Mining Data Graph

COMMUNITY DETECTION

Lecturer: Le Ngoc Thanh

Email: lnthanh@fit.hcmus.edu.vn

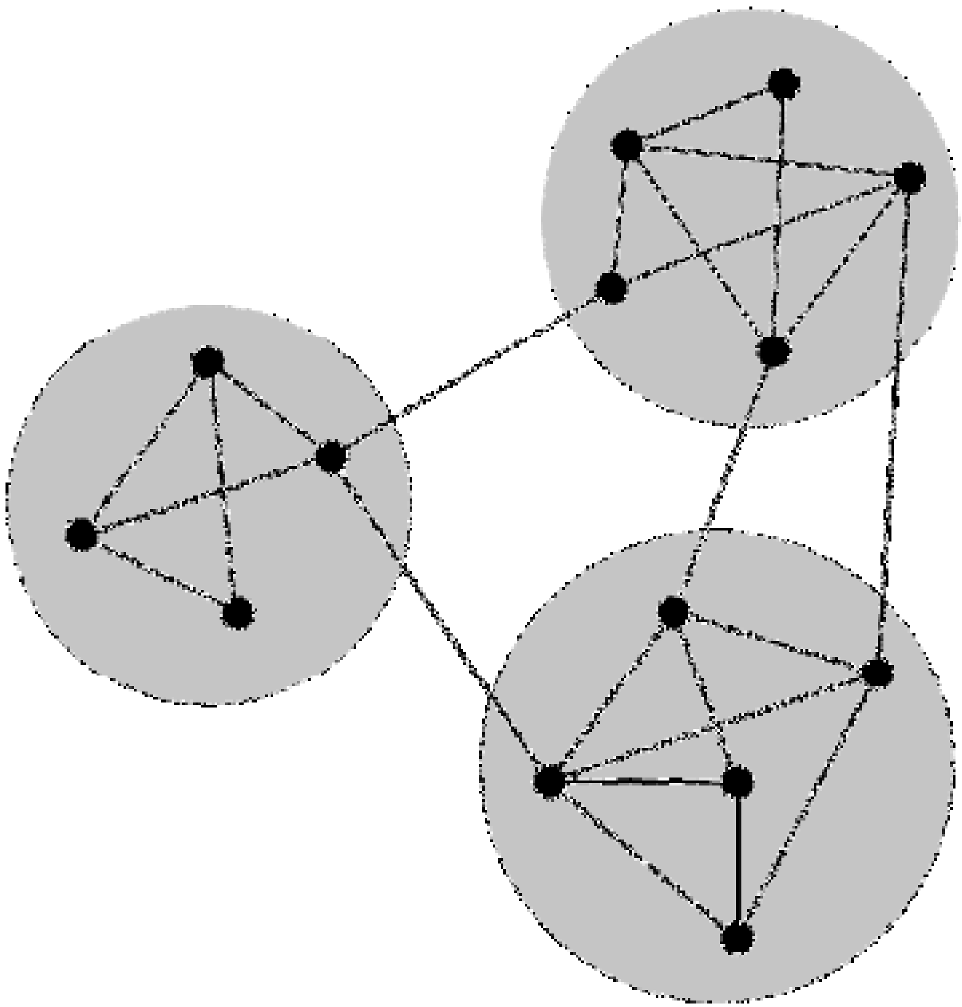
fit@hcmus

Content

- **Concept**
- Community optimization
- Community detection methods
  - Methods based on minimal slices
  - Intermediate-based method
  - Methods based on RandomWalk

Community

- Community is the set of vertices where each vertex has more internal connections than outward connections.

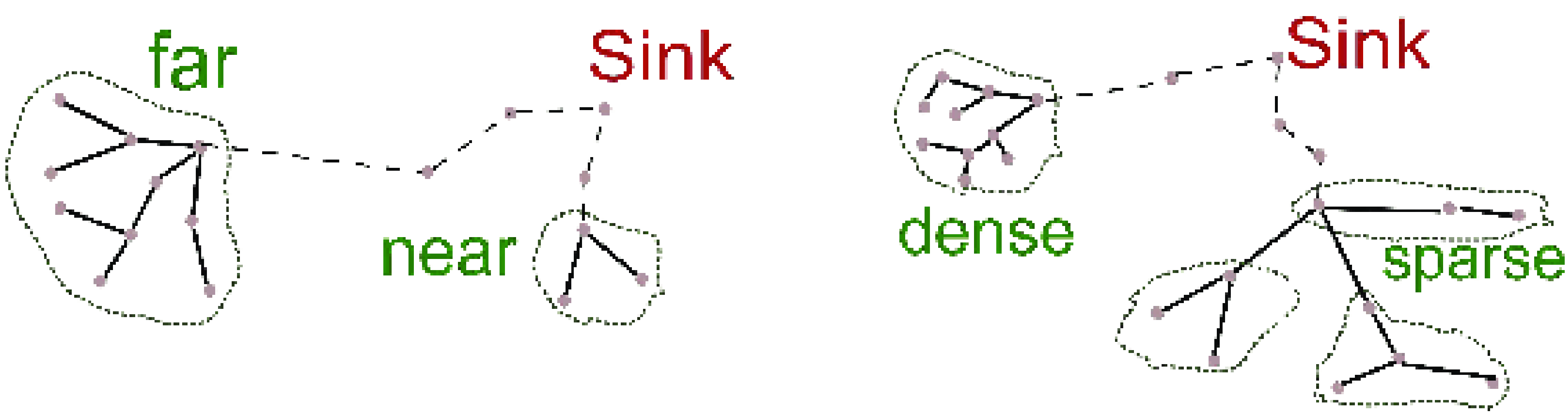


Content

- Concept
- **Community optimization**
- Community detection methods
  - Methods based on minimal slices
  - Intermediate-based method
  - Methods based on RandomWalk

Community optimization

- In addition to the quality and distance measurements, one has 2 more ways of measuring to assess density:
  - Intracluster density: the bigger the better
  - Intercluster density: the smaller the better

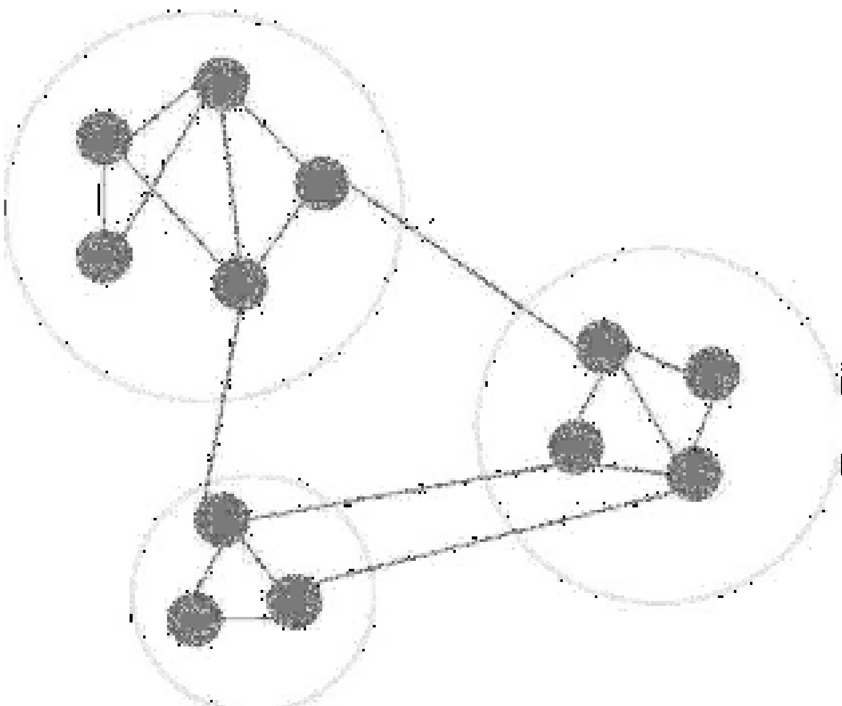


(a) distance (b) density

Density in the group

- Group density is the ratio between the number of edges inside the group and the maximum number of possible edges in the group.

- $\delta(C) = \dfrac{number\ of\ edges\ in\ the\ group}{N_C(N_C-1)/2}$



Example:

$$\delta(C_1) = \frac{7}{10} = 0.7$$

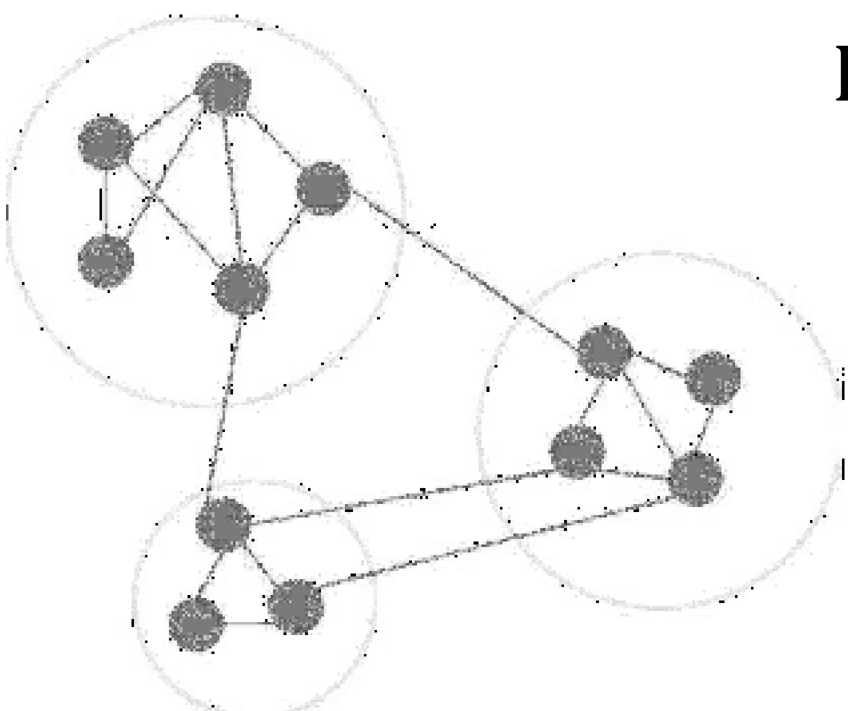
$$\delta(C_2) = \frac{4}{6} = 0.75$$

$$\delta(C_3) = \frac{3}{3} = 1.0$$

Out-of-group density

- Out-of-group density is the ratio of the number of out-group edges to the number of possible out-group edges.

- $\epsilon(C) = \dfrac{number\ of\ edges\ outside\ the\ group}{N_c(N - N_c)}$



Example:

$$\epsilon(C_1) = \frac{2}{35}$$

$$\epsilon(C_2) = \frac{3}{32}$$

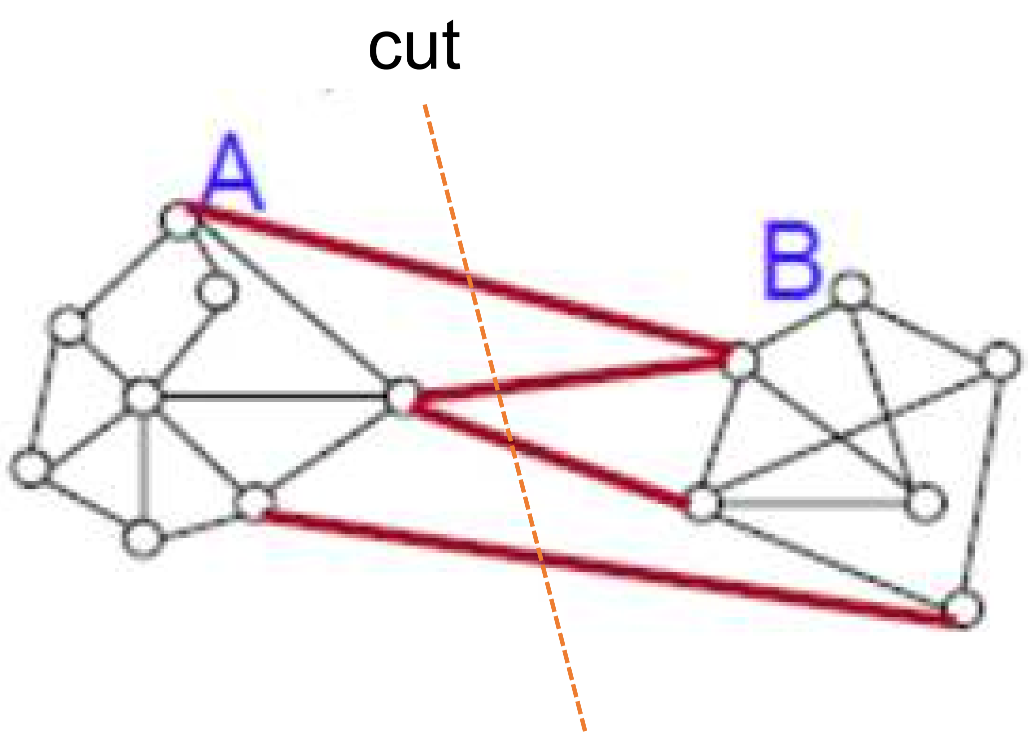
$$\epsilon(C_3) = \frac{3}{27}$$

Content

- Concept
- Community optimization
- **Community detection methods**
  - Methods based on minimal slices
  - Intermediate-based method
  - Methods based on RandomWalk

Minimum slice

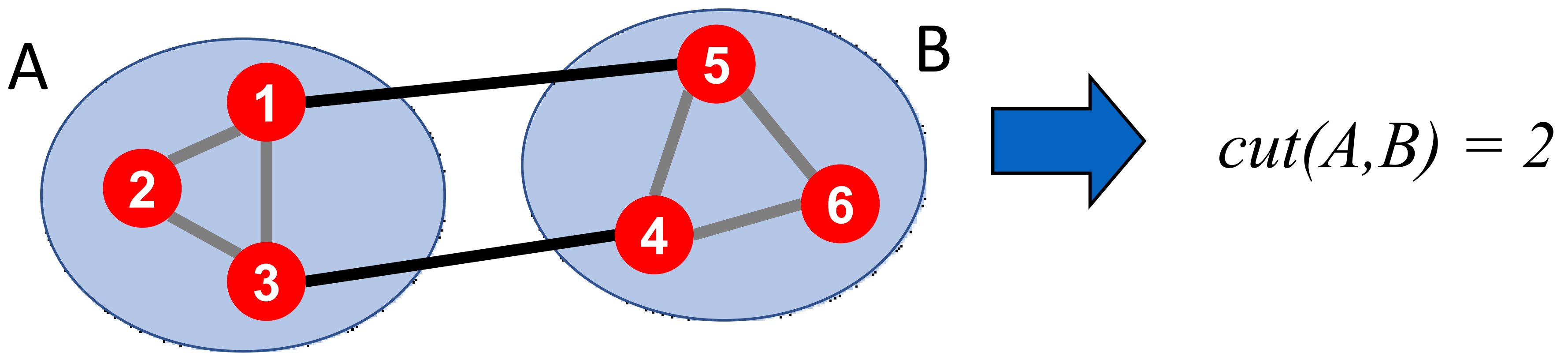
- The goal of the minimum slice is to find the least edge set that blocks the flow from source S to T.
  - The cutting size is the total weight of those edges



Minimum slice

- Slice:

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

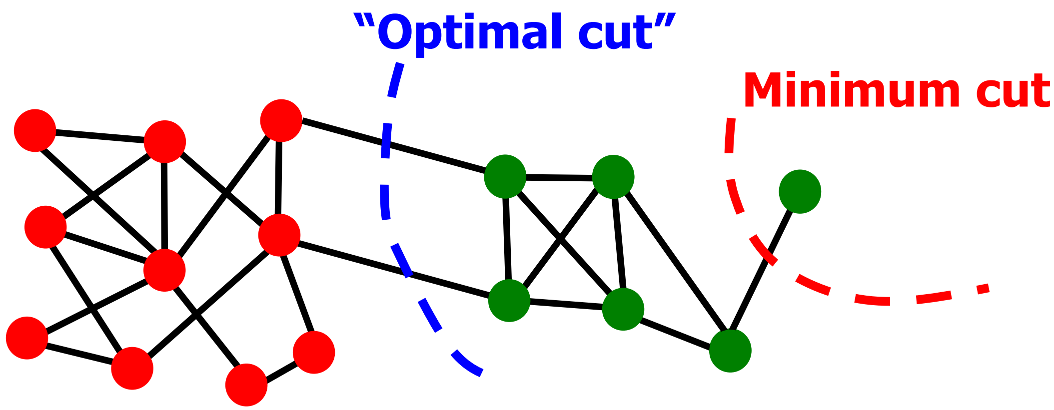


Community detection with minimal slice

- Goal: minimal slice

$$\arg\min_{A,B} cut(A,B)$$

- Problem:



The problem occurs due to not considering the internal connection

Slice normalization

- The normalized slice depends on the density in each group

$$ncut(A,B) = \frac{cut(A,B)}{vol(A)} + \frac{cut(A,B)}{vol(B)}$$

with $vol(A)$: total weight of edges with at least one end in A

$$vol(A) = \sum_{i \in A} k_i$$

Content

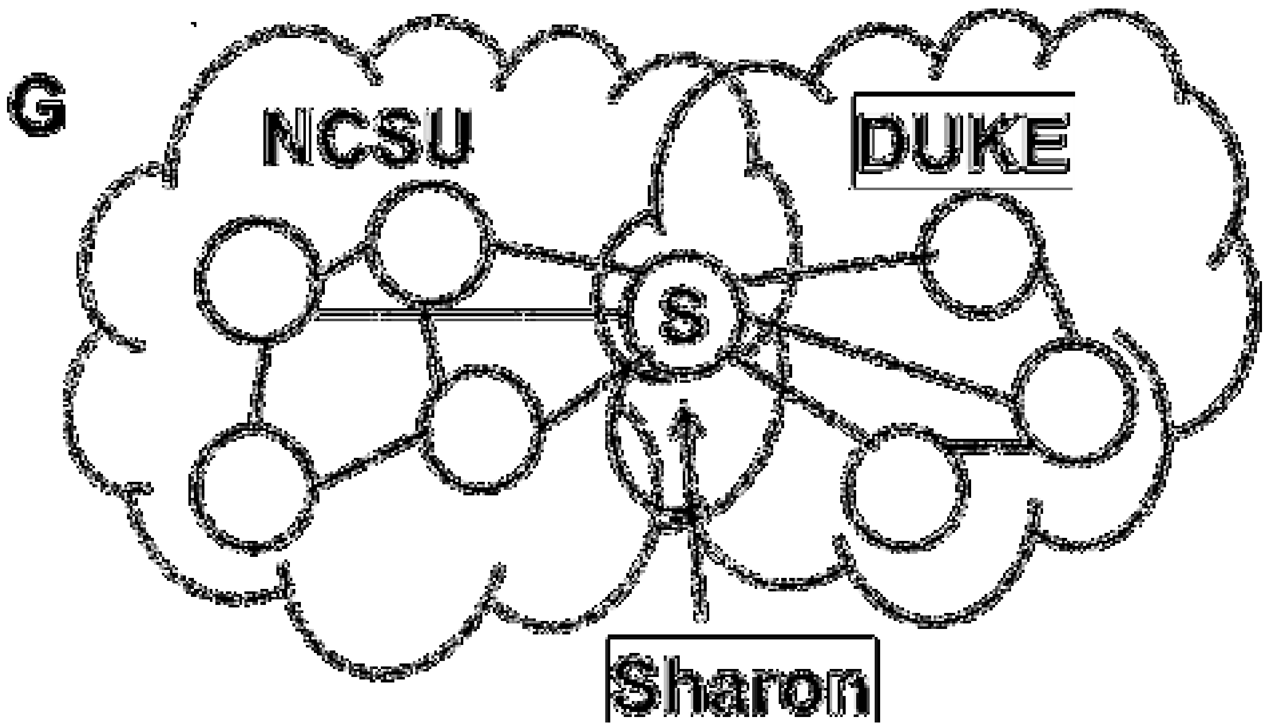
- Concept
- Community optimization
- **Community detection methods**
  - Methods based on minimal slices
  - Intermediate-based method
  - Methods based on RandomWalk

Community detection based on mediation

- Intermediate-based community detection performs the process of community identification by removing the highly intermediate vertex/edge.
- There are two types:
  - Based on peak intermediation
  - Based on edge intermediation.

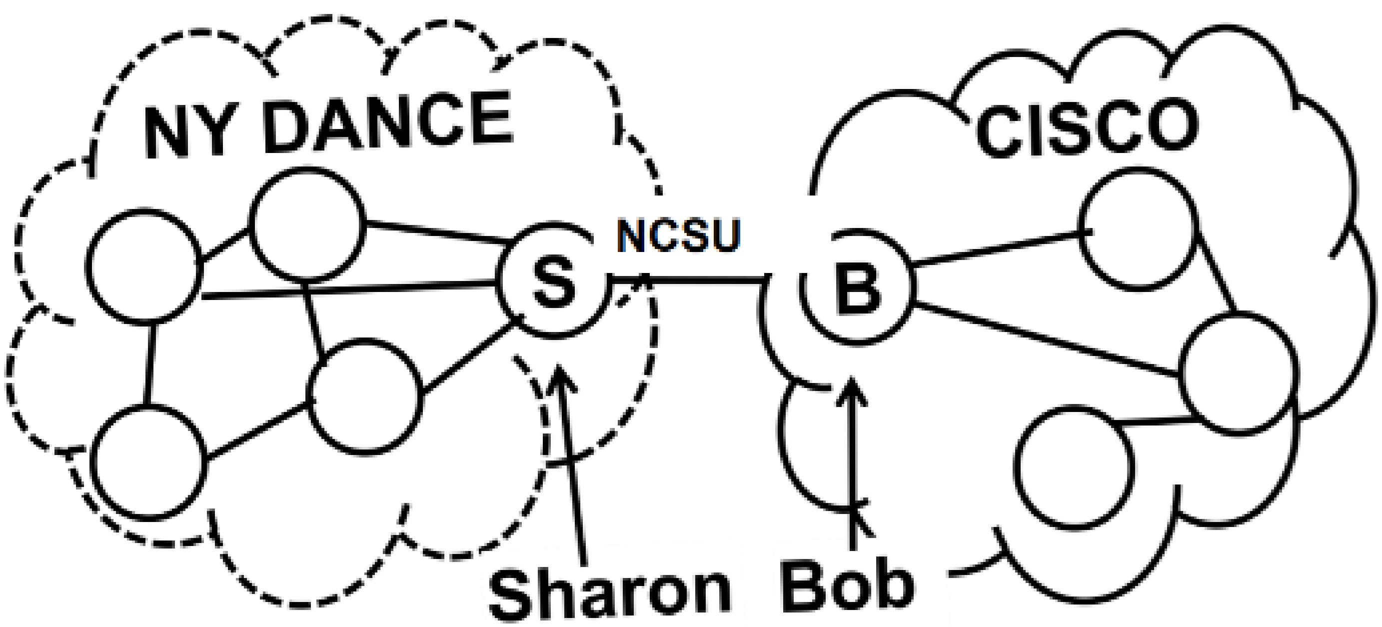
Peak intermediate

- The vertex intermediate is calculated based on the number of shortest paths in the graph that must pass through a given vertex.

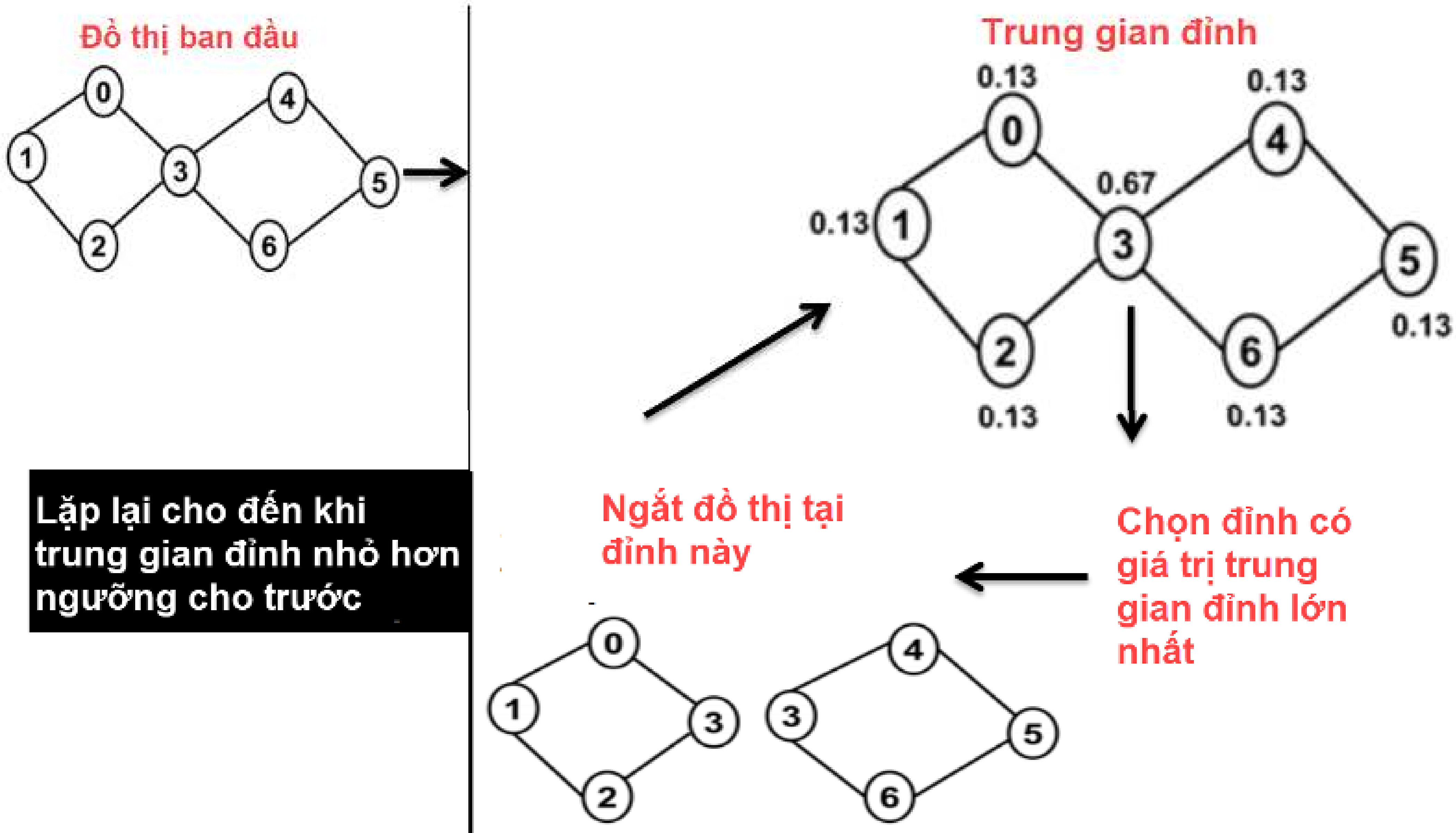


Edge intermediate

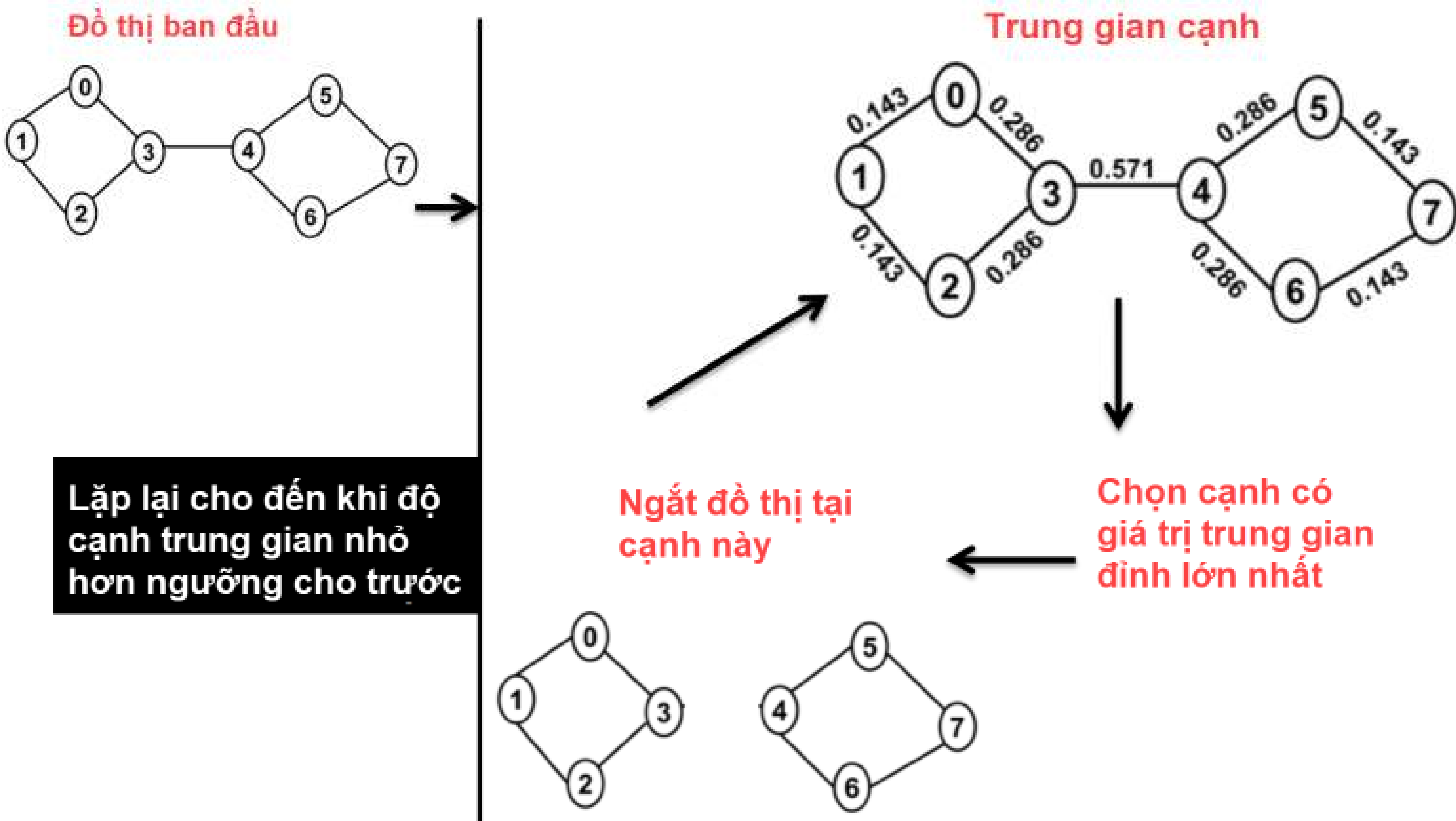
- Edge intermediates are evaluated based on the number of shortest paths that must pass through a given edge.



The vertex intermediate algorithm



Edge Intermediate Algorithm

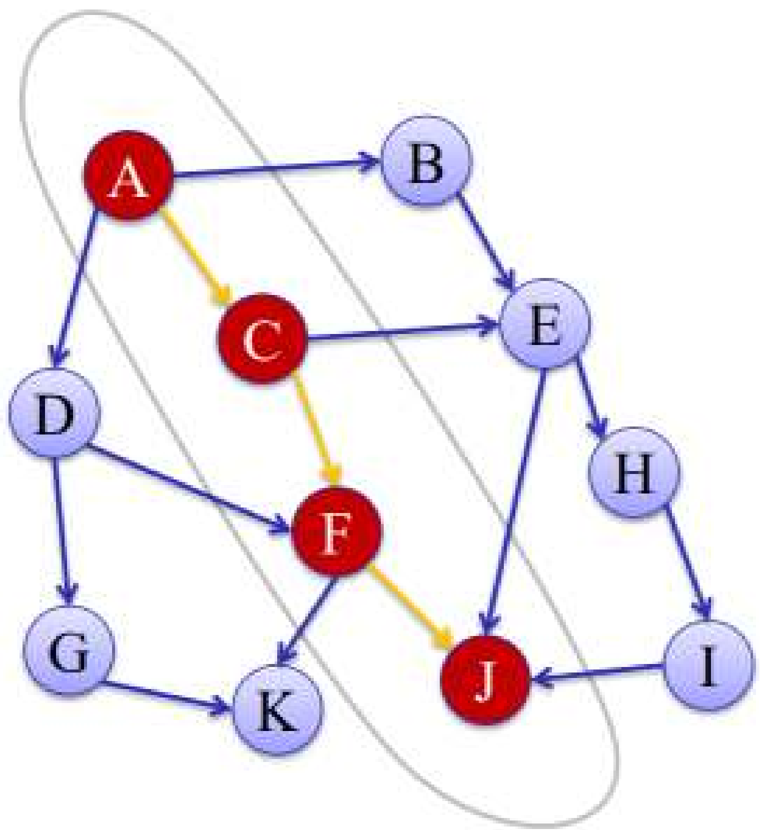


Content

- Concept
- Community optimization
- **Community detection methods**
  - Methods based on minimal slices
  - Intermediate-based method
  - Methods based on RandomWalk

Random Walk

- Given a graph, from a starting vertex, we randomly choose a vertex to continue. After t such iterations, we will have a random walk of size t.

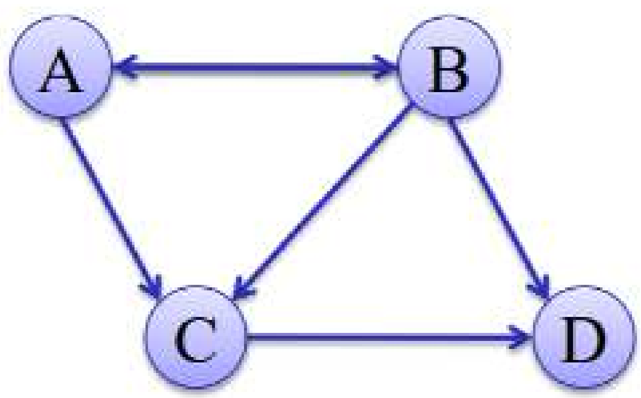


Random Walk

- The probability of selecting an edge can be evaluated based on the correlation between the two texts and is normalized.

Transition Matrix

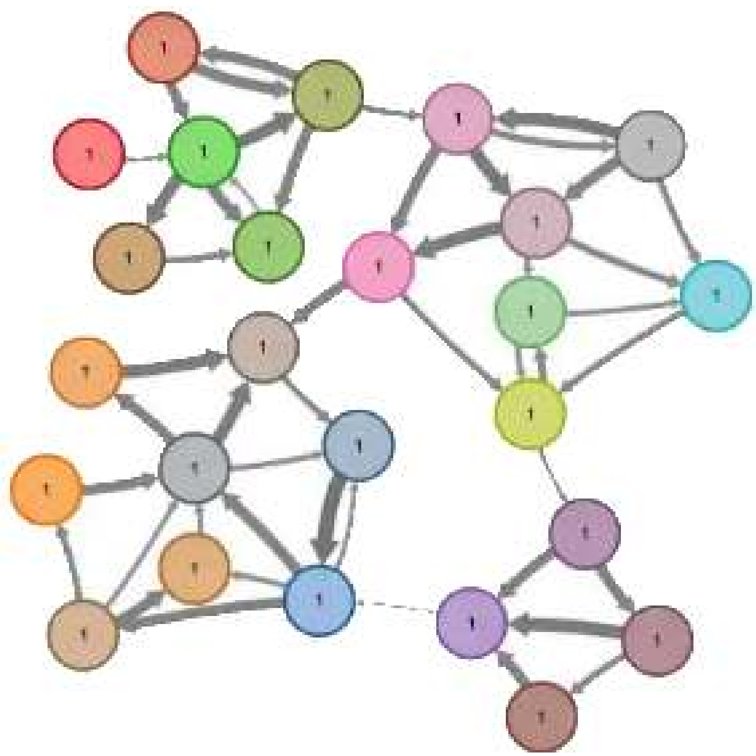
$$\begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$



Community detection based on random walk

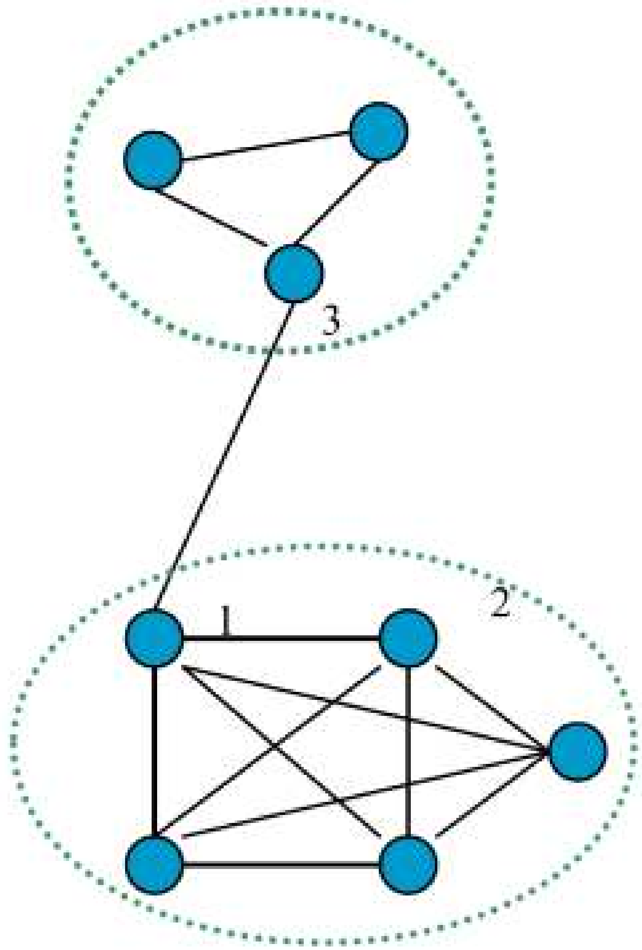
- Community detection based on random walk is done based on the following statement:

  A random path that begins at a vertex is more likely to move in one community than in another.



Example

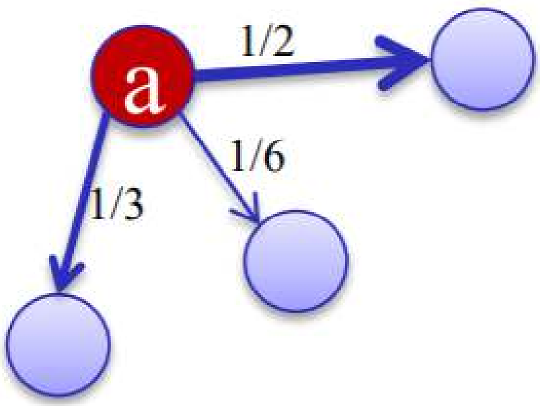
Node	Prob. Next Step within cluster	Prob. Next Step between clusters
1	80%	20%
2	100%	0%
3	67%	33%

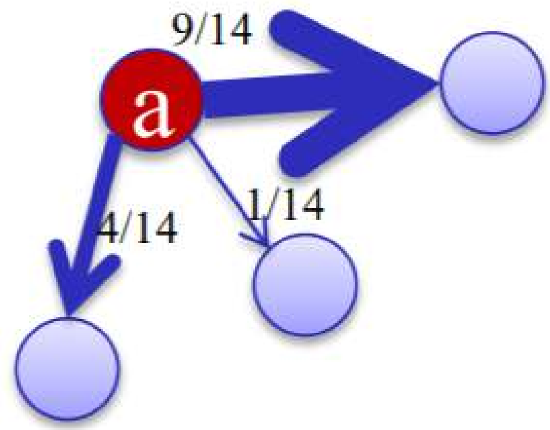


Algorithmic ideas

- Adjust the weight so that after a random walk of a given size, the likelihood of movement in the group will be high.
- The weighting is adjusted so that:
  - Stronger neighbors will get stronger
  - Weaker neighbors will get weaker
  - This process is called inflation

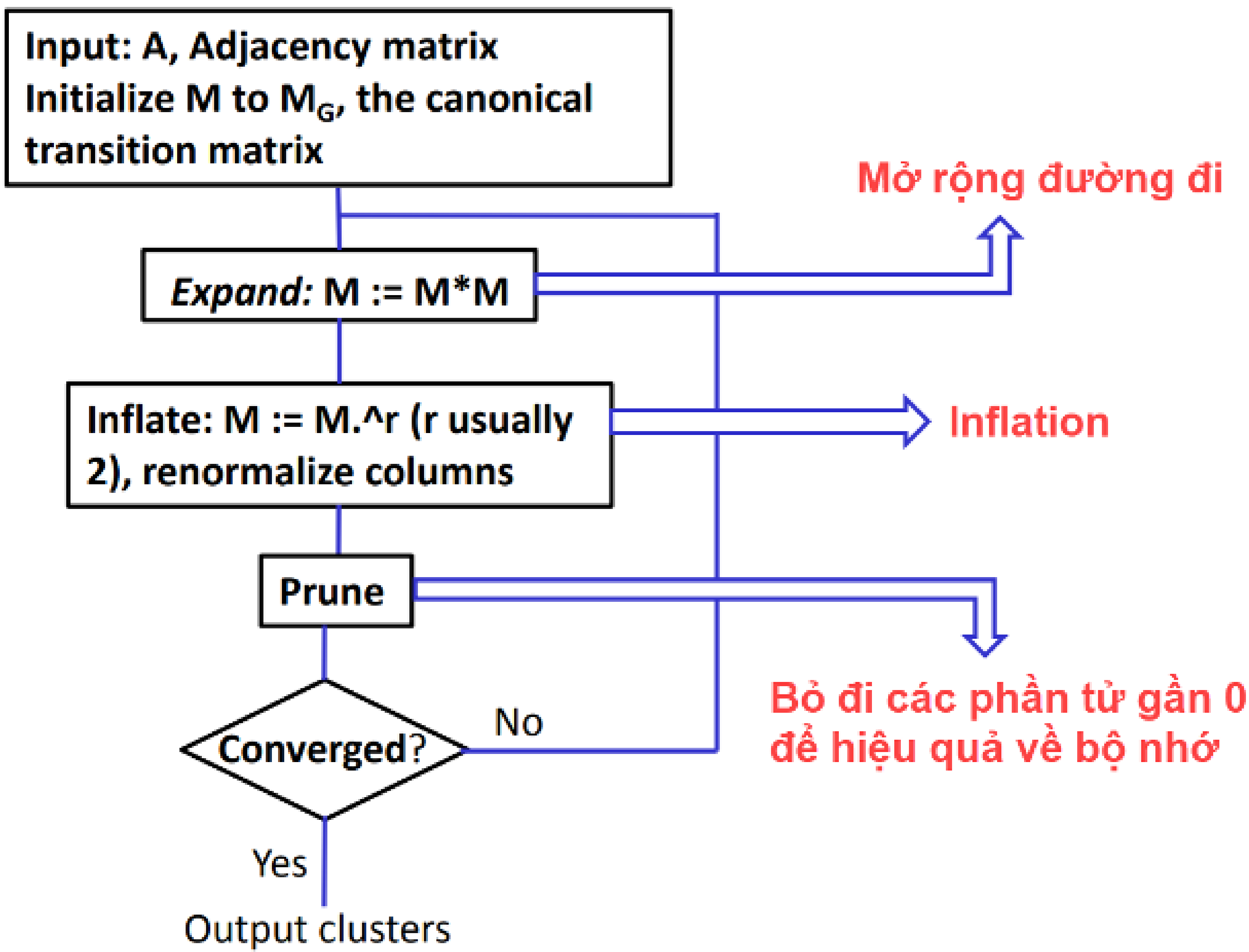
Inflationization



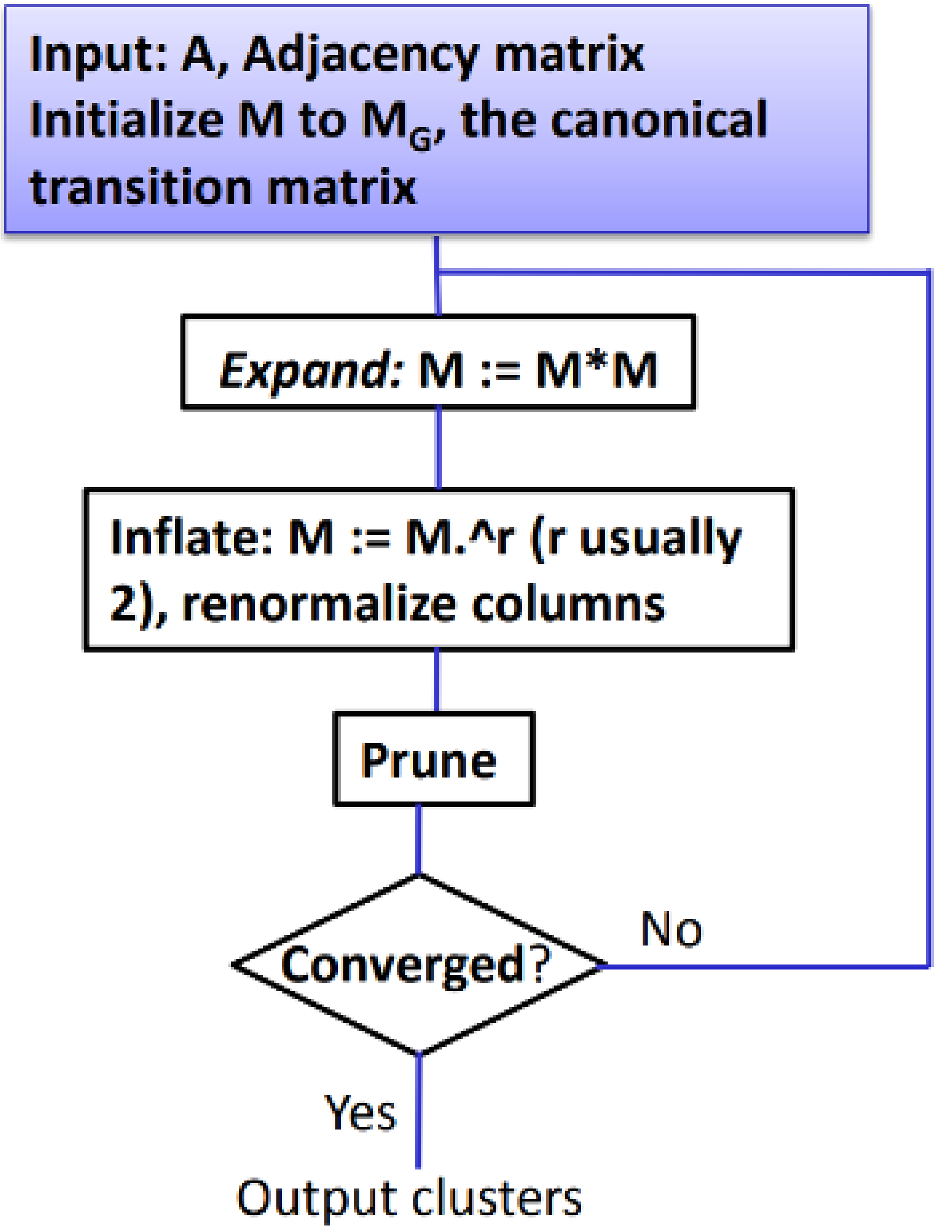


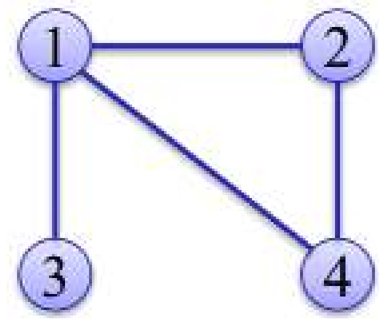
$$\begin{bmatrix}0\\1/2\\0\\1/6\\1/3\end{bmatrix} \xrightarrow{\text{Squaring}} \begin{bmatrix}0\\1/4\\0\\1/36\\1/9\end{bmatrix} \xrightarrow{\text{Normalization}} \begin{bmatrix}0\\9/14\\0\\1/14\\4/14\end{bmatrix}$$

Based on random walk



Based on random walk

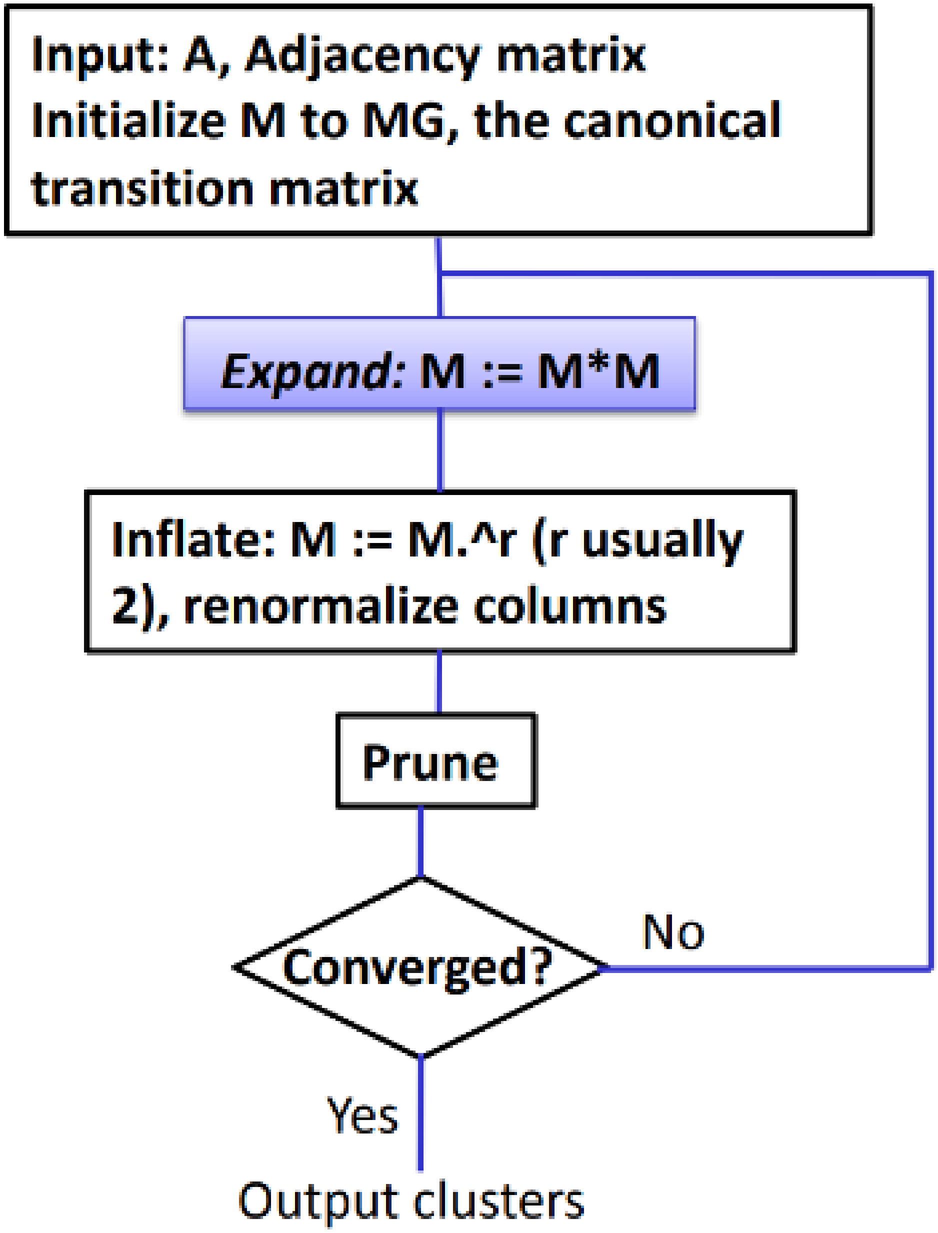


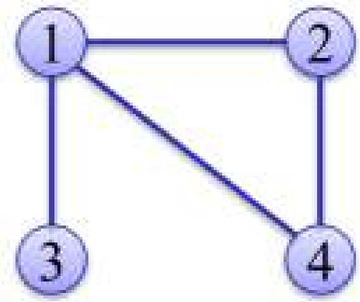


$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1/4 & 1/3 & 1/2 & 1/3 \\ 1/4 & 1/3 & 0 & 1/3 \\ 1/4 & 0 & 1/2 & 0 \\ 1/4 & 1/3 & 0 & 1/3 \end{bmatrix}$$

Based on random walk

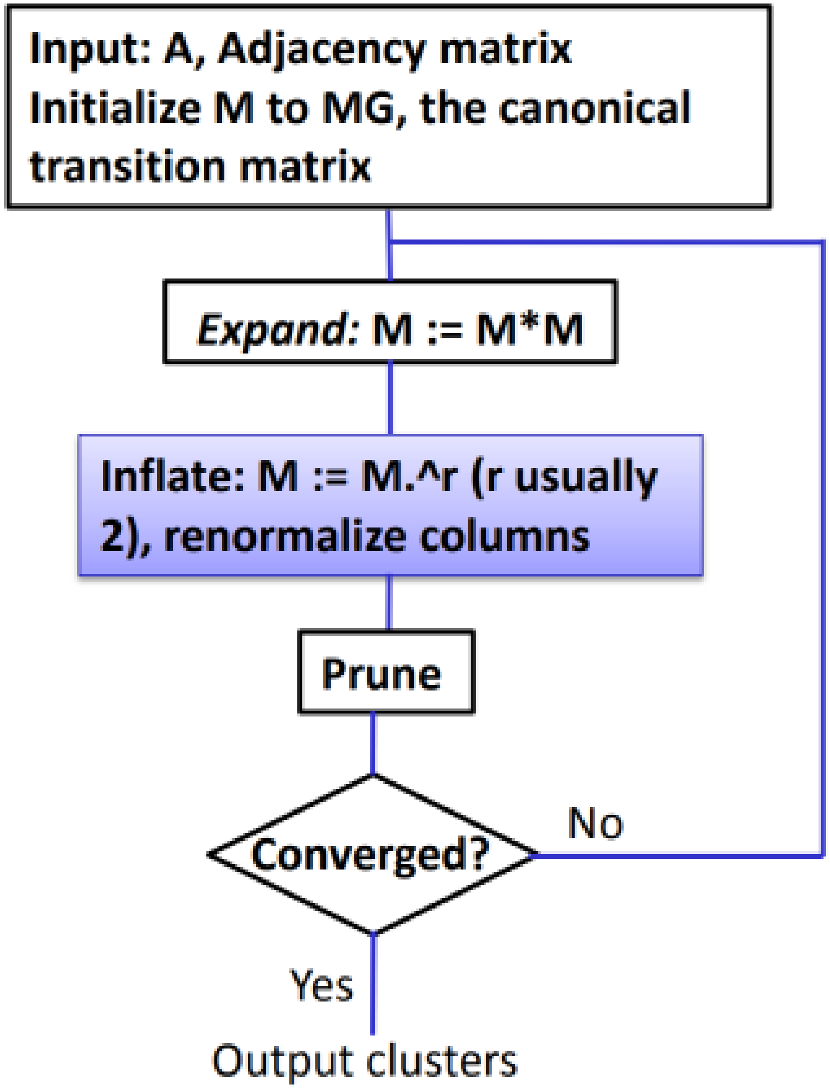


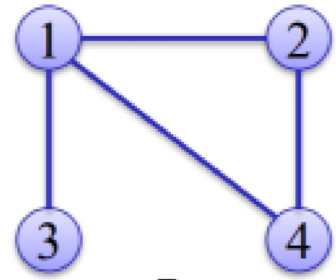


$$\begin{bmatrix} 1/4 & 1/3 & 1/2 & 1/3 \\ 1/4 & 1/3 & 0 & 1/3 \\ 1/4 & 0 & 1/2 & 0 \\ 1/4 & 1/3 & 0 & 1/3 \end{bmatrix} * \begin{bmatrix} 1/4 & 1/3 & 1/2 & 1/3 \\ 1/4 & 1/3 & 0 & 1/3 \\ 1/4 & 0 & 1/2 & 0 \\ 1/4 & 1/3 & 0 & 1/3 \end{bmatrix}$$

$$= \begin{bmatrix} 0.35 & 0.31 & 0.38 & 0.31 \\ 0.23 & 0.31 & 0.13 & 0.31 \\ 0.19 & 0.08 & 0.38 & 0.08 \\ 0.23 & 0.31 & 0.13 & 0.31 \end{bmatrix}$$

Based on random walk





$$\begin{bmatrix} 0.35 & 0.31 & 0.38 & 0.31 \\ 0.23 & 0.31 & 0.13 & 0.31 \\ 0.19 & 0.08 & 0.38 & 0.08 \\ 0.23 & 0.31 & 0.13 & 0.31 \end{bmatrix}$$

inflation

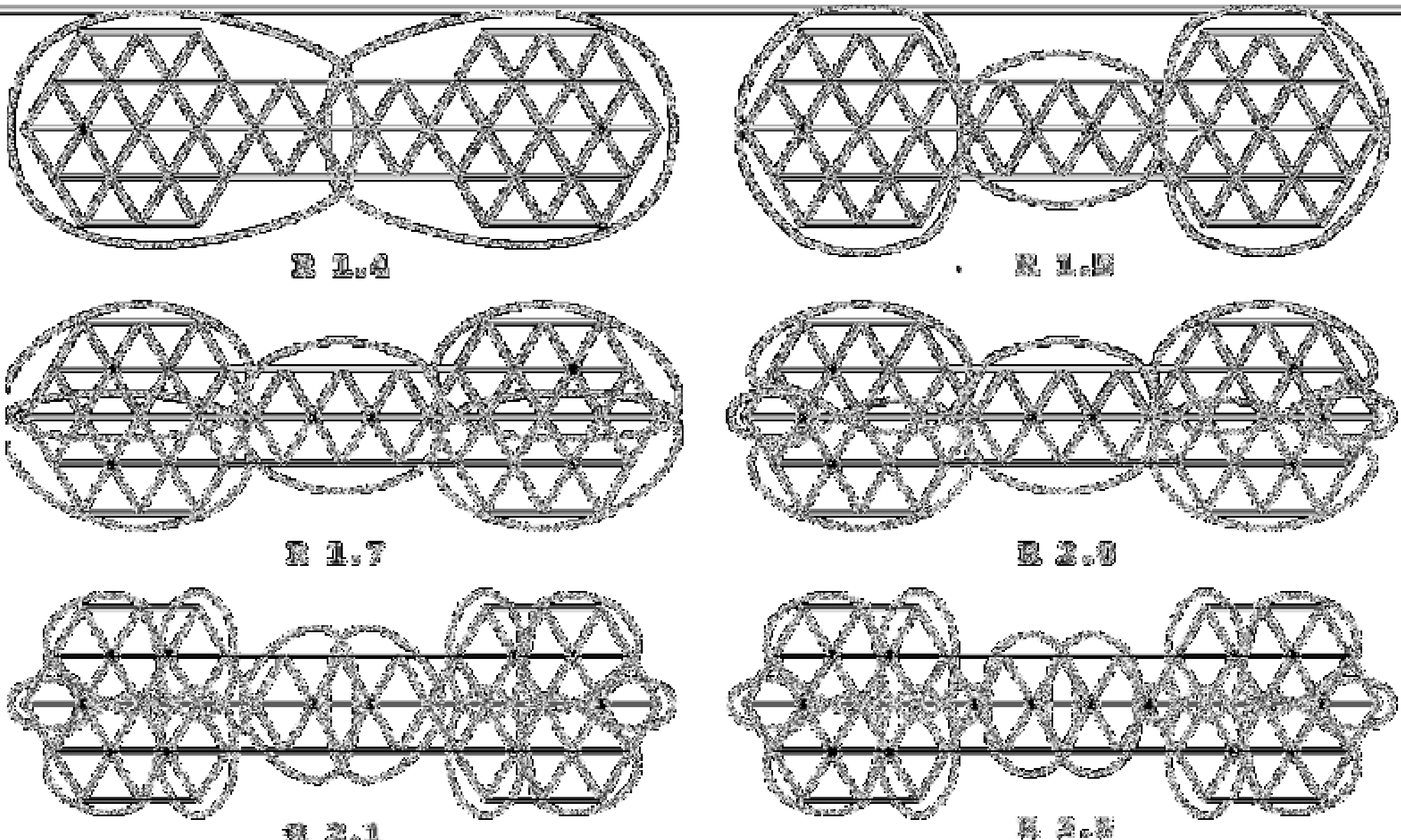
(Bình phương)

$$\begin{bmatrix} 0.13 & 0.09 & 0.14 & 0.09 \\ 0.05 & 0.09 & 0.02 & 0.09 \\ 0.04 & 0.01 & 0.14 & 0.01 \\ 0.05 & 0.09 & 0.02 & 0.09 \end{bmatrix}$$

normalization

$$\begin{bmatrix} 0.47 & 0.33 & 0.45 & 0.33 \\ 0.20 & 0.33 & 0.05 & 0.33 \\ 0.13 & 0.02 & 0.45 & 0.02 \\ 0.20 & 0.33 & 0.05 & 0.33 \end{bmatrix}$$

The Inflations



Reference

- Mihalcea, Rada, and Dragomir Radev. *Graph-based natural language processing and information retrieval*. Cambridge university press, 2011.