Khai Thác Dữ Liệu Đồ Thị

PHÁT HIỆN CỘNG ĐỒNG

Giảng viên: Lê Ngọc Thành

Email: lnthanh@fit.hcmus.edu.vn

fit@hcmus

Nội dung

- **Khái niệm**
- Tối ưu cộng đồng
- Các phương pháp phát hiện cộng đồng
  - Phương pháp dựa trên lát cắt tối thiểu
  - Phương pháp dựa trên trung gian
  - Phương pháp dựa trên RandomWalk

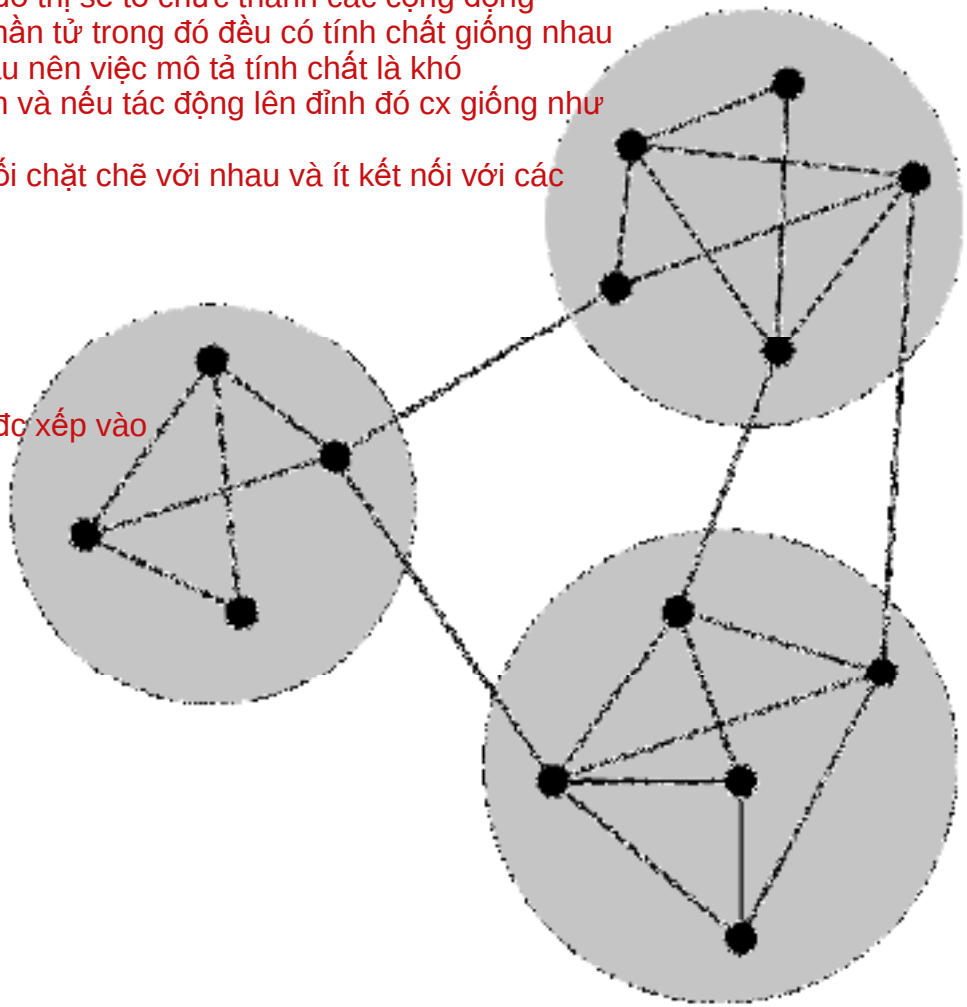
Cộng đồng

- **Cộng đồng** (community) là tập các đỉnh mà mỗi đỉnh có nhiều kết nối bên trong hơn kết nối ra ngoài.

- Các đồ thị trong thực tế ko phải là ngẫu nhiên
- Cta kỳ vọng rằng các đỉnh trong 1 đồ thị sẽ tổ chức thành các cộng động
- Phát hiện ra nhòm (CD) vì tất cả phần tử trong đó đều có tính chất giống nhau
- Mỗi nhóm sẽ có tính chất khác nhau nên việc mô tả tính chất là khó
- Gom các ptu trong CD thành 1 đỉnh và nếu tác động lên đỉnh đó cx giống như tác động trên nhóm đó
- Các đỉnh trong CD sẽ có mối kết nối chặt chẽ với nhau và ít kết nối với các cộng đồng khác

- Có 2 loại cộng đồng là:
+ Cộng dồng giao nhau (*)
+ Cộng đồng ko giao nhau

=> Bài toán gom nhóm (Clustering) đc xếp vào Unsupervise learning vì ko có nhãn. Nếu đã có nhãn thì nó thuộc về bài toán phân lớp (Classification)

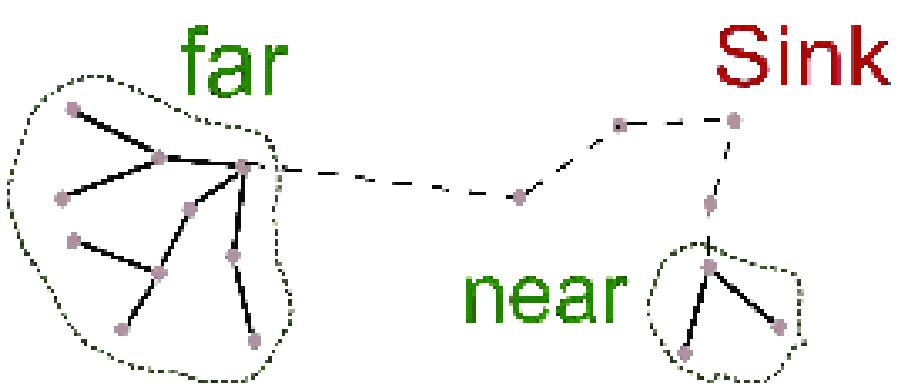


Nội dung

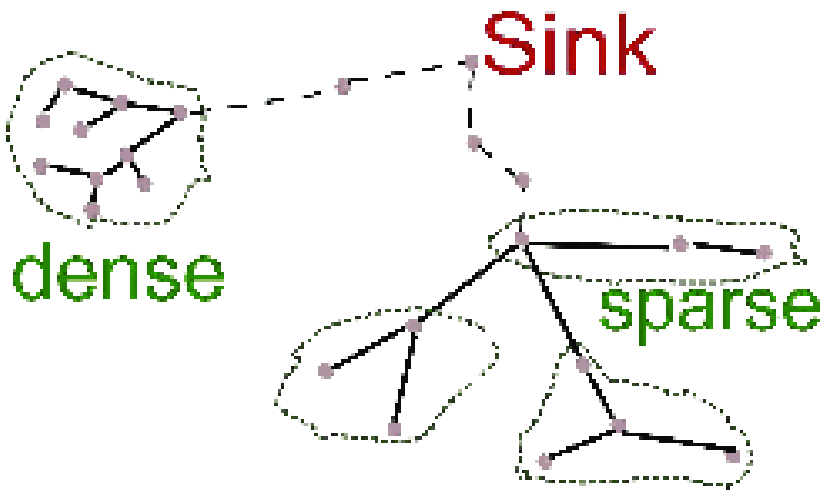
- Khái niệm
- **Tối ưu cộng đồng**
- Các ph□ơng pháp phát hiện cộng đồng
  - Ph□ơng pháp dựa trên lát cắt tối thiểu
  - Ph□ơng pháp dựa trên trung gian
  - Ph□ơng pháp dựa trên RandomWalk

Tối ưu cộng đồng

- Ngoài các độ đo chất l□ợng và khoảng cách, ng□ời ta có thêm 2 cách đo để đánh giá mật độ:
  - Mật độ trong nhóm (intracluster density): càng lớn càng tốt
  - Mật độ ngoài nhóm (intercluster density): càng nhỏ càng tốt



(a) distance



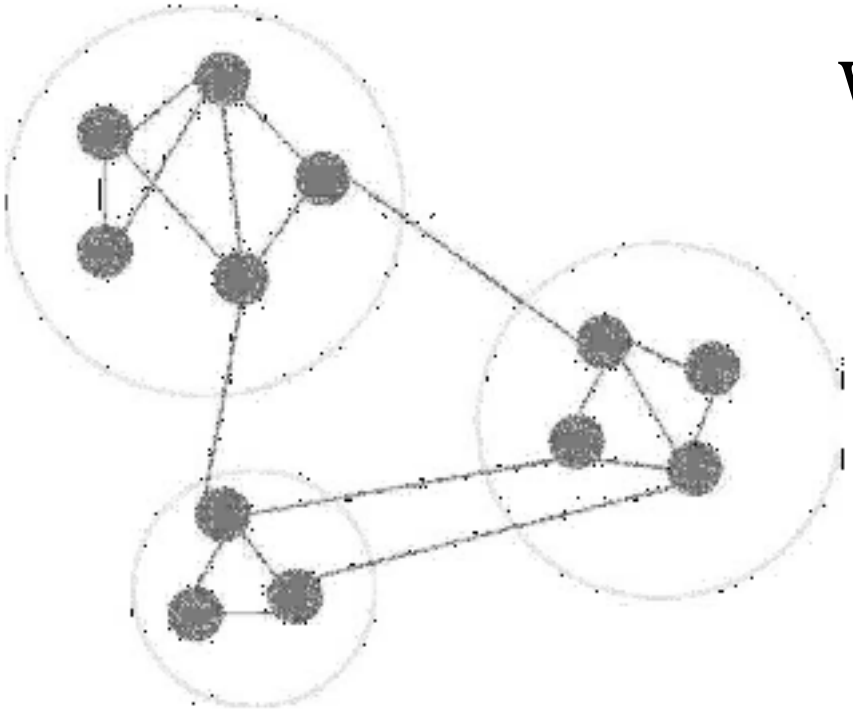
(b) density

Mật độ trong nhóm => THI

- Mật độ trong nhóm là tỉ số giữa số cạnh bên trong nhóm và số cạnh có thể tối đa trong nhóm.

Nc là số đỉnh trong nhóm

$$\delta(C) = \frac{số\ cạnh\ trong\ nhóm}{N_c(N_c - 1)/2}$$



Ví dụ:

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

$$\delta(C_2) = \frac{4}{6} = 0.75$$ => 5/6

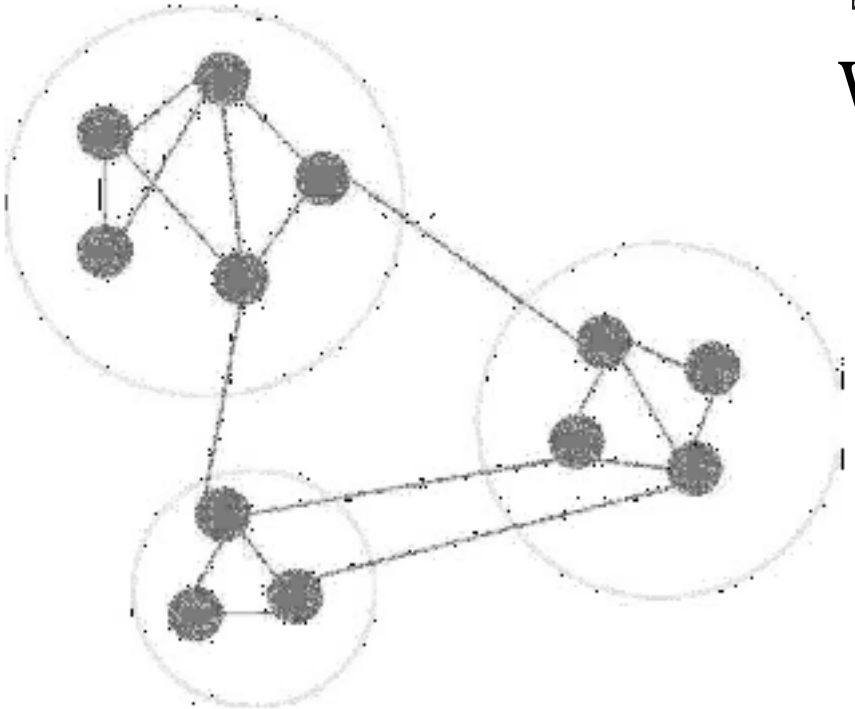
$$\delta(C_3) = \frac{3}{3} = 1.0$$

Mật độ ngoài nhóm

- Mật độ ngoài nhóm là tỉ số giữa số cạnh ngoài nhóm và số cạnh ngoài nhóm có thể.

Nc*(N-Nc): số đỉnh trog nhóm * số đỉnh ngoài nhóm

$$\epsilon(C) = \frac{số\ cạnh\ ngoài\ nhóm}{N_c(N - N_c)}$$



Ví dụ:

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

2/ (5*(3+4) = 2/35

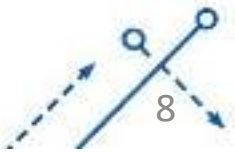
$$\epsilon(C_2) = \frac{3}{32}$$

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

=> Có nhiều cách gom nhóm khác nhau và để biết cách nào tốt thì họ dựa vào Modularity Maximum

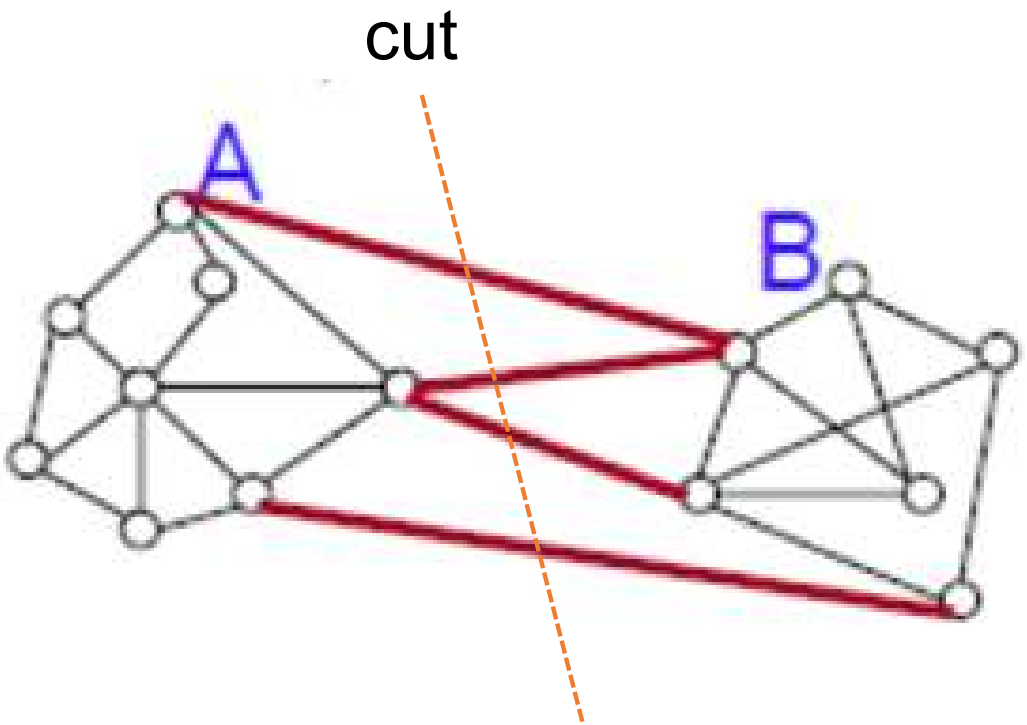
Nội dung

- Khái niệm
- Tối ưu cộng đồng
- **Các phương pháp phát hiện cộng đồng**
  - Phương pháp dựa trên lát cắt tối thiểu
  - Phương pháp dựa trên trung gian
  - Phương pháp dựa trên RandomWalk



Lát cắt tối thiểu

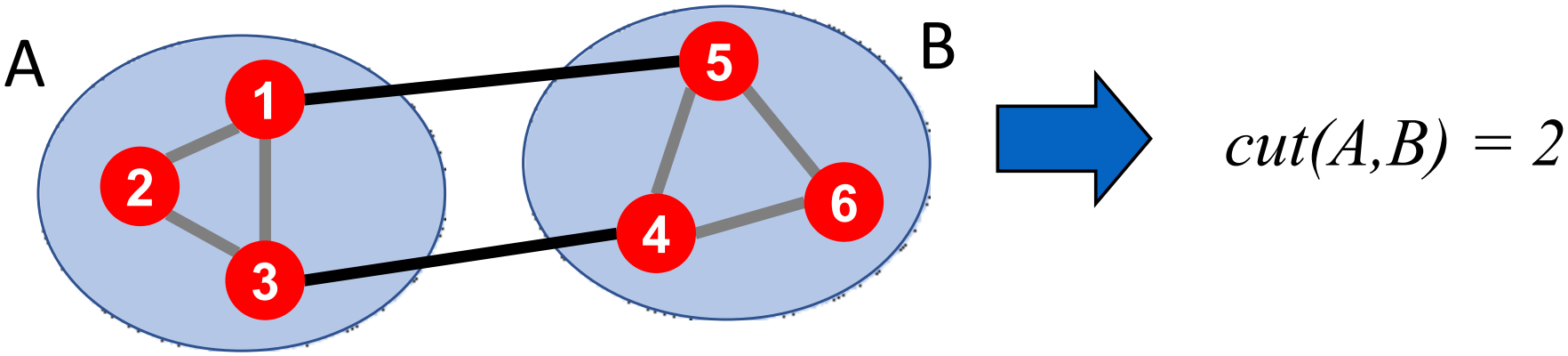
- Mục tiêu của lát cắt tối thiểu là tìm tập cạnh ít nhất mà chặn luồng từ nguồn S đến T.
  - Kích th□ớc cắt là tổng trọng l□ợng các cạnh đó



Lát cắt tối thiểu

- Lát cắt:

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

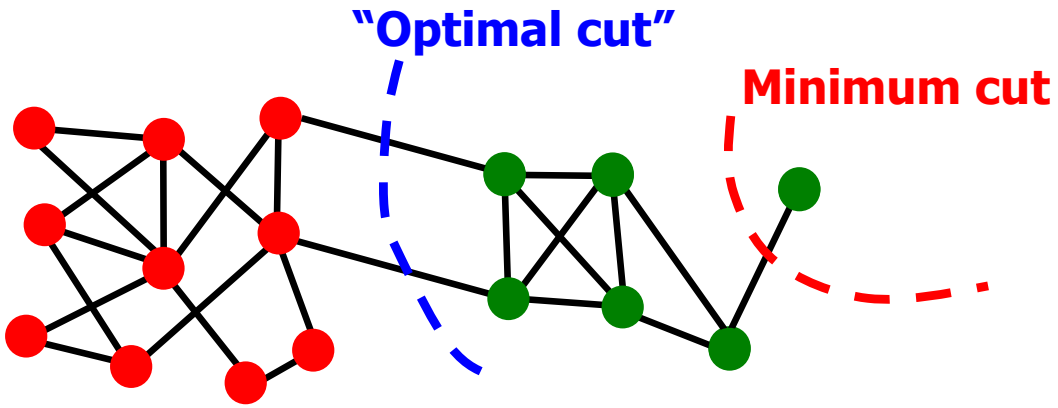


Phát hiện cộng đồng với lát cắt tối thiểu

- Mục tiêu: lát cắt tối thiểu

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

- Vấn đề:



Vấn đề xảy ra do không xem xét kết nối bên trong

Chuẩn hó□ lát cắt

- Lát cắt đ□ợc chuẩn hóa phụ thuộc vào mật độ trong từng nhóm

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

với $vol(A)$: tổng trọng số các cạnh với ít nhất một đầu ở trong A

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

Nội dung

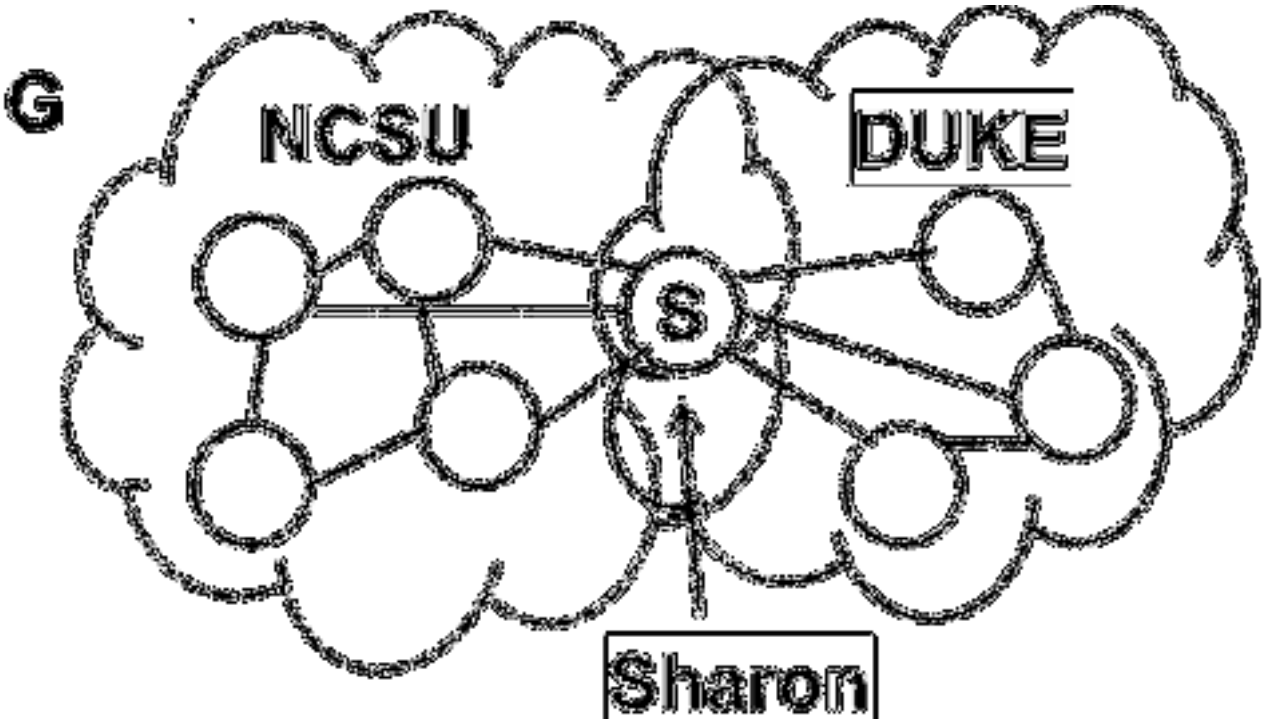
- Khái niệm
- Tối ưu cộng đồng
- **Các phương pháp phát hiện cộng đồng**
  - Phương pháp dựa trên lát cắt tối thiểu
  - Phương pháp dựa trên trung gian => Tập trung
  - Phương pháp dựa trên RandomWalk

Phát hiện cộng đồng dự□ trên tính trung gi□n

- Phát hiện cộng đồng dựa trên tính trung gian thực hiện quá trình xác định cộng đồng bằng cách lần l□ợt bỏ đi đỉnh/cạnh có tính trung gian cao.
- Có hai loại:
  - Dựa trên tính trung gian đỉnh
  - Dựa trên tính trung gian cạnh.

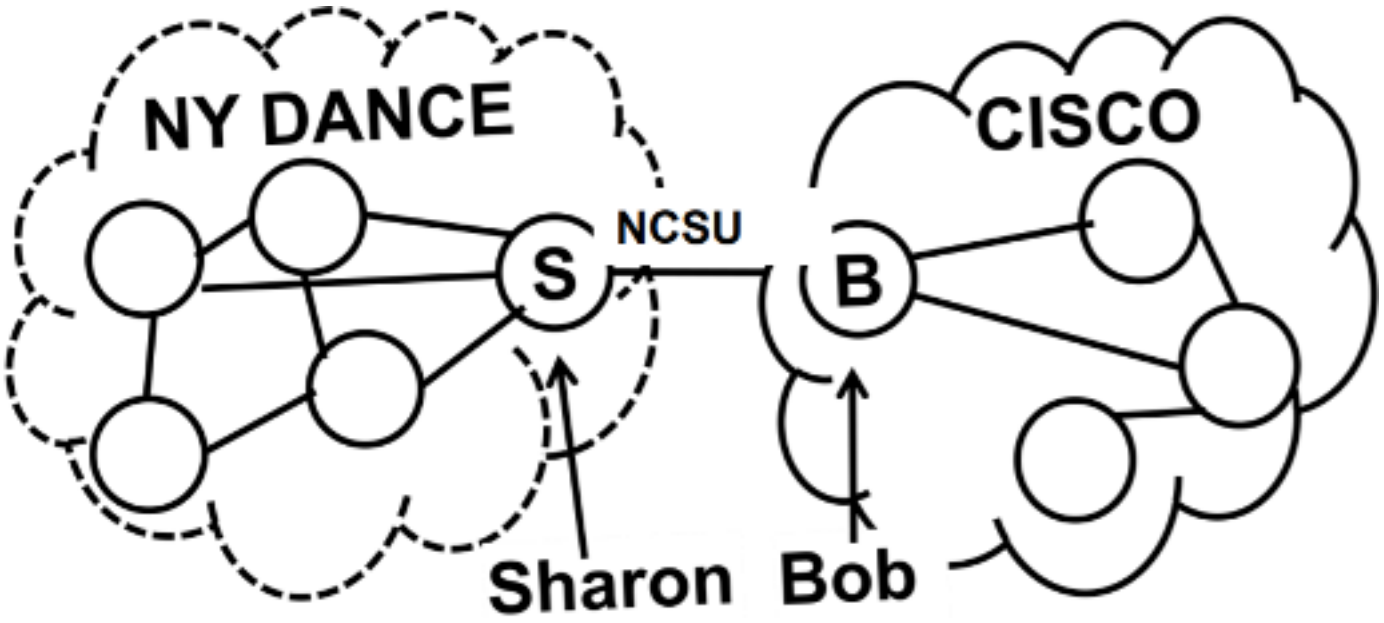
Trung gi□n đỉnh

- Trung gian đỉnh đ□ợc tính dựa trên số đ□ờng đi ngắn nhất trong đồ thị mà phải đi qua đỉnh cho tr□ớc.

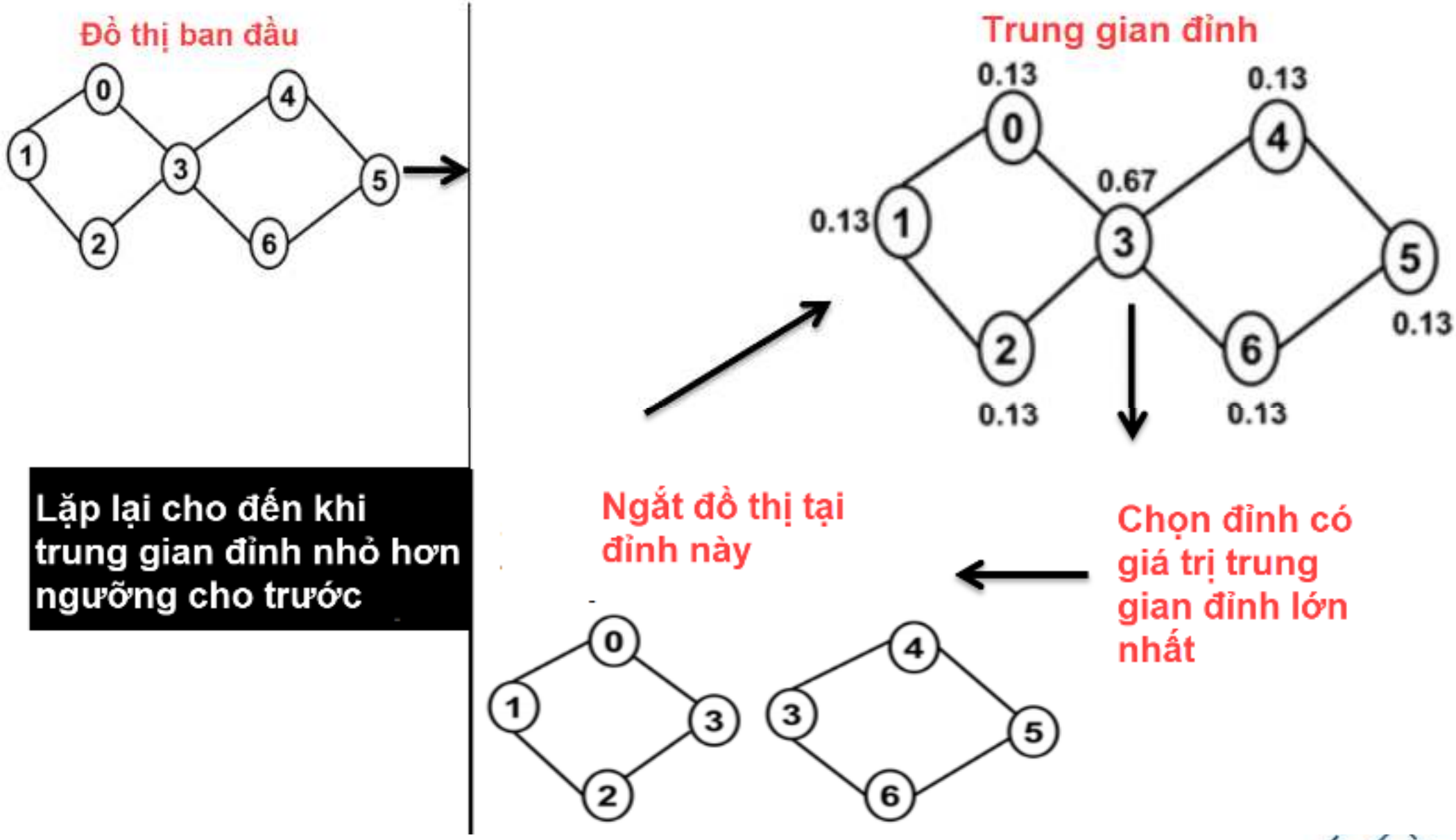


Trung gi□n cạnh

- Trung gian cạnh đ□ợc đánh giá dựa trên số đ□ờng đi ngắn nhất phải đi qua cạnh cho tr□ớc.



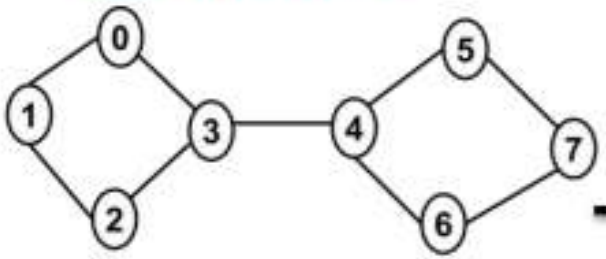
Thuật toán trung gi□n đỉnh



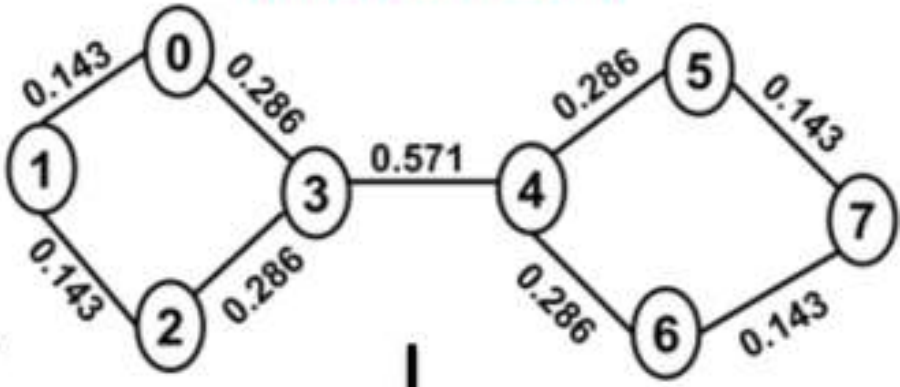
Thuật toán trung gi□n cạnh

=> Thuật toán Girvan Newman
=> Tập trung vào trung gian cạnh
=> Tập trung vào thuật toán Girvan Newman

Đồ thị ban đầu



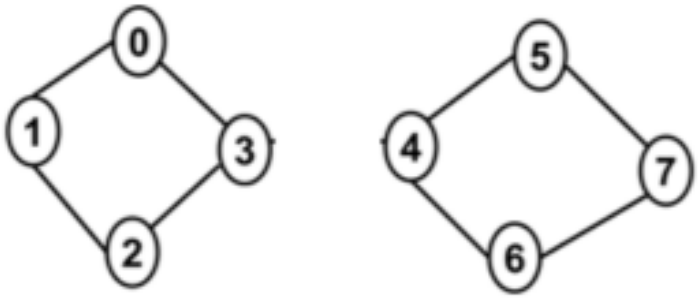
Trung gian cạnh



Chọn cạnh có giá trị trung gian đỉnh lớn nhất

Ngắt đồ thị tại cạnh này

Lặp lại cho đến khi độ cạnh trung gian nhỏ hơn ngưỡng cho trước



Cho đồ thị cho trước và ngưỡng modalarity và áp dụng thuật toán

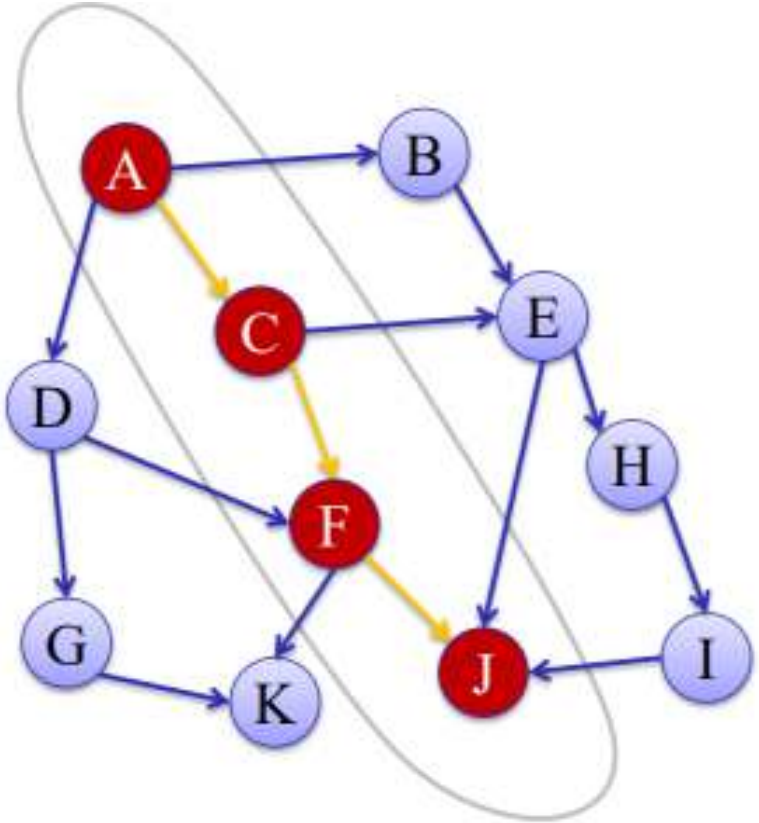
=> Hội tụ khi mà các nhóm chỉ có 1 đỉnh => Đg cắt đến từ tính Modalarity, đến 1 ngưỡng Modalarity thì cắt

Nội dung

- Khái niệm
- Tối ưu cộng đồng
- **Các phương pháp phát hiện cộng đồng**
  - Phương pháp dựa trên lát cắt tối thiểu
  - Phương pháp dựa trên trung gian
  - Phương pháp dựa trên RandomWalk

R□ndom W□lk

- Cho một đồ thị, từ một đỉnh bắt đầu, ta chọn ngẫu nhiên một đỉnh để đi tiếp. Sau t lần lặp lại nh□ vậy, ta sẽ có một đ□ờng đi ngẫu nhiên (random walk) kích th□ớc t.

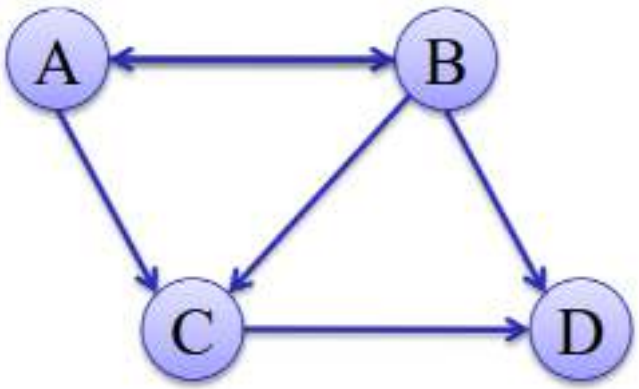


R□ndom W□lk

- Xác suất chọn cạnh có thể đánh giá dựa trên độ t□ơng quan giữa hai văn bản và đ□ợc chuẩn hóa.

Transition Matrix

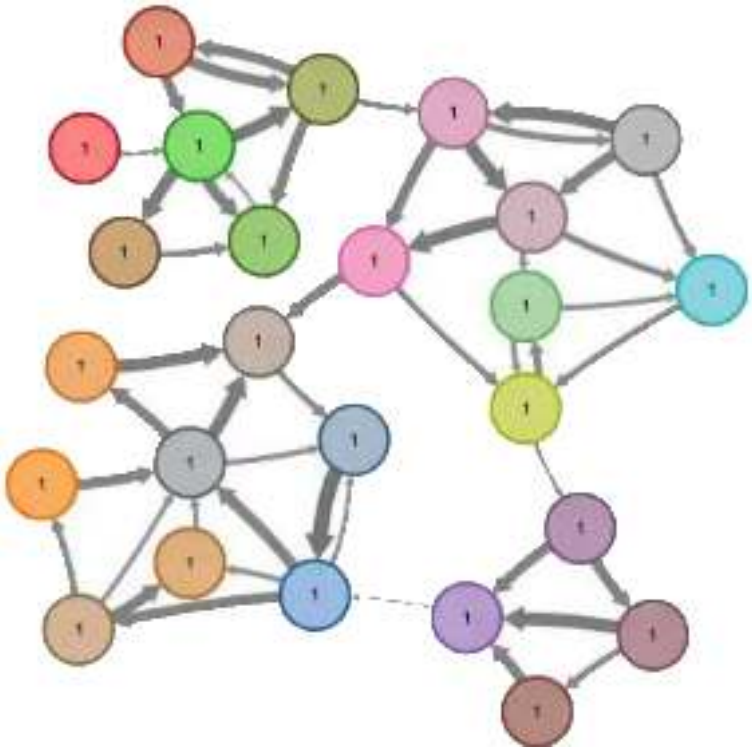
$$\begin{bmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$



Phát hiện cộng đồng dự□ trên r□ndom w□lk

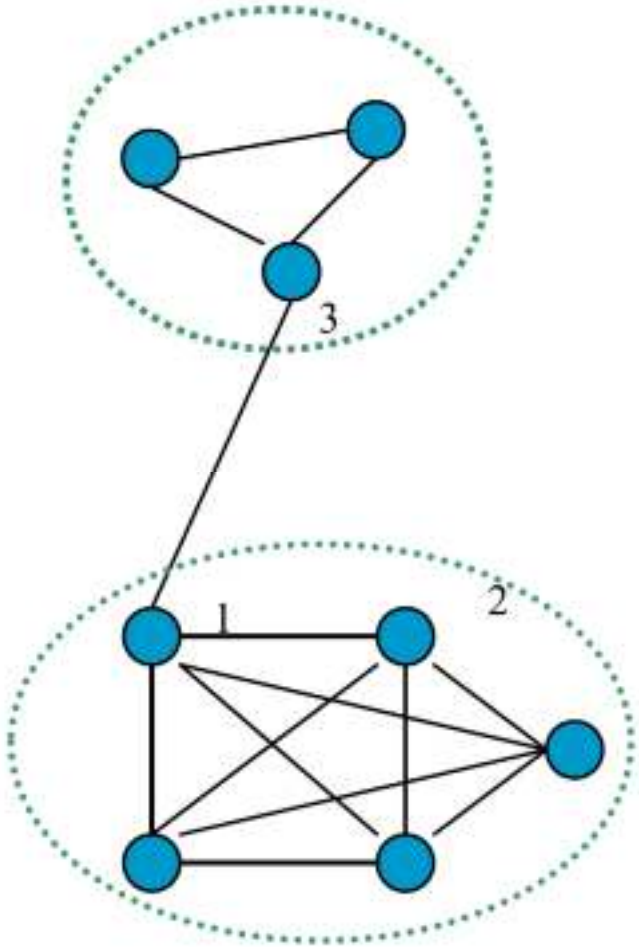
- Việc phát hiện cộng đồng dựa trên random walk đ?ợc thực hiện dựa trên nhận định:

  Một đường đi ngẫu nhiên bắt đầu ở một đỉnh nhiều khả năng sẽ di chuyển trong cộng đồng hơn là di chuyển trong cộng đồng khác.



Ví dụ

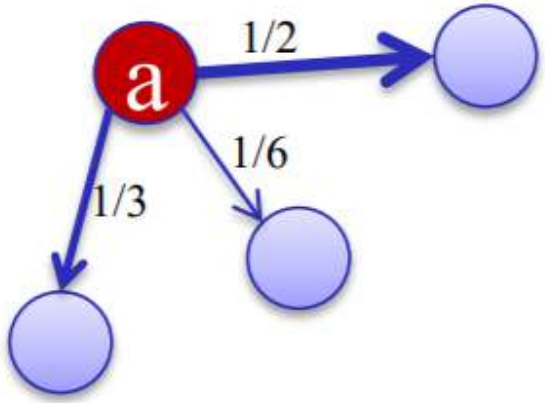
Node	Prob. Next Step within cluster	Prob. Next Step between clusters
1	80%	20%
2	100%	0%
3	67%	33%

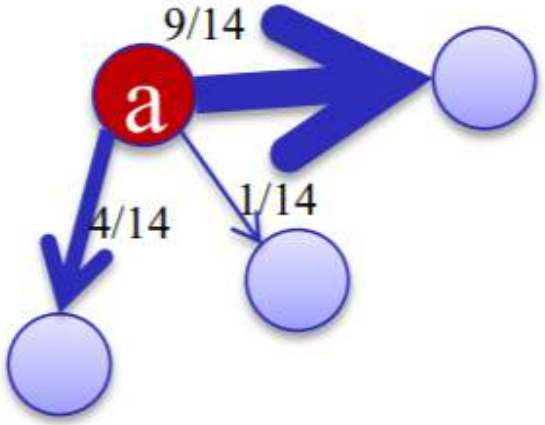


Ý tưởng thuật toán

- Điều chỉnh trọng số để sau b□ớc đi ngẫu nhiên kích th□ớc cho tr□ớc, khả năng di chuyển trong nhóm sẽ cao.
- Trọng số đ□ợc điều chỉnh để mà:
  - Các láng giềng mạnh sẽ càng mạnh lên
  - Các láng giềng yếu sẽ càng yếu đi
  - Tiến trình này đ□ợc gọi lại sự làm phát (inflation)

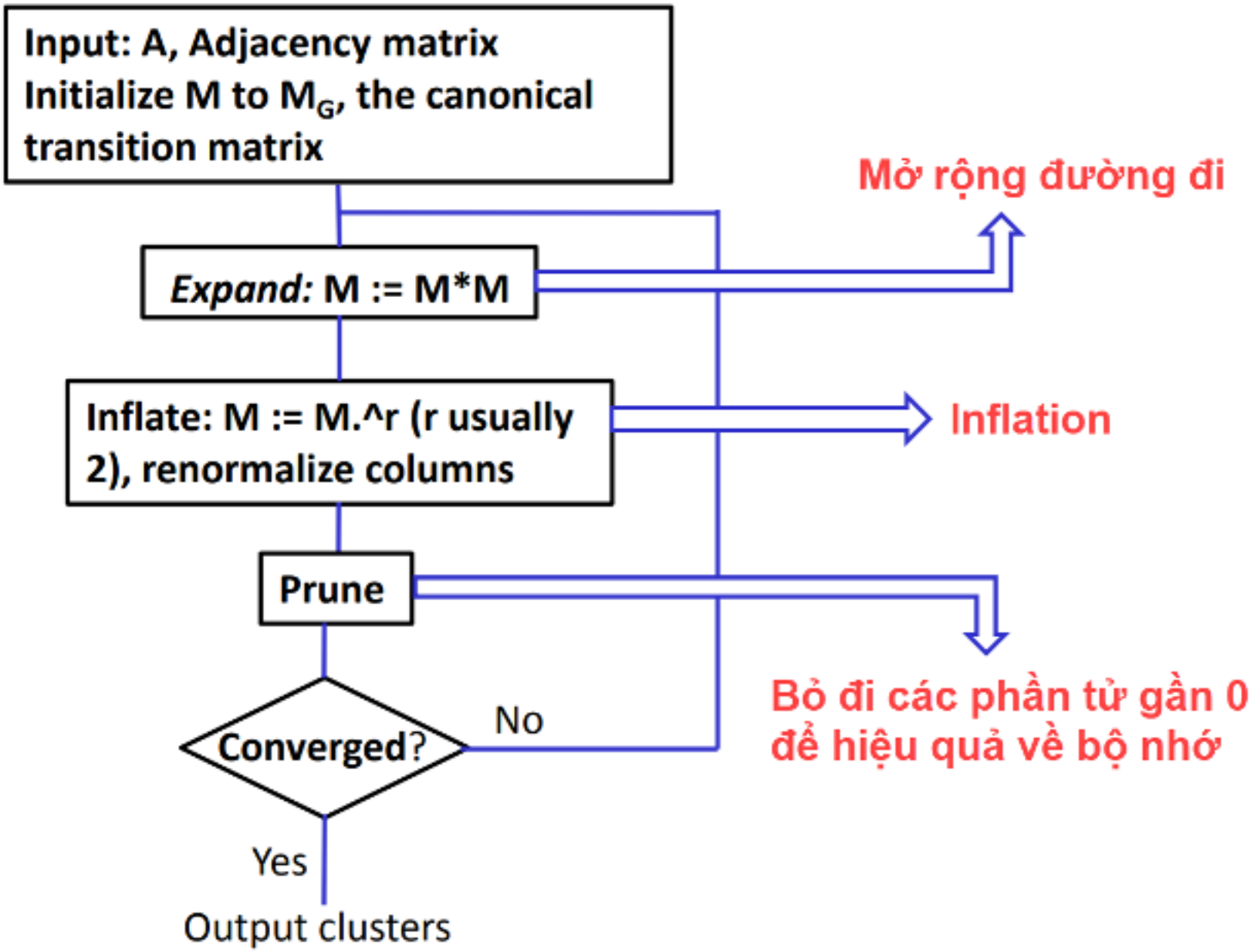
Lạm phát hó□



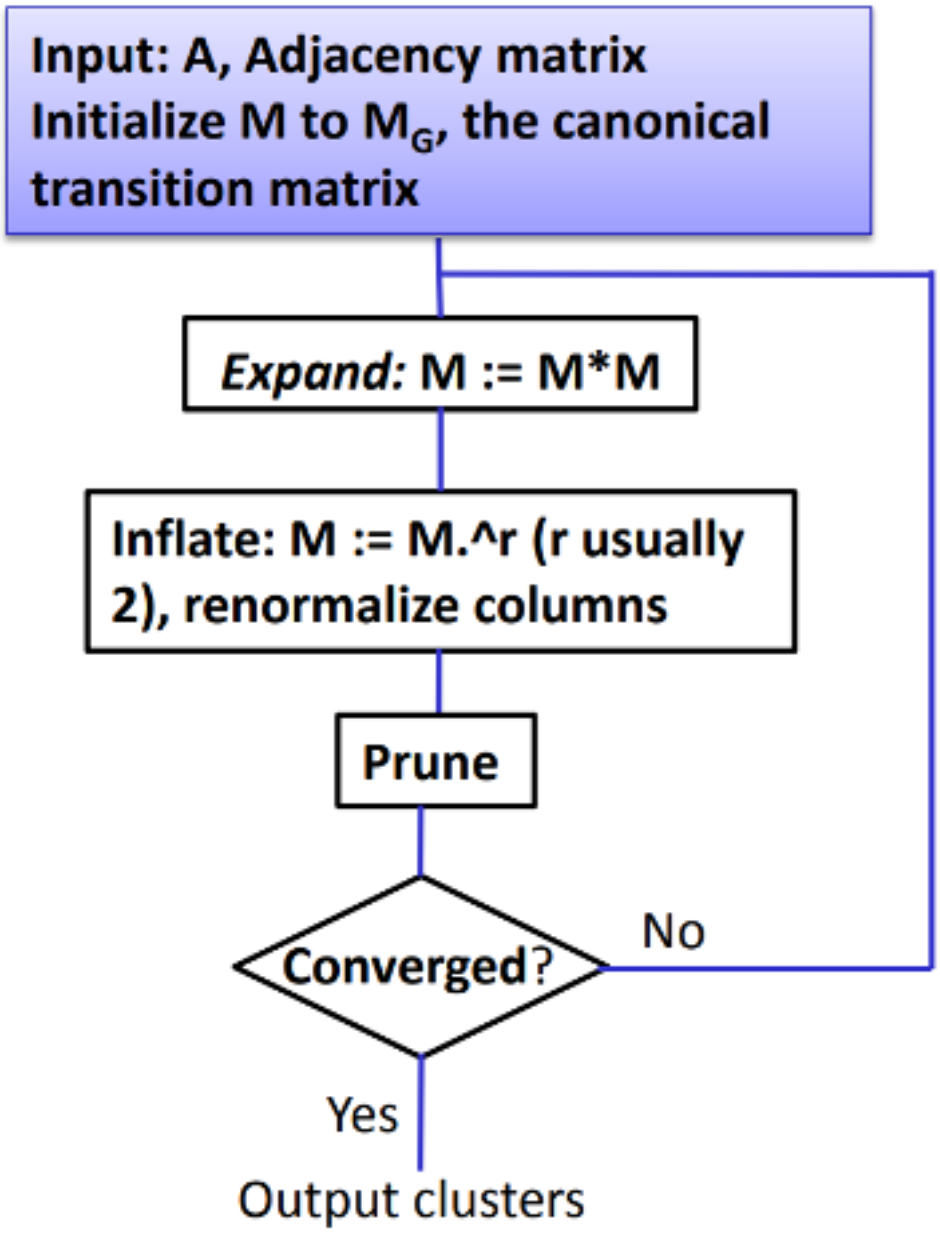


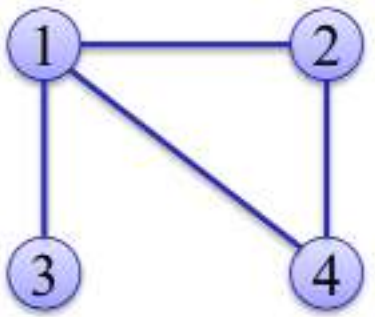
$$\begin{bmatrix} 0 \\ 1/2 \\ 0 \\ 1/6 \\ 1/3 \end{bmatrix} \xrightarrow{\text{Squaring}} \begin{bmatrix} 0 \\ 1/4 \\ 0 \\ 1/36 \\ 1/9 \end{bmatrix} \xrightarrow{\text{Normalization}} \begin{bmatrix} 0 \\ 9/14 \\ 0 \\ 1/14 \\ 4/14 \end{bmatrix}$$

Dự □ trên r □ ndom w □ lk



Dự□ trên r□ndom w□lk

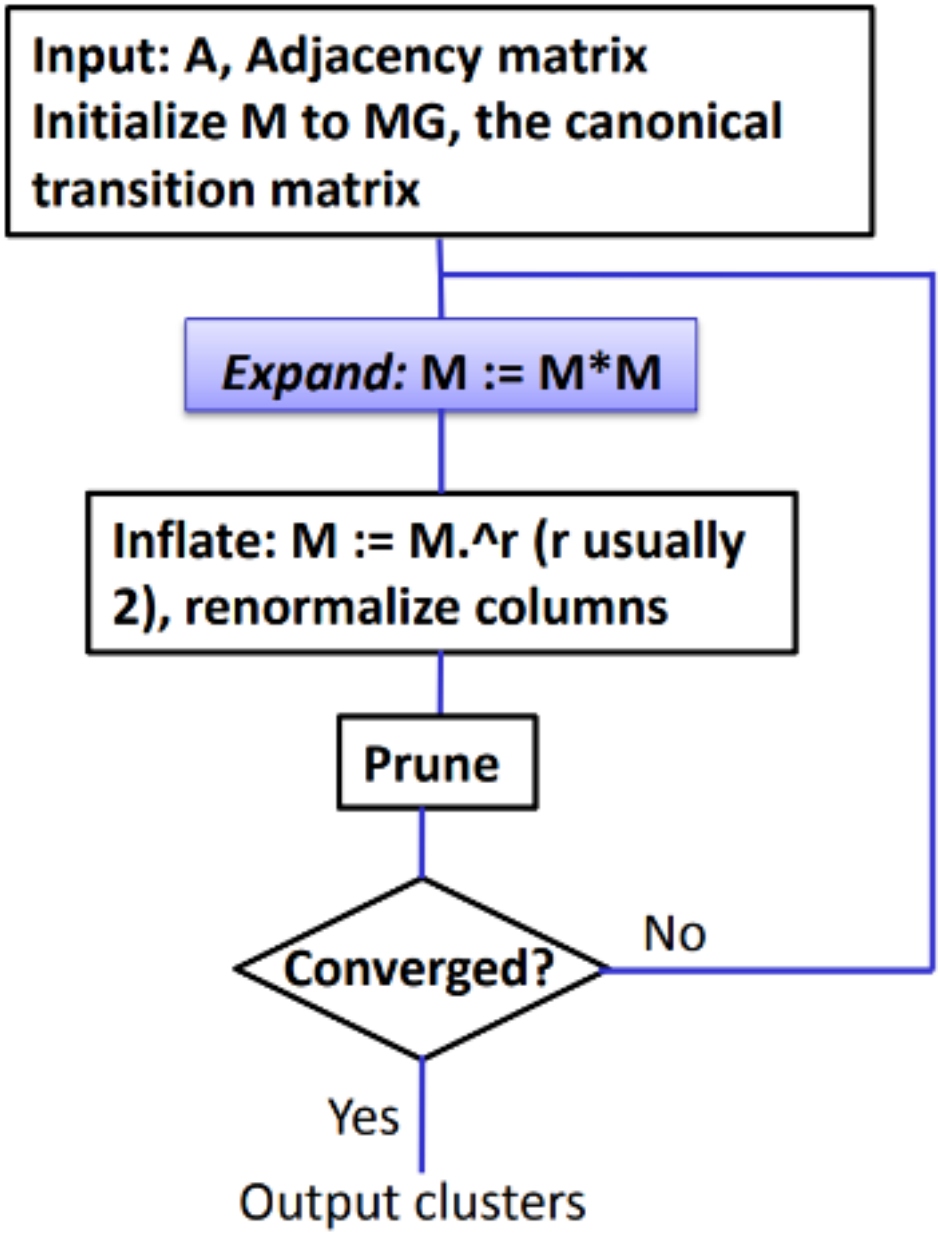


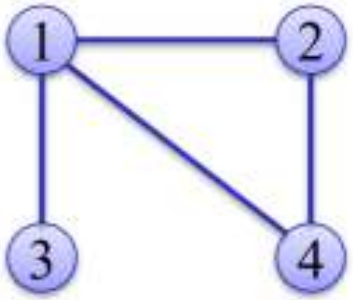


$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}$$

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

Dự□ trên r□ndom w□lk

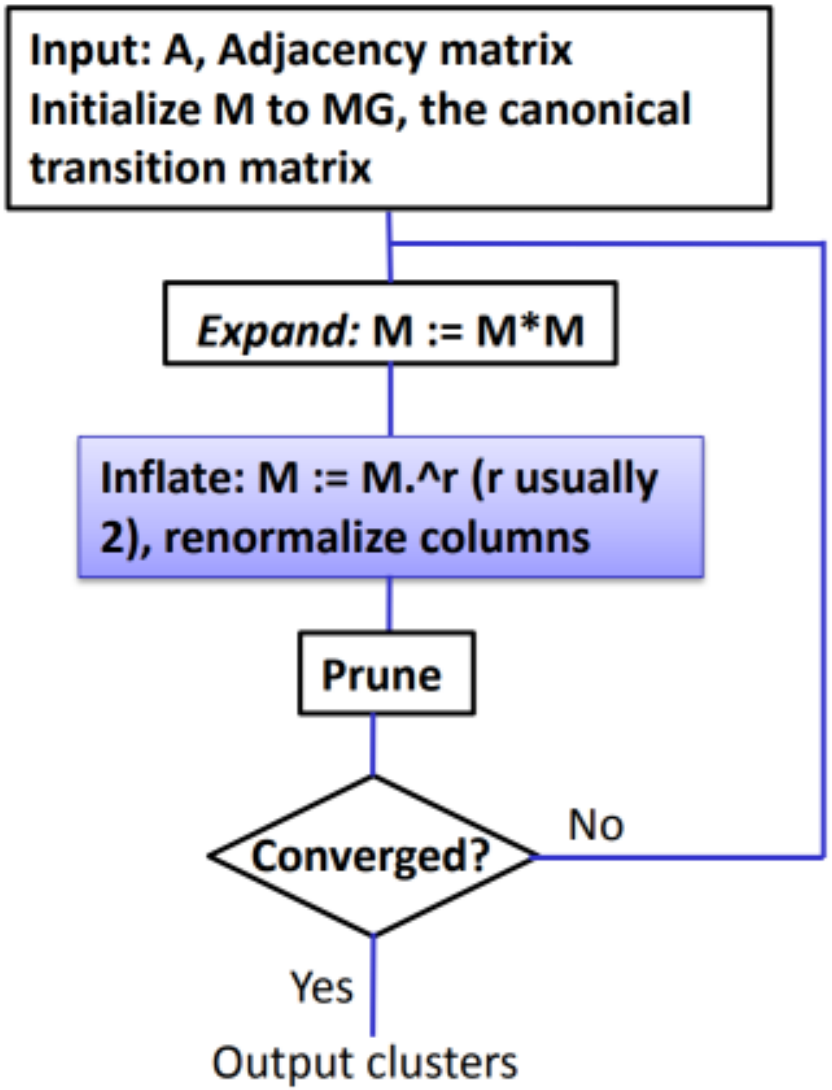


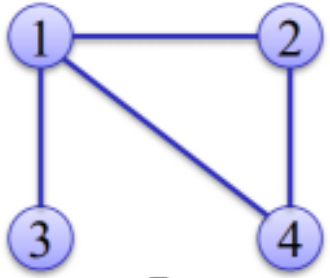


$$\begin{bmatrix} 1/4 & 1/3 & 1/2 & 1/3 \\ 1/4 & 1/3 & 0 & 1/3 \\ 1/4 & 0 & 1/2 & 0 \\ 1/4 & 1/3 & 0 & 1/3 \end{bmatrix} * \begin{bmatrix} 1/4 & 1/3 & 1/2 & 1/3 \\ 1/4 & 1/3 & 0 & 1/3 \\ 1/4 & 0 & 1/2 & 0 \\ 1/4 & 1/3 & 0 & 1/3 \end{bmatrix}$$

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

Dự□ trên r□ndom w□lk





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

inflation

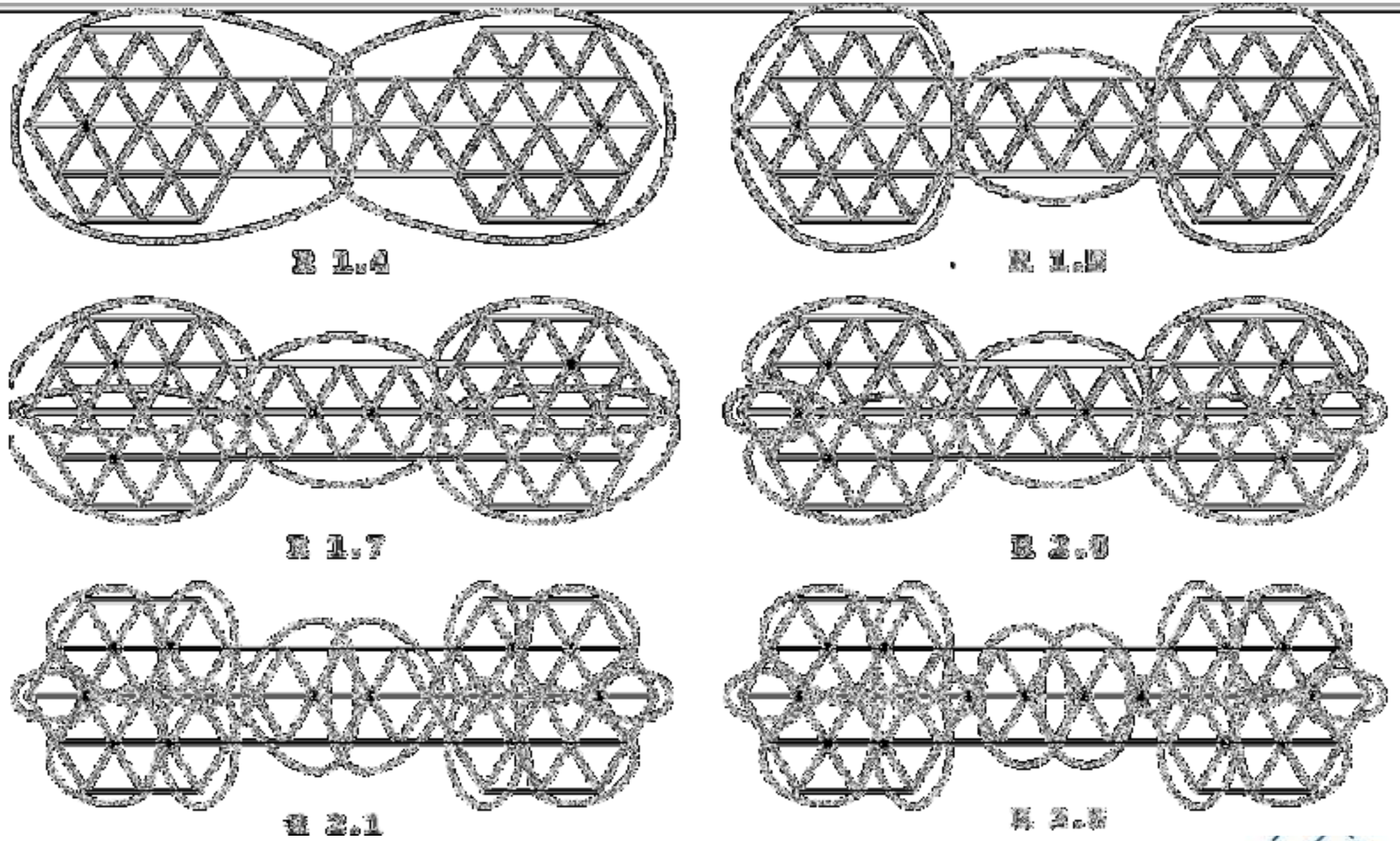
(Bình ph□ơng)

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

normalization

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

Các Infl□tion



Tài liệu th□m khảo

- Mih□lce□, R□d□, □nd Dr□gomir R□dev. *Graph-based natural language processing and information retrieval*. C□mbridge university press, 2011.