Community discovery in social network: Applications, methods and emerging trends

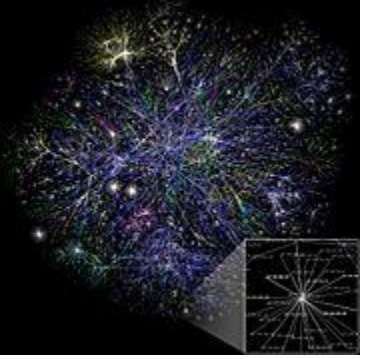
L16,17-SNA

- Graph representing a society, what do you see in layman’s world?
- Can you identify patterns?
- Analysis of patterns
- What is a community?

- Community is a closely knit group of nodes having common interest
- Entities of interest and interaction among them
- In this chapter we shall study the Methods of community discovery and its variants

- Different domains, different communities, several problems
- But they share some similar concepts

- Ex: Say one generalized algorithm developed that can be applied on various datasets



- Network Science
- Study of complex networks such as telecommunication, biological network, social network etc. This field draws its theories from graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science.

- Why is extracting a community challenging?
- Topological properties of network coupled with an uncertain setting limit the applicability of existing off the shelf techniques
- Requirements imposed by directed and dynamic network require research into appropriate solutions
- Scalability
  - Million node, billion edges

We focus on community detection in social networks

- Actionable patterns or tools one can derive from such an analysis on social network
- Role of communities in Twitter/ FB during emergency
- Hierarchical algorithms- do not scale well
  - Agglomerative
  - Divisive
- Hybrid algorithms
- Community discovery in heterogeneous social networks
- - community evolution, dispersion, merging

- Community discovery
  - In directed social network
  - That combines content and network information in natural manner
  - Ex: Topic driven community discovery, social media analysis

Communities in context

- One of the earliest studies in this context include work by Rice on the analysis of communities of individuals based on their political biases and voting patterns followed by many others who analysed network of various domains trying to interpret any patterns
- The Karate club study is a well known graph regularly commonly used as a bench mark for community detection algorithms
- A large majority of this study focuses on social structure and its evolution

- Study of various domains and the identification of influential herd leaders and how animals communicate and socialize to survive is studied
- Grouping web clients of similar interest and viral marketing
- E-commerce, personalized recommendations etc
- Community discovery used for Analysing online social media data
- - Ex: Depression prediction based on this

- Community discovery
  - Is helpful in understanding the social system
  - Helps in Summarizing interactions within the network and enforce better understanding in the social phenomenon
  - Actionable pattern discovery
  - Identification of influential nodes, sub-communities used for viral marketing within tele communication networks and ratings predictions
  - Emergency management
  - Mitigating the impact of disasters

Core Methods

- Informally, a community in a network is a group of nodes with greater ties internally than to the rest of the network.
- This intuitive definition has been formalized in a number of competing ways, usually by way of a **quality function**
- Quality function quantifies the goodness of a given division of the network into communities.
- Some of these quality metrics, such as Normalized Cuts and Modularity
- No single metric is applicable in all situations

- Algorithms for community discovery optimize a single quality metric
- Quality function
- A variety of quality functions or measures have been proposed in the literature to capture the goodness of a division of a graph into clusters.

The normalized cut of a group of vertices $S \subset V$ is defined as[96, 67

$$Ncut(S) = \frac{\sum_{i \in S, j \in \bar{S}} A(i,j)}{\sum_{i \in S} degree(i)} + \frac{\sum_{i \in S, j \in \bar{S}} A(i,j)}{\sum_{j \in \bar{S}} degree(j)}$$

Groups with low normalized cut make for good communities, as they are well connected amongst themselves but are sparsely connected to the rest of the graph.

The *conductance* of a group of vertices $S \subset V$ is closely related and is defined as [50]

$$Conductance(S) = \frac{\sum_{i \in S, j \in \bar{S}} A(i,j)}{\min(\sum_{i \in S} degree(i), \sum_{i \in \bar{S}} degree(i))} \tag{4.2}$$

- The normalized cut (or conductance) of a division of the graph into *k* clusters
- *V*1, . . . , *Vk* is the sum of the normalized cuts (or conductances) of each of
- the clusters *Vi{i* = 1*, . . . , k}* .
- The *Kernighan-Lin (KL) objective* looks to minimize the edge cut (or the sum of the inter-cluster edge weights) under the constraint that all clusters be of the same size

$$KLObj(V_1, \ldots, V_k) = \sum_{i \neq j} A(V_i, V_j) \text{subject to} |V_1| = |V_2| = \ldots = |V_k| \tag{4.3}$$

Here $A(V_i, V_j)$ denotes the sum of edge affinities between vertices in V_i and V_j, i.e. $A(V_i, V_j) = \sum_{u \in V_i, v \in V_j} A(u, v)$

- *Modularity* has recently become quite popular as a way to measure the goodness of a clustering of a graph.
- One of the advantages of modularity is that it is independent of the number of clusters that the graph is divided into.
- The intuition behind the definition of modularity is that the farther the subgraph corresponding to each community is from a random subgraph, the better or more significant the discovered community structure is.

The modularity Q for a division of the graph into k clusters $\{V_1, \ldots, V_k\}$ is given by:

$$Q = \sum_{c=1}^{k} \left[\frac{A(V_i, V_i)}{m} - \left(\frac{degree(V_i)}{2m} \right)^2 \right] \tag{4.4}$$

In the above, the V_is are the clusters, m is the number of edges in the graph and $degree(V_i)$ is the total degree of the cluster V_i. For each cluster, we take the difference between the fraction of edges internal to that cluster and the fraction of edges that would be expected to be inside a random cluster with the same total degree.

Optimizing any of these objective functions is NP-hard [39, 96, 18].

Basic definition

A **cut** $C = (S, T)$ is a partition of V of a graph $G = (V, E)$ into two subsets *S* and *T*. The **cut-set** of a cut $C = (S, T)$ is the set $\{(u, v) \in E \mid u \in S, v \in T\}$ of edges that have one endpoint in *S* and the other endpoint in *T*. If *s* and *t* are specified vertices of the graph *G*, then an ***s*–*t* cut** is a cut in which *s* belongs to the set *S* and *t* belongs to the set *T*.

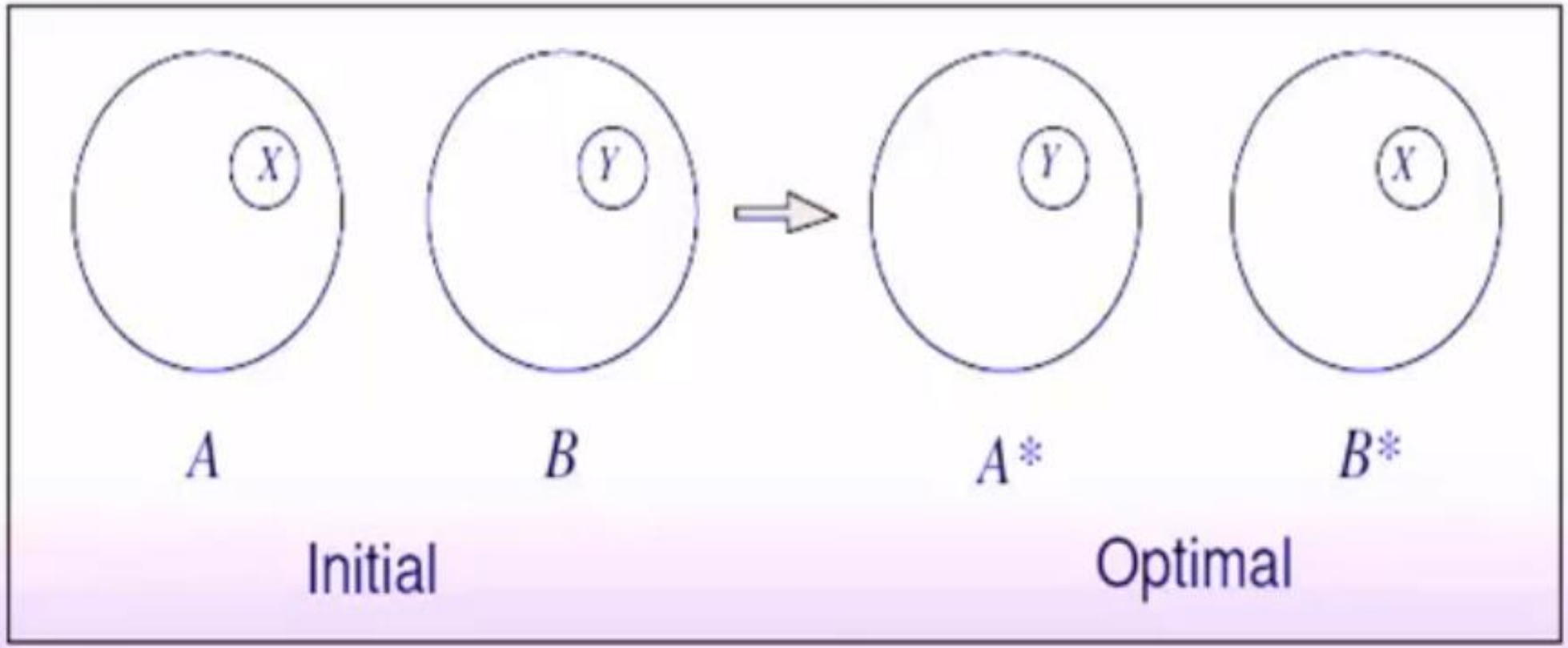
Kernighan-Lin Algorithm

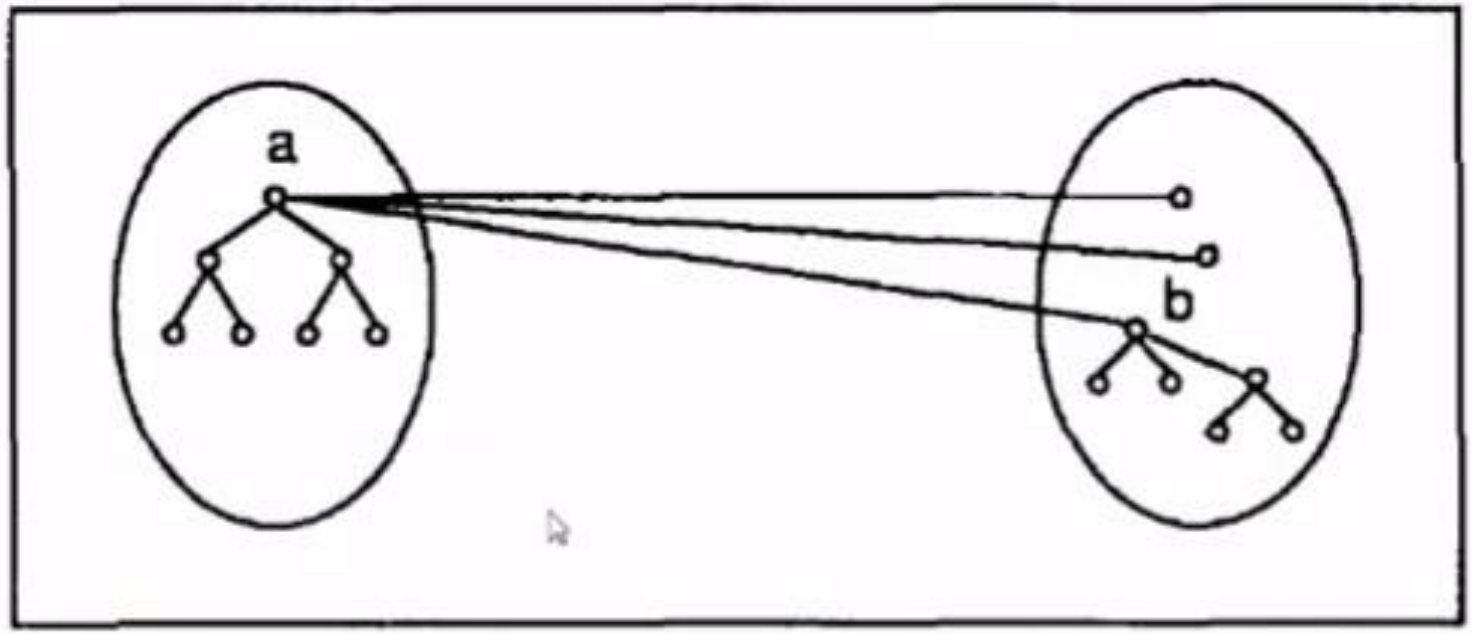
- Most popular algorithm for the two-way partitioning problem.
- Kernighan-Lin - iterative improvement algorithm. It starts from an initial partition (A,B) such that $|A| = n = |B|$, and $A \cap B = \emptyset$
- Let $P = \{A,B\}$ be the initial partition and $P^* = \{A^*,B^*\}$ be the optimum partition.
- Then, in order to attain P^* from P, one has to swap a subset X of A with a subset Y of B such that,

  (1) $|X| = |Y|$

  (2) $X = A \cap B^*$

  (3) $Y = A^* \cap B$


X
Y
Y
X
A
B
A*
B*
Initial
Optimal



Definition 1:

Consider any node a in block A. The contribution of node a to the cutset is called the external cost of a and is denoted as E_a where

$$E_a = \sum_{v \in B} c_{av}$$

Definition 2:

The internal cost I_a of node $a \in A$ is defined as follows.

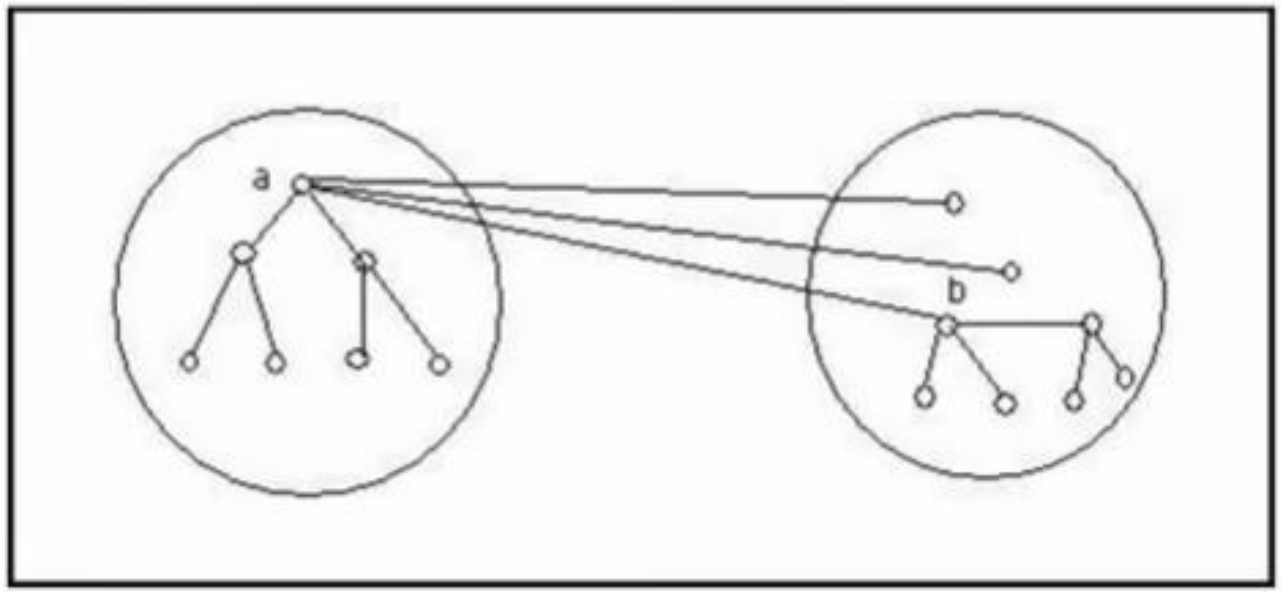
$$I_a = \sum_{v \in A} c_{av}$$

Definitions

Moving node **a** from block A to block B would increase the value of the cutset by I_a and decrease it by E_a.
Therefore, the benefit/profit of moving **a** from A to B is
$D_a = E_a - I_a$

Example

Consider the figure with, I_a=2, I_b=3, E_a=3, E_b =1, D_a = 1, and D_b = -2.



Internal cost versus external costs

Example contd..

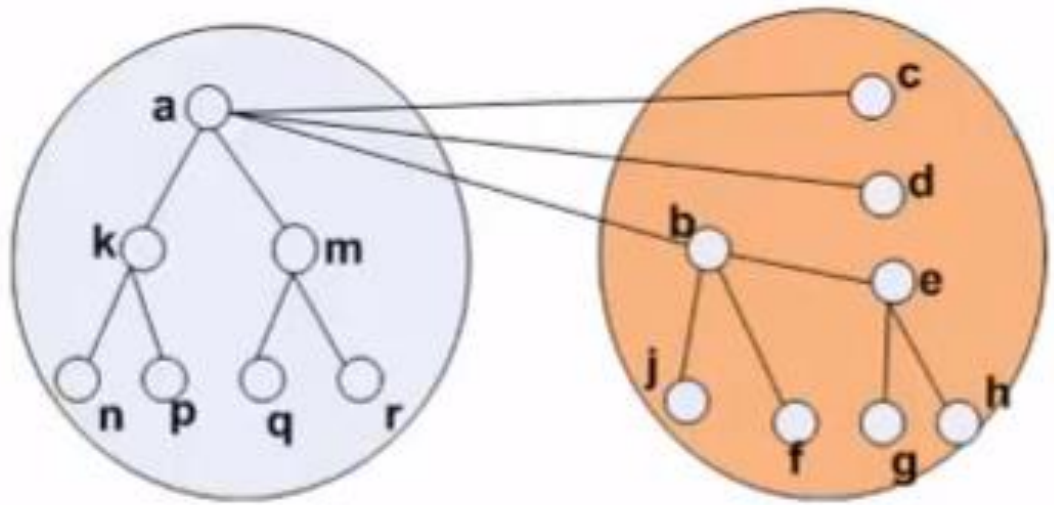
To maintain balanced partition, we must move a node from B to A each time we move a node from A to B.

The effect of swapping two modules a ε A with b ε B is characterized by the following lemma.

Lemma 1:

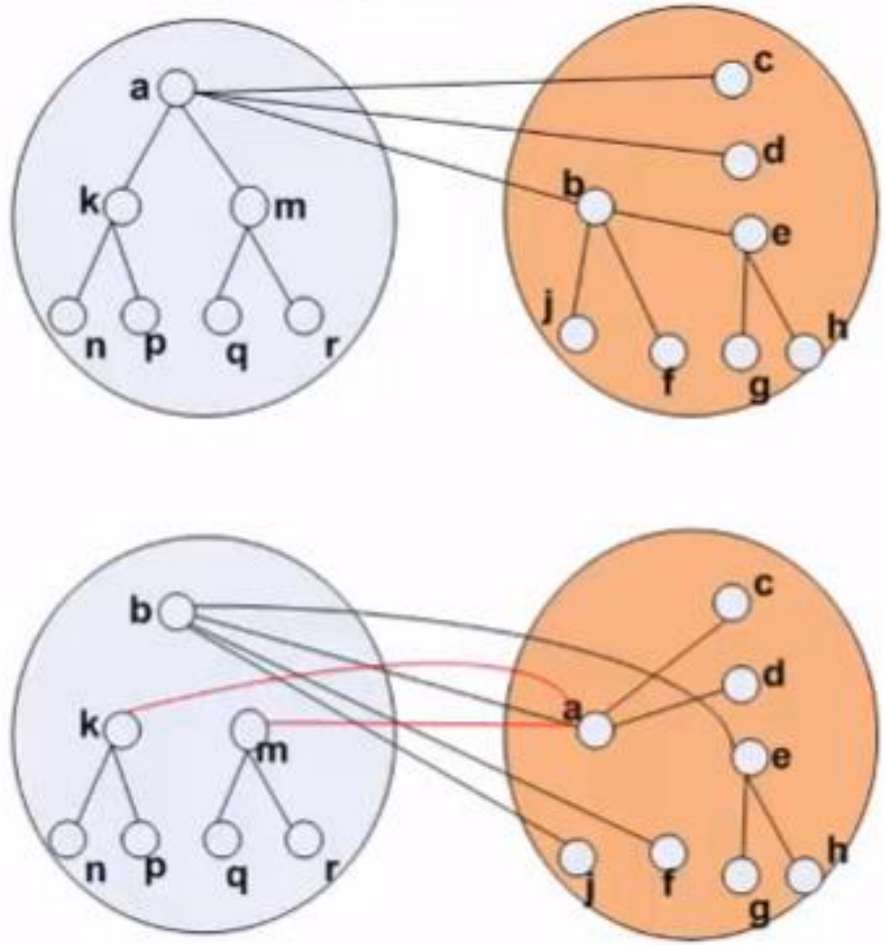
If two elements a ϵ A *and* b ϵ B *are interchanged, the reduction in the cost is given by*

$$g_{ab} = D_a + D_b - 2c_{ab}$$

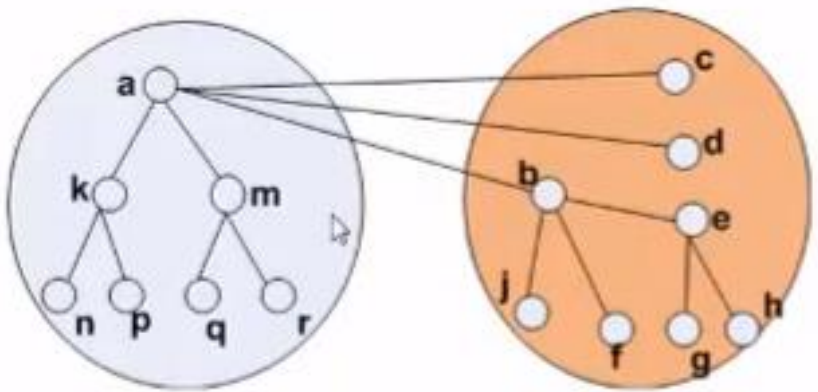


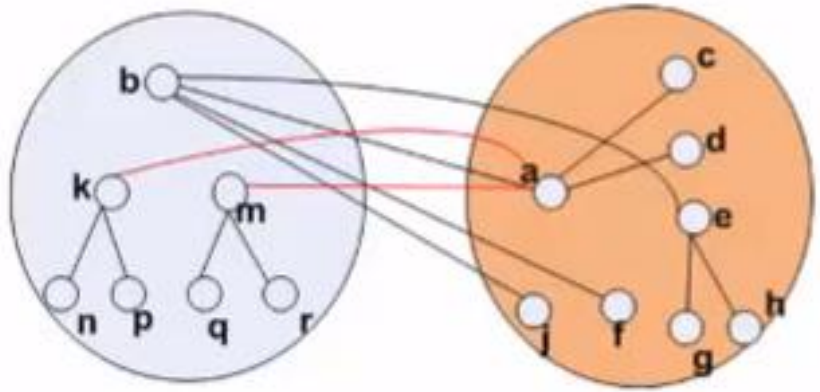
Swap a and b across the partition and find the cutset

Cut set=3



If a and b are exchanged across the partition, the cut set becomes 6





$g_{ab} = D_a + D_b - 2c_{ab}$

$D_a=1; D_b=-2$

$g_{ab} = 1 + (-2) - 2$

$g_{ab} = -3$

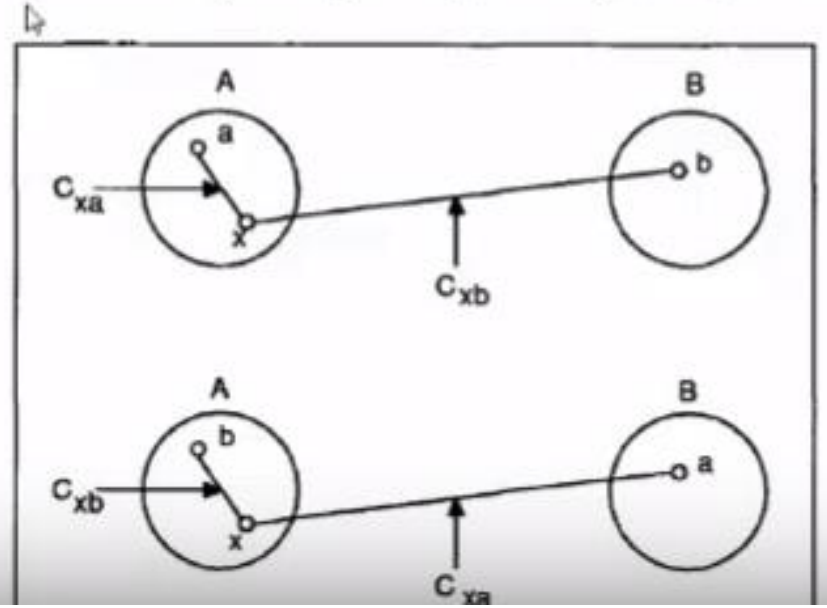
Cut set has increased by 3
Gain = -3

Swapping affects nodes attached to the swapped nodes

Lemma 2.2 *If two elements $a \in A$ and $b \in B$ are interchanged, then the new D−values, indicated by D', are given by*

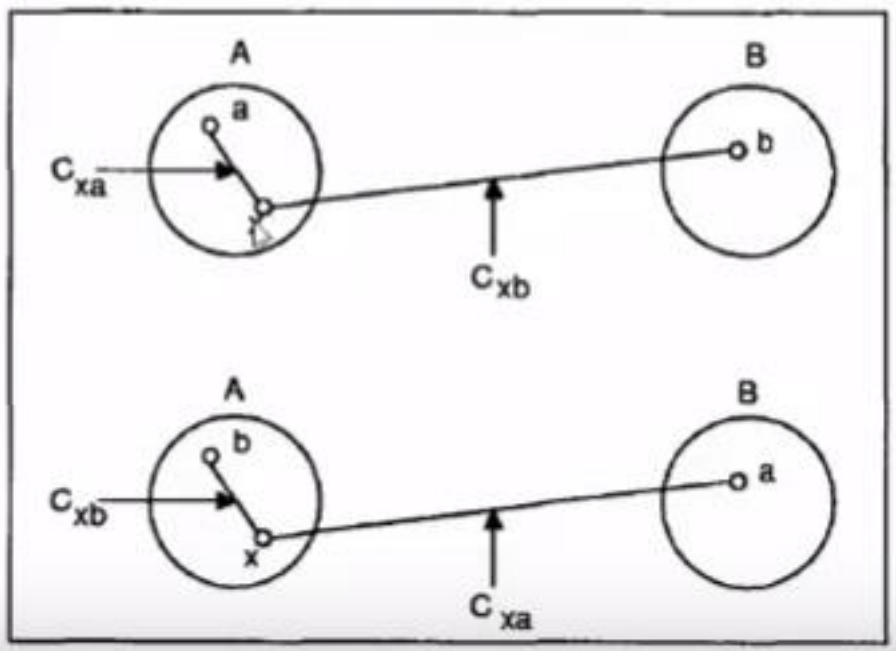
$$D'_x = D_x + 2c_{xa} - 2c_{xb}, \quad \forall x \in A - \{a\} \tag{2.17}$$

$$D'_y = D_y + 2c_{yb} - 2c_{ya}, \quad \forall y \in B - \{b\} \tag{2.18}$$



- Consider a node $x \varepsilon A - \{a\}$: Since b has entered partition A, the internal cost of x increases by c_{xb}.
- Similarly, since a has entered the opposite partition B; the internal cost of x must be decreased by c_{xa}.
- The new internal cost of x therefore is

$$I_x' = I_x - c_{xa} + c_{xb}$$

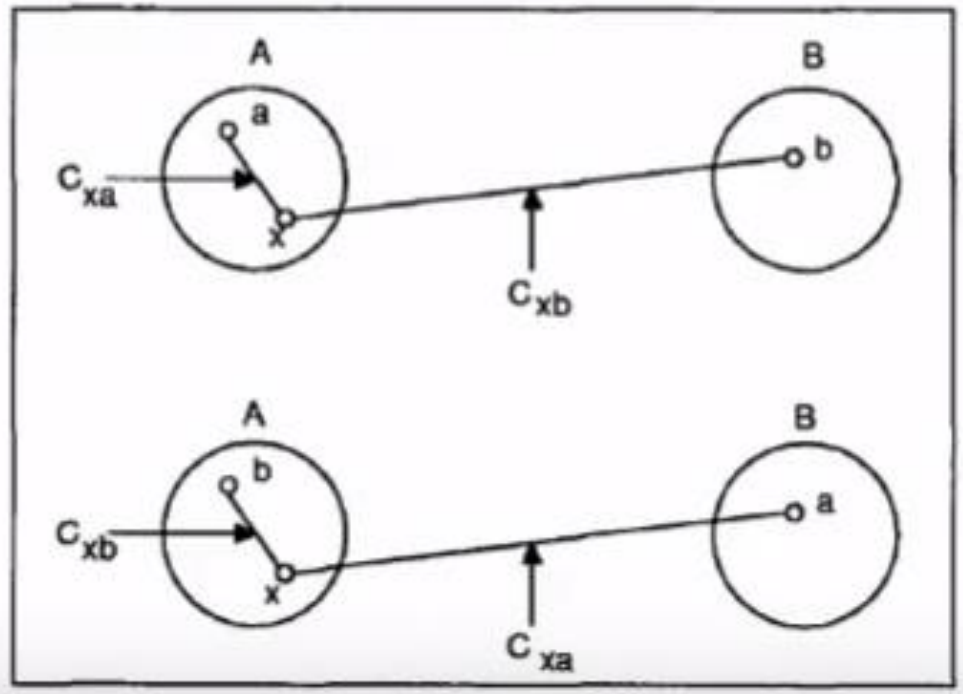


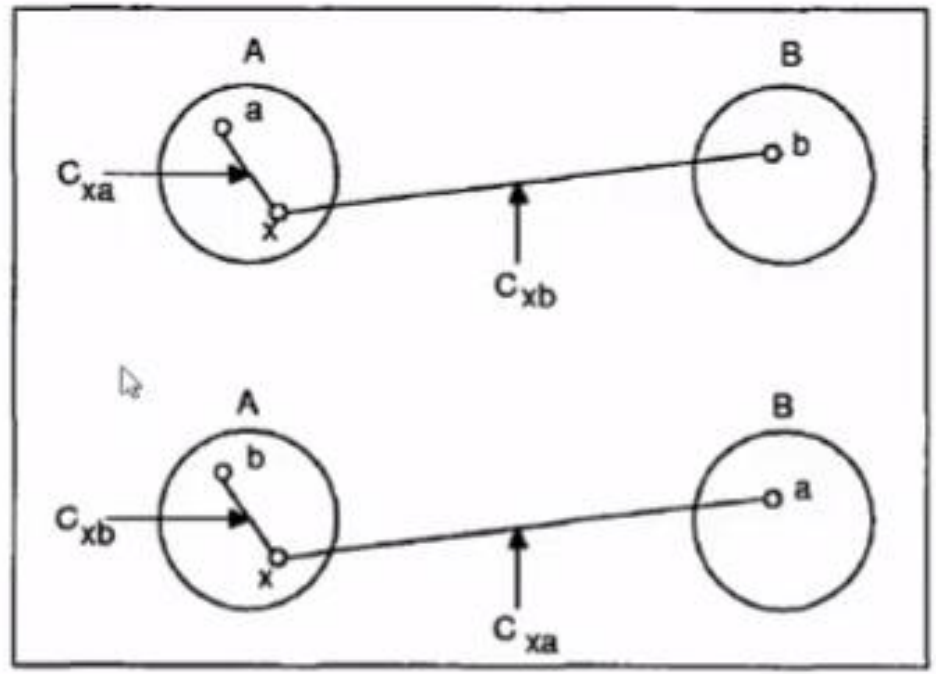
$E_x' = E_x + c_{xa} - c_{xb}$ and $I_x' = I_x - c_{xa} + c_{xb}$

For any node $x \varepsilon A-\{a\}$, we define updated D value as

$D_x' = E_x' - I_x'$

$D_x' = D_x + 2\, c_{xa} - 2\, c_{xb}$





Similarly, the new $D-$value of $y \in B - \{b\}$ is

$$D_y^{'} = E_y^{'} - I_y^{'} = D_y + 2c_{yb} - 2c_{ya}$$

Notice that if a module 'x' is neither connected to 'a' nor to 'b' then $c_{xa} = c_{xb} = 0$, and, $D_x^{'} = D_x$.

Overview of K-L algorithm

- Compute g_{ab} for all a ε A and b ε B.
- Select the pair (a_1 , b_1) with maximum gain g_1 and lock a_1 and b_1.
- Update the D values of remaining free nodes and re-compute the gains.
- Then a second pair (a_2, b_2) with maximum gain g_2 is selected and locked. Hence, the gain of swapping the pair (a_1, b_1) followed by the (a_2, b_2) swap is $G_2 = g_1 + g_2$.
- Continue selecting (a_3, b_3), ….., (a_i, b_i), …, (a_n , b_n) with gains g_3,….., g_i, ….., g_n.
- The gain of making the swap of the first k pairs is $G_k = E\ g_i$. If there is no k such that $G_k > 0$ then the current partition cannot be improved; otherwise choose the k that maximizes G_k, and make the interchange of {a_1, a_2, ….., a_k} with {b_1, b_2, … ,b_k} permanent.