

MẠNG XÃ HỘI

Bài 3. CỘNG ĐỒNG XÃ HỘI

ThS. Lê Nhật Tùng

❶ 3.1. Khái niệm cộng đồng

❷ 3.2. Khám phá cộng đồng

1 3.1. Khái niệm cộng đồng

2 3.2. Khám phá cộng đồng

3.1. Khái niệm cộng đồng

Khái niệm cơ bản:

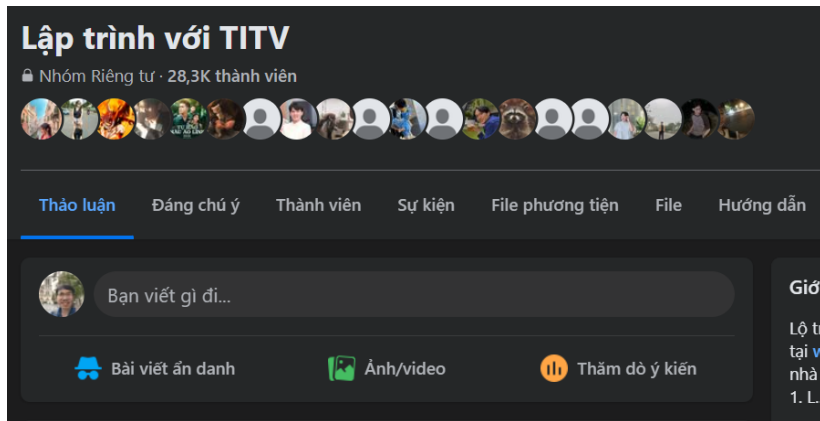
- Cộng đồng là tập hợp các cá nhân có mối liên kết chặt chẽ
- Các thành viên trong cộng đồng tương tác với nhau thường xuyên hơn so với bên ngoài
- Mỗi cộng đồng thường có đặc điểm và mục tiêu chung

3.1. Khái niệm cộng đồng

Ví dụ:

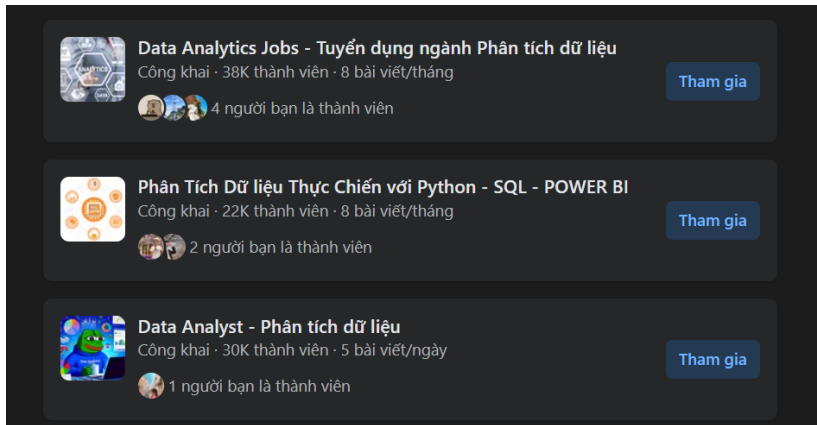
- Nhóm bạn học cùng lớp
- Câu lạc bộ thể thao
- Nhóm làm việc trong công ty

3.1. Khái niệm cộng đồng



Hình: Lập trình với TITV

3.1. Khái niệm cộng đồng



The screenshot displays three community listings on a dark background. Each listing includes a square icon, a title, a description of members and posts, and a 'Tham gia' (Join) button.

- Community 1:**
 - Icon:** A hexagonal graphic with the word 'ANALYTICS' in the center.
 - Title:** Data Analytics Jobs - Tuyển dụng ngành Phân tích dữ liệu
 - Description:** Công khai · 38K thành viên · 8 bài viết/tháng
 - Members:** 4 người bạn là thành viên (indicated by 4 small profile icons)
 - Button:** Tham gia
- Community 2:**
 - Icon:** A circular graphic with orange and white segments.
 - Title:** Phân Tích Dữ liệu Thực Chiến với Python - SQL - POWER BI
 - Description:** Công khai · 22K thành viên · 8 bài viết/tháng
 - Members:** 2 người bạn là thành viên (indicated by 2 small profile icons)
 - Button:** Tham gia
- Community 3:**
 - Icon:** A colorful graphic featuring a blue dinosaur-like character.
 - Title:** Data Analyst - Phân tích dữ liệu
 - Description:** Công khai · 30K thành viên · 5 bài viết/ngày
 - Members:** 1 người bạn là thành viên (indicated by 1 small profile icon)
 - Button:** Tham gia

Hình: Ví dụ một số cộng đồng

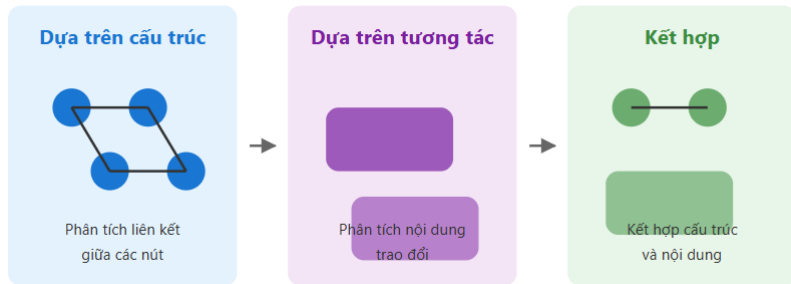
3.1. Khái niệm cộng đồng

Khám phá các cộng đồng trên mạng xã hội:

- Khám phá cộng đồng là tìm các nhóm trong mạng xã hội với hàm thành viên của nhóm không được xác định trước
- Chúng ta cần khám phá các cộng đồng trên mạng xã hội vì:
 - Con người có tương tác trong xã hội
 - Mạng xã hội cho phép con người mở rộng đời sống xã hội theo nhiều cách khác nhau
 - Trong thế giới thực, việc tìm và gặp bạn bè nhằm tìm bạn có cùng sở thích khó hơn tìm và trao đổi trong mạng xã hội

3.1. Khái niệm cộng đồng

Các phương pháp khám phá cộng đồng:



- **Dựa trên cấu trúc:** Phân tích topology mạng
- **Dựa trên tương tác:** Phân tích nội dung trao đổi
- **Kết hợp:** Tận dụng cả hai nguồn thông tin

1 3.1. Khái niệm cộng đồng

2 3.2. Khám phá cộng đồng

3.2. Khám phá cộng đồng

- Một trong những bài toán thường gặp khi phân tích một mạng xã hội là việc phát hiện ra các cộng đồng theo một số tính chất nào đó trên mạng xã hội.

Các hướng tiếp cận khám phá cộng đồng

- **Node-Centric Community:**

- Mỗi node trong cộng đồng sẽ thỏa một số tính nào đó

- **Group-Centric Community:**

- Xem tất cả các liên kết trong một nhóm là một liên kết duy nhất
- Trong cộng đồng thì mỗi nhóm sẽ phải thỏa một số tính chất nào đó
- Ta không xem xét từng node mà xem một nhóm các node

- **Network-Centric Community:**

- Chia mạng xã hội thành những tập con các node không liên thông nhau

- **Hierarchy-Centric Community:**

- Xây dựng một cấu trúc phân cấp từ mạng xã hội

Các thuật toán khám phá cộng đồng

Các thuật toán chính

- **Thuật toán Girvan – Newman**
 - Phát hiện cấu trúc dựa trên độ tương tự của đỉnh
- **Thuật toán dựa trên độ đo tương tự**
 - Phát hiện cấu trúc cộng đồng thông qua đánh giá mức độ tương tự giữa các node
- **Thuật toán lan truyền nhãn**
 - Xác định cộng đồng thông qua quá trình lan truyền thông tin giữa các node

Chú ý

Mỗi thuật toán có ưu và nhược điểm riêng, việc lựa chọn phụ thuộc vào đặc điểm của bài toán cụ thể.

3.2.1 Modularity

Định nghĩa

Số đo modularity (Q), được đề xuất bởi Girvan và Newman dùng làm thước đo khám phá cộng đồng.

- Thuật toán này theo hướng Node-Centric Community
- Sử dụng để đánh giá chất lượng của việc phân chia cộng đồng
- Giá trị Modularity càng lớn, cộng đồng có cấu trúc càng tốt

Sức mạnh của cộng đồng

Khái niệm

Sức mạnh của cộng đồng được đo bằng sự chênh lệch giữa số lượng kết nối thực tế và số lượng kết nối kỳ vọng trong cộng đồng đó.

Công thức

$$\sum_{i \in CI, j \in CI} (A_{ij} - \frac{d_i d_j}{2m})$$

CI Tập các node thuộc cùng một cộng đồng

A_{ij} Ma trận kề biểu diễn kết nối thực tế:

- = 1 nếu có kết nối giữa i và j
- = 0 nếu không có kết nối

d_i, d_j Bậc của đỉnh i và j (số lượng kết nối)

m Số cạnh của đồ thị ; $[\frac{d_i d_j}{2m}]$ Số kết nối kỳ vọng giữa i và j

Modularity toàn mạng

Khái niệm

Modularity (Q) đo lường chất lượng của việc phân chia mạng thành các cộng đồng, bằng cách tổng hợp sức mạnh của tất cả các cộng đồng trong mạng.

Công thức

$$Q = \frac{1}{2m} \sum_{l=1}^k \sum_{i \in C_l, j \in C_l} (A_{ij} - \frac{d_i d_j}{2m})$$

k : Số lượng cộng đồng được phát hiện trong mạng

l : Chỉ số của cộng đồng, $l = 1, 2, \dots, k$

C_l : Tập hợp các node thuộc cộng đồng thứ l

A_{ij} : Ma trận kề (1: có kết nối, 0: không có kết nối)

d_i, d_j : Bậc của node i và node j

m : Tổng số cạnh trong mạng

Ý nghĩa giá trị Modularity

Phạm vi giá trị

Q là thước đo định lượng cho chất lượng phân chia cộng đồng, có giá trị nằm trong khoảng $[-1, 1]$

Đánh giá cấu trúc

- $Q > 0$: Cấu trúc cộng đồng tốt
- $Q = 0$: Cấu trúc ngẫu nhiên
- $Q < 0$: Cấu trúc cộng đồng kém

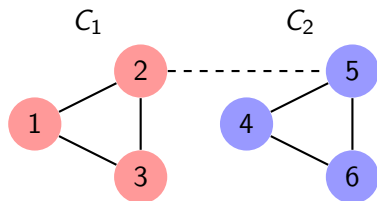
Trong thực tế

- $Q > 0.3$: Cấu trúc có ý nghĩa
- $Q > 0.7$: Cấu trúc rất rõ ràng
- $Q < 0.3$: Cấu trúc yếu

Ứng dụng

- Đánh giá chất lượng các thuật toán phát hiện cộng đồng
- So sánh các cách phân chia cộng đồng khác nhau
- Xác định số lượng cộng đồng tối ưu

Ví dụ tính toán Modularity



Thông số đồ thị:

- $m = 7$ cạnh
- $k = 2$ cộng đồng
- Bậc các đỉnh:
 - $d_1 = 2$
 - $d_2 = 3$
 - $d_3 = 2$
 - $d_4 = 2$
 - $d_5 = 3$
 - $d_6 = 2$

Tính toán Modularity - Bước 1

Tính toán cho cộng đồng C_1

$$\sum_{i,j \in C_1} (A_{ij} - \frac{d_i d_j}{2m})$$

- ❶ Các cặp cạnh trong C_1 : (1,2), (1,3), (2,3)
- ❷ $A_{ij} = 1$ cho các cặp này
- ❸ Kỳ vọng cho từng cặp:
 - (1,2): $\frac{2 \cdot 3}{14} = \frac{6}{14}$
 - (1,3): $\frac{2 \cdot 2}{14} = \frac{4}{14}$
 - (2,3): $\frac{3 \cdot 2}{14} = \frac{6}{14}$
- ❹ Tổng: $3 - (\frac{6}{14} + \frac{4}{14} + \frac{6}{14}) = 3 - \frac{16}{14}$

Tính toán Modularity - Bước 2

Tính toán cho cộng đồng C_2

$$\sum_{i,j \in C_2} (A_{ij} - \frac{d_i d_j}{2m})$$

- ❶ Các cặp cạnh trong C_2 : (4,5), (4,6), (5,6)
- ❷ $A_{ij} = 1$ cho các cặp này
- ❸ Kỳ vọng cho từng cặp:
 - (4,5): $\frac{2 \cdot 3}{14} = \frac{6}{14}$
 - (4,6): $\frac{2 \cdot 2}{14} = \frac{4}{14}$
 - (5,6): $\frac{3 \cdot 2}{14} = \frac{6}{14}$
- ❹ Tổng: $3 - (\frac{6}{14} + \frac{4}{14} + \frac{6}{14}) = 3 - \frac{16}{14}$

Tính toán Modularity - Kết quả

Công thức tổng quát

$$Q = \frac{1}{2m} [\text{Tổng } C_1 + \text{Tổng } C_2]$$

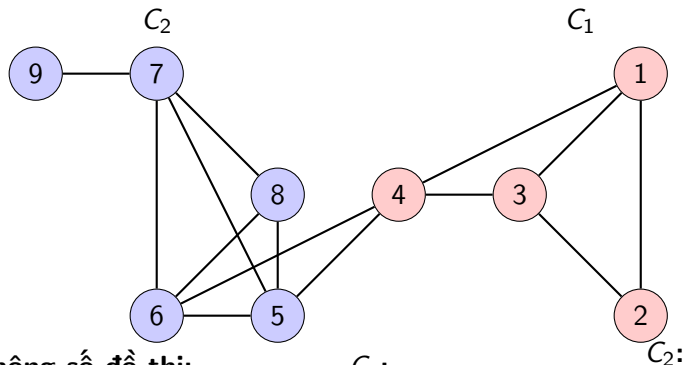
Thay số

$$\begin{aligned} Q &= \frac{1}{14} \left[\left(3 - \frac{16}{14} \right) + \left(3 - \frac{16}{14} \right) \right] \\ &= \frac{1}{14} \left[6 - \frac{32}{14} \right] \approx 0.265 \end{aligned}$$

Nhận xét

- $Q < 0.3$: Cấu trúc cộng đồng yếu

Ví dụ phức tạp - Tính Modularity



Thông số đồ thị:

- $m = 14$ cạnh
- $k = 2$ cộng đồng
- $C_1: \{1, 2, 3, 4\}$
- $C_2: \{5, 6, 7, 8, 9\}$

C_1 :

- $d_1 = 3$
- $d_2 = 2$
- $d_3 = 3$
- $d_4 = 4$

C_2 :

- $d_5 = 4$
- $d_6 = 4$
- $d_7 = 4$
- $d_8 = 3$
- $d_9 = 1$

Tính toán Modularity cho C_1

Công thức tính

$$\sum_{i \in C_1, j \in C_1} (A_{ij} - \frac{d_i d_j}{2m})$$

Tính $(A_{ij} - \frac{d_i d_j}{2m})$ cho từng cặp:

① $(1,2): 1 - \frac{3 \cdot 2}{28} = 1 - \frac{6}{28}$

② $(1,3): 1 - \frac{3 \cdot 3}{28} = 1 - \frac{9}{28}$

③ $(1,4): 1 - \frac{3 \cdot 4}{28} = 1 - \frac{12}{28}$

④ $(2,3): 1 - \frac{2 \cdot 3}{28} = 1 - \frac{6}{28}$

⑤ $(2,4): 0 - \frac{2 \cdot 4}{28} = 0 - \frac{8}{28}$

⑥ $(3,4): 1 - \frac{3 \cdot 4}{28} = 1 - \frac{12}{28}$

Tổng cho C_1

$$5 - (\frac{6}{28} + \frac{9}{28} + \frac{12}{28} + \frac{6}{28} + \frac{12}{28} + \frac{8}{28}) = 5 - \frac{53}{28}$$

Tính toán Modularity cho C_2

Tính cho từng cặp:

$$① (5,6): 1 - \frac{4 \cdot 4}{28} = 1 - \frac{16}{28}$$

$$② (5,7): 1 - \frac{4 \cdot 4}{28} = 1 - \frac{16}{28}$$

$$③ (5,8): 1 - \frac{4 \cdot 3}{28} = 1 - \frac{12}{28}$$

$$④ (5,9): 0 - \frac{4 \cdot 1}{28} = 0 - \frac{4}{28}$$

$$⑤ (6,7): 1 - \frac{4 \cdot 4}{28} = 1 - \frac{16}{28}$$

$$⑤ (6,8): 1 - \frac{4 \cdot 3}{28} = 1 - \frac{12}{28}$$

$$⑥ (6,9): 0 - \frac{4 \cdot 1}{28} = 0 - \frac{4}{28}$$

$$⑦ (7,8): 1 - \frac{4 \cdot 3}{28} = 1 - \frac{12}{28}$$

$$⑧ (7,9): 1 - \frac{4 \cdot 1}{28} = 1 - \frac{4}{28}$$

$$⑨ (8,9): 0 - \frac{3 \cdot 1}{28} = 0 - \frac{3}{28}$$

Tổng cho C_2

$$7 - \left(\frac{16}{28} + \frac{16}{28} + \frac{12}{28} + \frac{16}{28} + \frac{12}{28} + \frac{12}{28} + \frac{4}{28} + \frac{4}{28} + \frac{4}{28} + \frac{3}{28} \right) = 7 - \frac{99}{28}$$

Kết quả tính Modularity

Tính Q

$$Q = \frac{1}{2m} [\text{Tổng } C_1 + \text{Tổng } C_2]$$
$$= \frac{1}{28} \left(12 - \frac{133}{28} \right) \approx 0.16$$

Phân tích kết quả

- $Q = 0.16 < 0.3$: Cấu trúc cộng đồng tương đối yếu
- Nguyên nhân:
 - Node 4 có nhiều kết nối với C_2 (2 cạnh)
 - Mật độ kết nối giữa các node trong C_2 không cao
 - Node 9 chỉ có 1 kết nối, làm giảm tính kết dính của C_2

Cải thiện cấu trúc cộng đồng

Gợi ý cải thiện:

- Di chuyển node 4 sang C_2
- Hoặc tách thành 3 cộng đồng:
 - $C'_1: \{1,2,3\}$
 - $C'_2: \{4,5,6,8\}$
 - $C'_3: \{7,9\}$

Lợi ích:

- Tăng mật độ kết nối nội bộ
- Giảm kết nối giữa các cộng đồng
- Cấu trúc cộng đồng rõ ràng hơn
- Giá trị Q có thể tăng lên

Kết luận

Việc chọn phân vùng tối ưu cần cân nhắc:

- Số lượng cộng đồng phù hợp
- Cân bằng giữa kích thước các cộng đồng
- Ý nghĩa thực tế của việc phân chia

Ưu và nhược điểm của Modularity

Ưu điểm

- Đánh giá khách quan chất lượng phân chia cộng đồng
- Không yêu cầu biết trước số lượng cộng đồng
- Dễ hiểu và triển khai

Nhược điểm

- Resolution limit: khó phát hiện cộng đồng nhỏ
- Chi phí tính toán cao với mạng lớn
- Có thể bỏ sót một số cấu trúc cộng đồng

Modularity Matrix

Khái niệm

Modularity matrix là một ma trận được sử dụng để đánh giá độ mạnh của các kết nối trong mạng so với một mô hình ngẫu nhiên. Ma trận này so sánh số lượng kết nối thực tế giữa các đỉnh với số lượng kết nối kỳ vọng dựa trên bậc của các đỉnh.

Định nghĩa toán học

Modularity matrix B được xác định bởi hiệu của hai ma trận:

- Ma trận kề A : biểu diễn kết nối thực tế
- Ma trận kỳ vọng $dd^T/2m$: biểu diễn kết nối kỳ vọng

$$B = A - dd^T/2m$$

$$B_{ij} = A_{ij} - \frac{d_i d_j}{2m}$$

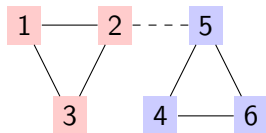
Các thành phần

- A**: Ma trận kề của đồ thị ($A_{ij} = 1$ nếu có cạnh nối i, j ; $A_{ij} = 0$ nếu không)
- d**: Vector bậc của các đỉnh (số cạnh nối với mỗi đỉnh)
- m**: Tổng số cạnh trong đồ thị
- dd^T** : Tích ma trận của vector d với chuyển vị của nó

Ý nghĩa

- B_{ij} so sánh số kết nối thực tế (A_{ij}) với số kết nối kỳ vọng ($\frac{d_i d_j}{2m}$)
- $B_{ij} > 0$: Kết nối mạnh hơn kỳ vọng
- $B_{ij} < 0$: Kết nối yếu hơn kỳ vọng

Ví dụ đơn giản - Tính Modularity Matrix



Thông số:

- $m = 7$ cạnh
- Bậc: $d_1 = 2, d_2 = 3, d_3 = 2$
- $d_4 = 2, d_5 = 3, d_6 = 2$

Ma trận kề A

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Tính toán ma trận kỳ vọng $dd^T/2m$

Vector bậc d

$$d = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 2 \\ 3 \\ 2 \end{pmatrix}$$

$$d^T = (2 \quad 3 \quad 2 \quad 2 \quad 3 \quad 2)$$

Tính dd^T

$$dd^T = \begin{pmatrix} 2 \\ 3 \\ 2 \\ 2 \\ 3 \\ 2 \end{pmatrix} (2 \ 3 \ 2 \ 2 \ 3 \ 2)$$
$$= \begin{pmatrix} 2 \cdot 2 & 2 \cdot 3 & 2 \cdot 2 & 2 \cdot 2 & 2 \cdot 3 & 2 \cdot 2 \\ 3 \cdot 2 & 3 \cdot 3 & 3 \cdot 2 & 3 \cdot 2 & 3 \cdot 3 & 3 \cdot 2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Ma trận kỳ vọng $dd^T/2m$

Kết quả dd^T

$$dd^T = \begin{pmatrix} 4 & 6 & 4 & 4 & 6 & 4 \\ 6 & 9 & 6 & 6 & 9 & 6 \\ 4 & 6 & 4 & 4 & 6 & 4 \\ 4 & 6 & 4 & 4 & 6 & 4 \\ 6 & 9 & 6 & 6 & 9 & 6 \\ 4 & 6 & 4 & 4 & 6 & 4 \end{pmatrix}$$

Ma trận kỳ vọng

$$\frac{dd^T}{2m} = \frac{1}{14} \begin{pmatrix} 4 & 6 & 4 & 4 & 6 & 4 \\ 6 & 9 & 6 & 6 & 9 & 6 \\ 4 & 6 & 4 & 4 & 6 & 4 \\ 4 & 6 & 4 & 4 & 6 & 4 \\ 6 & 9 & 6 & 6 & 9 & 6 \\ 4 & 6 & 4 & 4 & 6 & 4 \end{pmatrix}$$

Tính toán ma trận Modularity B

Ma trận kề **A**:

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Ma trận kỳ vọng $\frac{dd^T}{2m}$:

$$\approx \begin{pmatrix} 0.29 & 0.43 & 0.29 & 0.29 & 0.43 & 0.29 \\ 0.43 & 0.64 & 0.43 & 0.43 & 0.64 & 0.43 \\ 0.29 & 0.43 & 0.29 & 0.29 & 0.43 & 0.29 \\ 0.29 & 0.43 & 0.29 & 0.29 & 0.43 & 0.29 \\ 0.43 & 0.64 & 0.43 & 0.43 & 0.64 & 0.43 \\ 0.29 & 0.43 & 0.29 & 0.29 & 0.43 & 0.29 \end{pmatrix}$$

Ma trận Modularity B - Kết quả

$$B = A - \frac{dd^T}{2m}$$

$$B = \begin{pmatrix} -0.29 & 0.57 & 0.71 & -0.29 & -0.43 & -0.29 \\ 0.57 & -0.64 & 0.57 & -0.43 & 0.36 & -0.43 \\ 0.71 & 0.57 & -0.29 & -0.29 & -0.43 & -0.29 \\ -0.29 & -0.43 & -0.29 & -0.29 & 0.57 & 0.71 \\ -0.43 & 0.36 & -0.43 & 0.57 & -0.64 & 0.57 \\ -0.29 & -0.43 & -0.29 & 0.71 & 0.57 & -0.29 \end{pmatrix}$$

Phân tích kết quả

- Giá trị dương: kết nối thực tế > kỳ vọng (ví dụ: $B_{13} = 0.71$)
- Giá trị âm: kết nối thực tế < kỳ vọng (ví dụ: $B_{15} = -0.43$)
- Giá trị cao trong cùng cộng đồng (ví dụ: $B_{12} = 0.57$)
- Giá trị thấp giữa các cộng đồng (ví dụ: $B_{14} = -0.29$)

3.2.2 Nhát cắt

Khái niệm

Nhát cắt là phương pháp phân hoạch các đỉnh của đồ thị thành hai tập không giao nhau bằng cách cắt các cạnh nối giữa chúng.

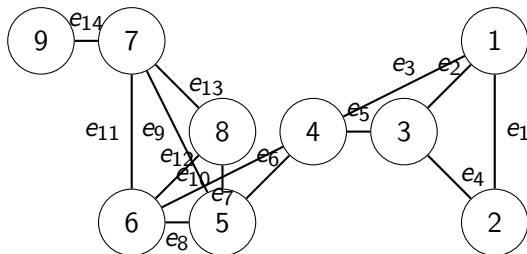
- Mục tiêu: Tìm cách phân chia đồ thị sao cho số cạnh cần cắt là tối thiểu
- Áp dụng: Phát hiện cộng đồng theo hướng Node-Centric
- Kết quả: Các cộng đồng được xác định bởi các nhóm đỉnh sau khi cắt

Đặc điểm quan trọng

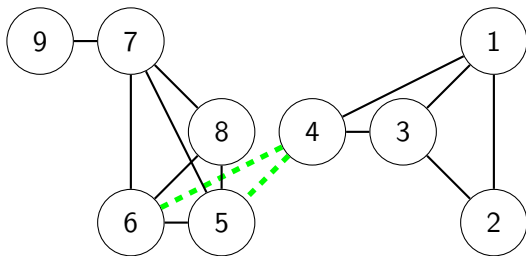
Nhát cắt tối thiểu thường tạo ra các phân hoạch không cân bằng:

- Một tập hợp có thể chỉ chứa một đỉnh (singleton)
- Kích thước các cộng đồng có thể chênh lệch lớn

Bước 1: Đồ thị ban đầu



Bước 2: Xác định nhát cắt chính

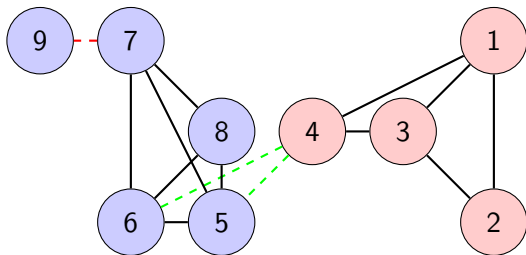


Hình: Nhát cắt chính phân chia thành hai cộng đồng

Nhát cắt chính

- Cắt cạnh e_6 và e_7 (được đánh dấu màu xanh)
- Tạo ra hai nhóm: $\{1, 2, 3, 4\}$ và $\{5, 6, 7, 8, 9\}$

Bước 3: Kết quả cuối cùng



Hình: Kết quả phân chia cộng đồng cuối cùng

- **Nhát cắt xanh:** Phân chia hai cộng đồng chính
- **Nhát cắt đỏ:** Tách node singleton (node 9)
- Kết quả: Ba cộng đồng với kích thước 4-4-1

Các loại nhát cắt - Phần 1

Nhát cắt tối thiểu (Min-Cut)

- Tìm tập cạnh nhỏ nhất cần cắt để chia đồ thị thành 2 phần
- Nhược điểm: Thường cho kết quả phân hoạch không cân bằng
- Có thể tạo ra singleton (một đỉnh đơn lẻ)
- Ví dụ: Node 9 với một cạnh cắt $\{7,9\}$

Nhu cầu cải tiến

Cần thay đổi hàm mục tiêu để xem xét:

- Kích thước của cộng đồng
- Cân bằng giữa các phần được tạo ra
- Tổng số kết nối trong mỗi cộng đồng

Các loại nhất cắt - Phần 2

Ratio Cut

$$\text{Ratio Cut}(\pi) = \frac{1}{k} \sum_{i=1}^k \frac{\text{cut}(C_i, \overline{C_i})}{|C_i|}$$

Trong đó:

- C_i : cộng đồng thứ i
- $|C_i|$: số node trong cộng đồng C_i
- $\text{cut}(C_i, \overline{C_i})$: số cạnh cắt để tạo C_i
- k : số lượng cộng đồng cần tách

Đặc điểm của Ratio Cut: Giá trị càng nhỏ càng tốt

- Cân bằng kích thước của các cộng đồng
- Ưu điểm: Tránh tạo ra singleton; Các cộng đồng có số lượng node tương đương nhau

Normalized Cut

$$\text{Normalized Cut}(\pi) = \frac{1}{k} \sum_{i=1}^k \frac{\text{cut}(C_i, \overline{C_i})}{\text{vol}(C_i)}$$

Trong đó:

- $\text{vol}(C_i)$: tổng số bậc của các node trong C_i
- $\overline{C_i}$: tập các node không thuộc cộng đồng C_i
- $\text{cut}(C_i, \overline{C_i})$: số cạnh nối giữa C_i và phần còn lại

Ưu điểm của Normalized Cut

- Cân nhắc mật độ kết nối của cộng đồng:
 - Không chỉ quan tâm số lượng node
 - Xét đến số lượng kết nối của mỗi node
- Thích hợp cho mạng không đồng nhất:
 - Một số node có nhiều kết nối
 - Một số node có ít kết nối
- Giá trị nhỏ chỉ ra phân hoạch tốt:
 - Ít cạnh cắt
 - Kết nối nội bộ cộng đồng mạnh

Phân hoạch π_1 (cắt $\{7,9\}$):

- Ratio Cut:

$$\frac{1}{2}\left(\frac{1}{1} + \frac{1}{8}\right) = \frac{9}{16} = 0.56$$

- Normalized Cut:

$$\frac{1}{2}\left(\frac{1}{1} + \frac{1}{27}\right) = \frac{14}{27} = 0.52$$

Phân hoạch π_2 (cắt $\{4,5\}$, $\{4,6\}$):

- Ratio Cut:

$$\frac{1}{2}\left(\frac{2}{4} + \frac{2}{5}\right) = \frac{9}{20} = 0.45$$

- Normalized Cut:

$$\frac{1}{2}\left(\frac{2}{12} + \frac{2}{16}\right) = \frac{7}{48} = 0.15$$

Kết luận

- Ratio Cut(π_2) = 0.45 < Ratio Cut(π_1) = 0.56
- Normalized Cut(π_2) = 0.15 < Normalized Cut(π_1) = 0.52
- Chọn phân hoạch π_2 (cắt $\{4,5\}$, $\{4,6\}$) vì cho giá trị nhỏ hơn

Ưu và nhược điểm của phương pháp nhát cắt

Ưu điểm:

- Đơn giản, dễ hiểu
- Hiệu quả với đồ thị nhỏ
- Có nhiều biến thể cải tiến
- Cơ sở cho các thuật toán phức tạp hơn

Nhược điểm:

- Có thể tạo cộng đồng không cân bằng
- Dễ tạo ra singleton (node đơn lẻ)
- Chỉ phân chia thành 2 phần
- Không hiệu quả với đồ thị lớn

3.2.3 Thuật toán Girvan Newman

Giới thiệu

- Được đề xuất bởi Michelle Girvan và Mark Newman (2002)
- Dùng độ đo "betweenness centrality" để phát hiện cộng đồng
- Đặc biệt hiệu quả trong việc phân tích cấu trúc mạng xã hội

Ý tưởng chính

- Tính "edge betweenness" cho mỗi cạnh
- Loại bỏ lần lượt các cạnh có betweenness cao nhất
- Cập nhật lại giá trị betweenness sau mỗi lần loại bỏ

Định nghĩa

Edge Betweenness của cạnh e là tỷ số:

$$BC(e) = \sum_{\substack{u, w \in V \\ u \neq w}} \frac{\sigma_{uw}(e)}{\sigma_{uw}}$$

Trong đó:

- σ_{uw} : số đường đi ngắn nhất giữa u và w
- $\sigma_{uw}(e)$: số đường đi ngắn nhất giữa u và w đi qua cạnh e

Ý nghĩa

- Đo lường tầm quan trọng của cạnh trong việc kết nối các đỉnh
- Cạnh có betweenness cao thường là cầu nối giữa các cộng đồng
- Loại bỏ các cạnh này sẽ tách được các cộng đồng

Thuật toán Girvan Newman

Đầu vào

- Mạng xã hội $G(V, E)$
- V : tập đỉnh
- E : tập cạnh

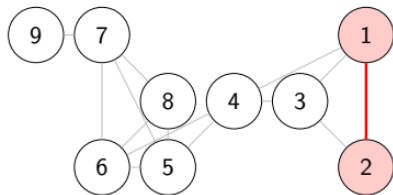
Đầu ra

- Tập các cộng đồng V_1, V_2, \dots, V_k
- $\cup_{i=1}^k V_i = V$ (bao phủ toàn bộ đồ thị)

Các bước thực hiện

- 1 Tính "edge betweenness" cho tất cả các cạnh trong mạng
- 2 Tìm và xóa cạnh có "edge betweenness" cao nhất
- 3 Tính lại "edge betweenness" cho các cạnh bị ảnh hưởng
- 4 Lặp lại bước 2 cho đến khi không còn cạnh nào

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 1



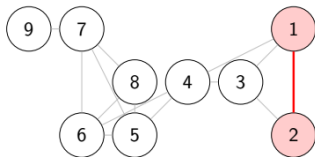
Xét đường đi ngắn nhất

- Shortest Path = $1 \rightarrow 2$
- Độ dài đường đi = 1

Tính toán

- Số đường đi ngắn nhất = 1
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 1



Đóng góp Edge Betweenness

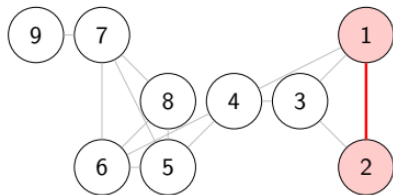
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{1} = 1.00 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 1.00$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 2



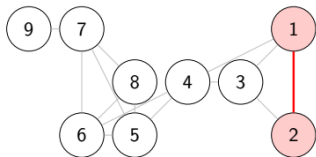
Xét đường đi ngắn nhất

- Shortest Path = $4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 2

Tính toán

- Số đường đi ngắn nhất = 2
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 2



Đóng góp Edge Betweenness

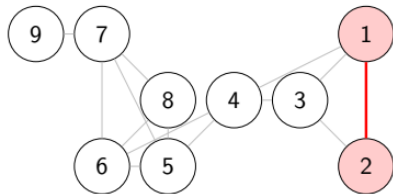
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{2} = 0.50 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 1.00 + 0.50 = 1.50$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 3



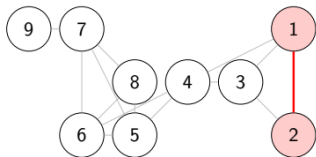
Xét đường đi ngắn nhất

- Shortest Path = $5 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 3

Tính toán

- Số đường đi ngắn nhất = 2
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 3



Đóng góp Edge Betweenness

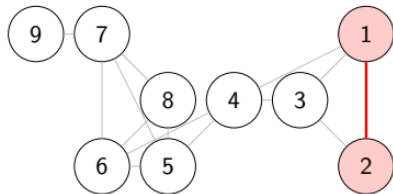
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{2} = 0.50 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 1.50 + 0.50 = 2.00$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 4



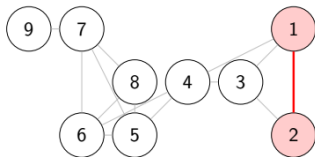
Xét đường đi ngắn nhất

- Shortest Path = $6 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 3

Tính toán

- Số đường đi ngắn nhất = 2
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 4



Đóng góp Edge Betweenness

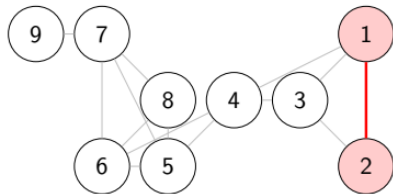
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{2} = 0.50 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 2.00 + 0.50 = 2.50$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 5



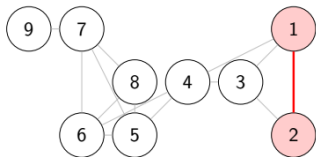
Xét đường đi ngắn nhất

- Shortest Path = $7 \rightarrow 5 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 4

Tính toán

- Số đường đi ngắn nhất = 4
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 5



Đóng góp Edge Betweenness

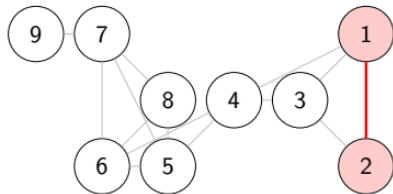
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{4} = 0.25 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 2.50 + 0.25 = 2.75$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 6



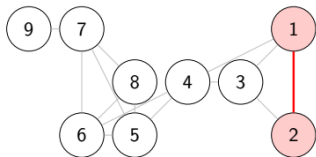
Xét đường đi ngắn nhất

- Shortest Path = $7 \rightarrow 6 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 4

Tính toán

- Số đường đi ngắn nhất = 4
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 6



Đóng góp Edge Betweenness

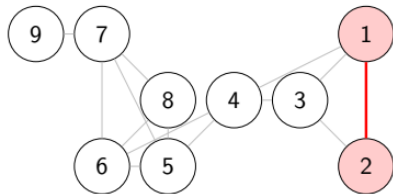
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{4} = 0.25 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 2.75 + 0.25 = 3.00$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 7



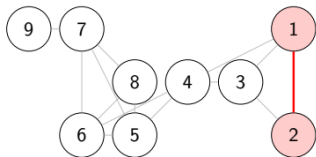
Xét đường đi ngắn nhất

- Shortest Path = $8 \rightarrow 5 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 4

Tính toán

- Số đường đi ngắn nhất = 4
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 7



Đóng góp Edge Betweenness

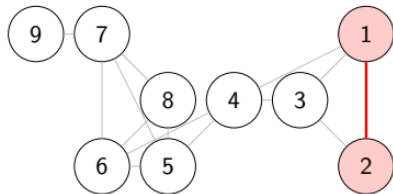
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{4} = 0.25 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 3.00 + 0.25 = 3.25$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 8



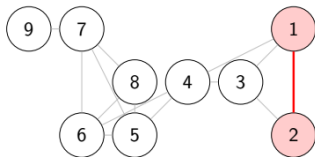
Xét đường đi ngắn nhất

- Shortest Path = $8 \rightarrow 6 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 4

Tính toán

- Số đường đi ngắn nhất = 4
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 8



Đóng góp Edge Betweenness

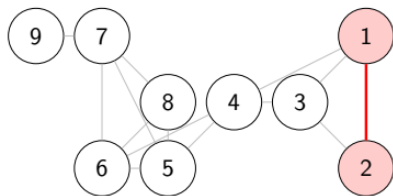
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{4} = 0.25 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 3.25 + 0.25 = 3.50$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 9



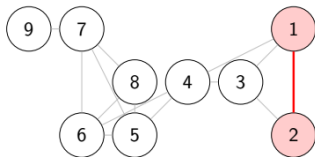
Xét đường đi ngắn nhất

- Shortest Path = $9 \rightarrow 7 \rightarrow 5 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 5

Tính toán

- Số đường đi ngắn nhất = 4
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán 9



Đóng góp Edge Betweenness

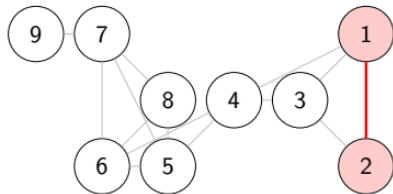
Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{4} = 0.25 \end{aligned}$$

Tổng Edge Betweenness hiện tại

$$BC(\langle 1, 2 \rangle) = 3.50 + 0.25 = 3.75$$

Xét cạnh $\langle 1, 2 \rangle$ - Shortest Path 10 (Cuối cùng)



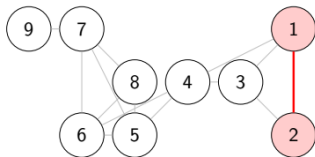
Xét đường đi ngắn nhất

- Shortest Path = $9 \rightarrow 7 \rightarrow 6 \rightarrow 4 \rightarrow 1 \rightarrow 2$
- Độ dài đường đi = 5

Tính toán

- Số đường đi ngắn nhất = 4
- Số đường đi qua cạnh = 1

Xét cạnh $\langle 1, 2 \rangle$ - Kết quả tính toán cuối cùng



Đóng góp Edge Betweenness

Gia tăng Edge Betweenness

$$\begin{aligned} &= \frac{\text{Số đường đi qua cạnh}}{\text{Tổng số đường đi}} \\ &= \frac{1}{4} = 0.25 \end{aligned}$$

Tổng Edge Betweenness cuối cùng

$$BC(\langle 1, 2 \rangle) = 3.75 + 0.25 = 4.00$$

Bảng 3.1. Bảng Edge Betweenness

	1	2	3	4	5	6	7	8	9
1	0	4	1	9	0	0	0	0	0
2	4	0	4	0	0	0	0	0	0
3	1	4	0	9	0	0	0	0	0
4	9	0	9	0	10	10	0	0	0
5	0	0	0	10	0	1	6	3	0
6	0	0	0	10	1	0	6	3	0
7	0	0	0	0	6	6	0	2	8
8	0	0	0	0	3	3	2	0	0
9	0	0	0	0	0	0	8	0	0

- Giá trị tại vị trí (i,j) là Edge Betweenness của cạnh nối đỉnh i và j
- Giá trị 0 cho biết không có cạnh trực tiếp giữa hai đỉnh
- Các giá trị cao nhất (10) ở cạnh $(4,5)$ và $(4,6)$ - các cạnh cầu nối quan trọng

Phân tích kết quả từ Bảng 3.1

- Hai cạnh có Edge Betweenness cao nhất:
 - Cạnh $\langle 4, 5 \rangle$: giá trị = 10
 - Cạnh $\langle 4, 6 \rangle$: giá trị = 10

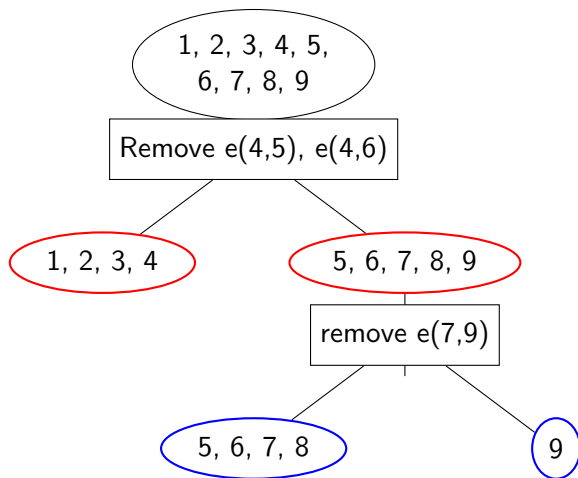
Bước 2: Loại bỏ cạnh

- Chọn cạnh $\langle 4, 5 \rangle$ để loại bỏ
- Lý do: Có Edge Betweenness cao nhất

Bước 3: Cập nhật Edge Betweenness

- Sau khi loại bỏ $\langle 4, 5 \rangle$:
 - Cạnh $\langle 4, 6 \rangle$ trở thành cạnh quan trọng nhất
 - $BC(4, 6)$ tăng lên 20

Quá trình phát hiện cộng đồng



- Bước 1: Loại bỏ cạnh $e(4,5)$ và $e(4,6)$ → Tạo 2 cộng đồng
- Bước 2: Loại bỏ cạnh $e(7,9)$ → Tách node đơn lẻ

Ưu và nhược điểm

Ưu điểm:

- Không cần biết trước số cộng đồng
- Phát hiện được cấu trúc phân cấp
- Kết quả trực quan, dễ hiểu
- Phù hợp với nhiều loại mạng

Nhược điểm:

- Độ phức tạp tính toán cao $O(m^2n)$
- m : số cạnh
- n : số đỉnh
- Không hiệu quả với mạng lớn

Lưu ý quan trọng

- Cần kết hợp với các độ đo chất lượng (như Modularity)
- Có thể dừng thuật toán khi đạt số cộng đồng mong muốn
- Phù hợp với mạng có cấu trúc cộng đồng rõ ràng

3.2.4. Thuật toán dựa trên độ tương tự của Node (Node Similarity)

- Độ tương tự của node dựa trên các node láng giềng. Thuật toán này theo hướng Node-Centric Community.
- Hai node có cấu trúc tương tự nếu chúng có chung tập các node láng giềng

Độ tương tự Jaccard

$$Jaccard(v_i, v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$$

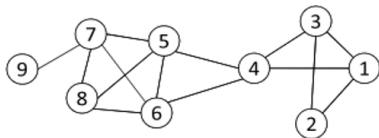
trong đó:

- N_i = Tập hợp các node láng giềng của node i
- N_j = Tập hợp các node láng giềng của node j
- $|N_i \cap N_j|$ = Số lượng node láng giềng chung
- $|N_i \cup N_j|$ = Tổng số node láng giềng duy nhất

Ý nghĩa:

- Giá trị nằm trong khoảng $[0,1]$
- Bằng 0: hai node không có node láng giềng chung
- Bằng 1: hai node có cùng tập node láng giềng
- Càng gần 1: cấu trúc láng giềng càng tương tự nhau
- Đo lường mức độ chồng lấp của các node láng giềng

Ví dụ: Độ tương tự Jaccard



Hình 3.1. Đồ thị để khám phá cộng đồng

Cho node 4 và node 6:

$$\begin{aligned} Jaccard(4, 6) &= \frac{|\{5\}|}{|\{1, 3, 4, 5, 6, 7, 8\}|} \\ &= \frac{1}{7} \end{aligned}$$

Kết quả này cho thấy node 4 và node 6 chỉ có một node láng giềng chung trong tổng số bảy node láng giềng.

Độ tương tự Cosine

$$\text{Cosine}(v_i, v_j) = \frac{|N_i \cap N_j|}{\sqrt{|N_i||N_j|}}$$

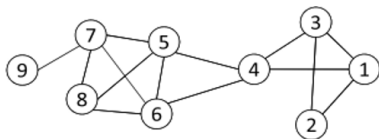
trong đó:

- N_i = Tập hợp các node láng giềng của node i
- N_j = Tập hợp các node láng giềng của node j
- $|N_i \cap N_j|$ = Số lượng node láng giềng chung
- $|N_i|$ = Số lượng node láng giềng của node i
- $|N_j|$ = Số lượng node láng giềng của node j

Ý nghĩa:

- Đo lường góc giữa hai vector láng giềng, không phụ thuộc vào kích thước tuyệt đối; Phù hợp khi so sánh các node có số lượng láng giềng khác biệt nhiều; Cho phép so sánh công bằng hơn giữa các node có độ lớn khác nhau so với Jaccard

Ví dụ: Độ tương tự Cosine



Hình 3.1. Đồ thị để khám phá cộng đồng

Cho node 4 và node 6:

$$\begin{aligned}\text{Cosine}(4, 6) &= \frac{1}{\sqrt{4 \cdot 4}} \\ &= \frac{1}{4}\end{aligned}$$

Kết quả này cho thấy độ tương tự cosine giữa node 4 và node 6 là 0.25, thể hiện mức độ tương đồng cấu trúc tương đối thấp.

Sau khi tính toán các độ đo tương tự:

- Ta thu được ma trận độ tương tự giữa tất cả các node
- Ma trận này có thể được sử dụng làm đầu vào cho các thuật toán phân cụm
- Thuật toán k-means có thể được áp dụng để khám phá các cụm node
- Các cụm đại diện cho các nhóm node có tính chất cấu trúc tương tự nhau

Độ tương tự dựa trên sự gần gũi của node

- Xét sự gần gũi giữa các node thông qua vai trò trung gian
- Đánh giá khả năng truyền tải thông tin giữa các node
- Xem xét đường đi qua các node láng giềng trung gian

Công thức tính độ tương tự

$$S_{ij} = \sum_{z \in T(i) \cap T(j)} \frac{1}{k(z)}$$

trong đó:

- $T(i)$ = Tập các node láng giềng của node i
- $T(j)$ = Tập các node láng giềng của node j
- z = Node chung thuộc cả $T(i)$ và $T(j)$
- $k(z)$ = Số bậc của node z
- $S_{ij} = 0$ khi node i không kết nối trực tiếp với node j

Ý nghĩa của độ đo

- Đo lường mức độ gần gũi thông qua các node trung gian
- Node trung gian có bậc cao sẽ đóng góp ít hơn vào độ tương tự
- Node trung gian có bậc thấp thể hiện kết nối chuyên biệt hơn
- Phản ánh khả năng truyền thông tin giữa hai node
- Càng có nhiều đường đi ngắn qua node trung gian, độ tương tự càng cao

Ma trận độ tương tự

- Với mạng n node, tính S_{ij} cho mọi cặp node (i,j)
- Kết quả được lưu vào ma trận S kích thước $n \times n$
- S_{ij} là độ tương tự giữa node i và node j
- Ma trận S đối xứng: $S_{ij} = S_{ji}$
- Giá trị S_{ij} phụ thuộc vào:
 - Số lượng node trung gian chung
 - Bậc của các node trung gian