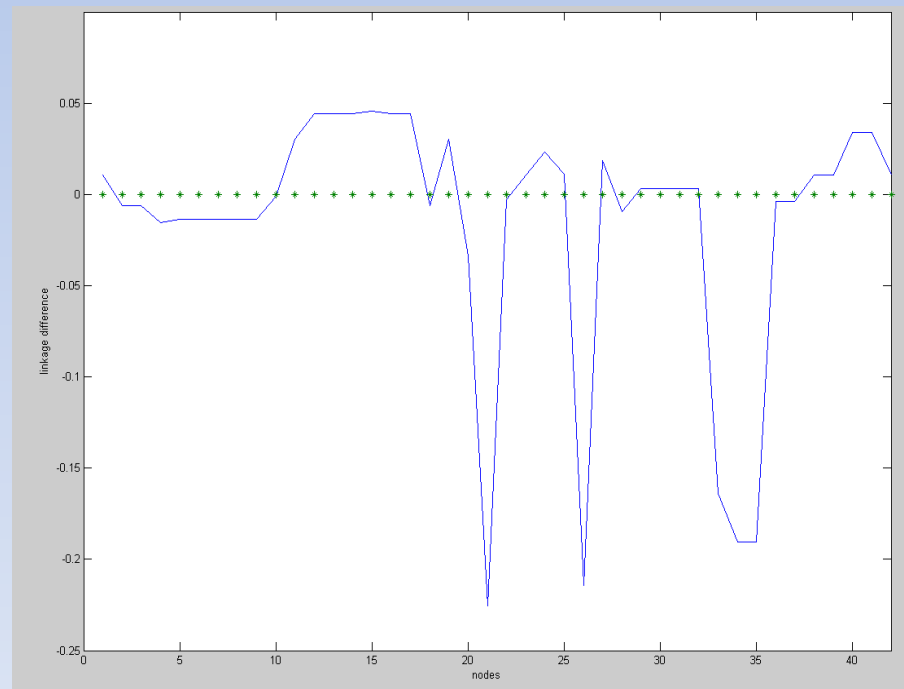
Objective Function

$M_{\text{cut}} = 3.3746$

Cluster A 36 nodes

Cluster B 6 nodes

The linkage difference of all nodes



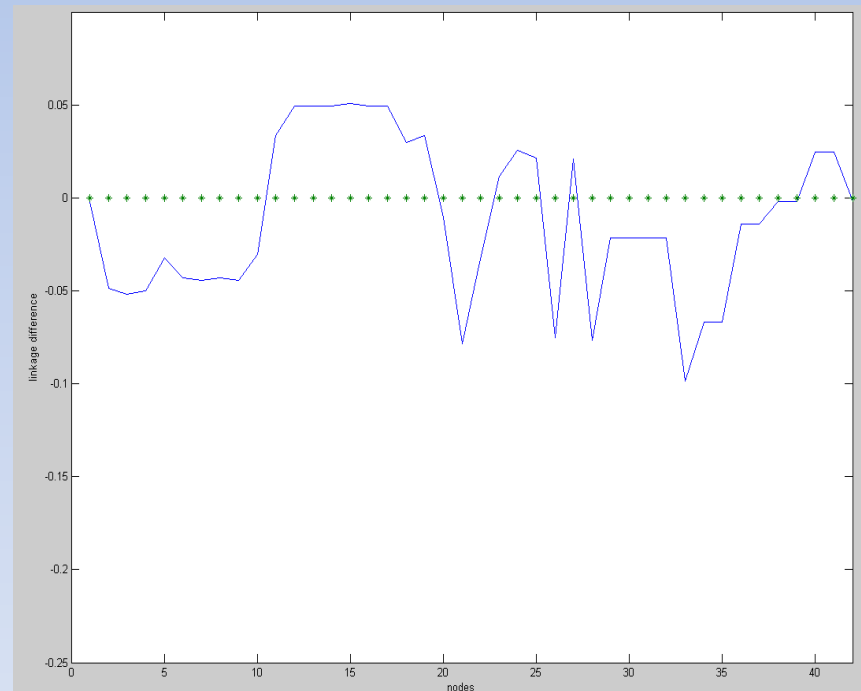
Improvement Results, Analyses

Mcut = 2.0986

Cluster A 32 Nodes

Cluster B 10 Nodes

The linkage difference of all nodes



References

- Reference:
 - * “Food Webs, Competition Graphs, and Habitat Formation” Margaret B. Cozzens, DIMACS, Rutgers University
 - * Nir Ailon, Moses Charikar, Alantha Newman.(2008). Aggregating inconsistent information: Ranking and clustering. J. ACM 55(5)
 - * Ding. C. HQ et al..(2001). A Min-max Cut Algorithm for Graph partition and clustering. IEEE conference Proceeding. pp . 107-114
 - * Chung, F. R. K. (1997). *Spectral graph theory*. Providence, RI: American Mathematical Society.
 - * Dy, J. G., & Brodley, C. E. (2004). Feature selection for unsupervised learning. *J. Mach. Learn. Res.*, 5, 845–889.
 - * Hagen, L., & Kahng, A. (1992). New spectral methods for radio cut partitioning and clustering. *IEEE Transactions on Computer-Aided Design*, II(9), 1074–1085.

Acknowledgements

- Special thanks to our mentor Dr. Urmi Ghosh-Dastidar who introduced us to the program and assisted in every step of the way.
- To Dr. Fiorini, the DIMACS and the MAA for giving us this opportunity.
- To our Professors : Dr. Arnavaz P Taraporevala and Dr. Norman Horowitz for their recommendation
- To all our fellow researchers for the wonderful time.

**Thank You All for
your Attention**