ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

HOMEWORK 03

Khai thác dữ liệu đồ thị

Sinh viên thực hiện: 21127229 - Dương Trường Bình

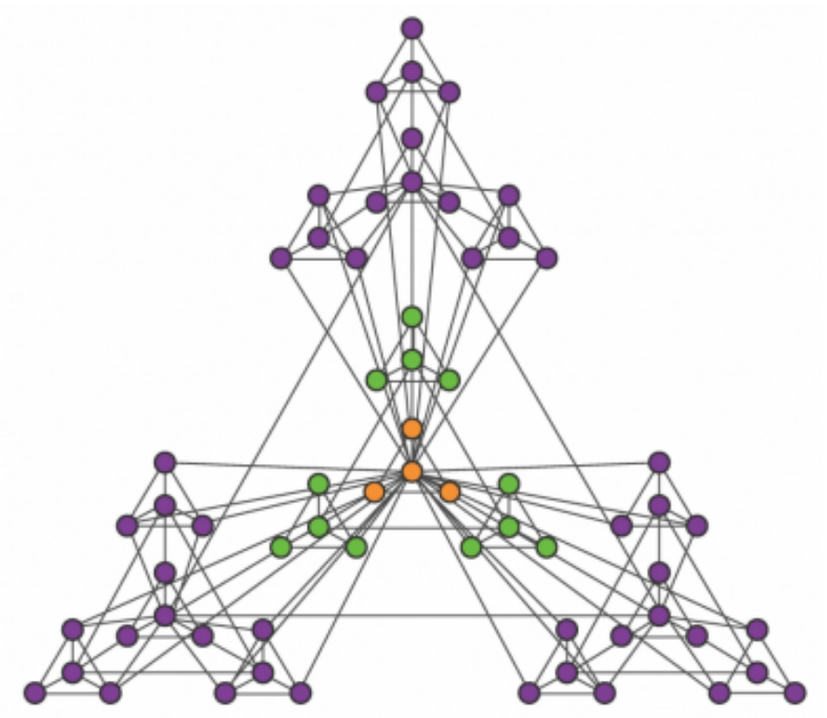
Giảng viên hướng dẫn: Lê Ngọc Thành
Lê Nhựt Nam

Lớp: 21KHDL

Mục lục

1 Problem 1

Câu hỏi: Tính toán lũy thừa bậc của bạn phân cấp dưới đây.



Trả lời

1.1 Phân tích cấu trúc mạng

Mạng phân cấp trong hình được chia thành 3 cấp, mỗi cấp được biểu diễn bằng một màu khác nhau:

- Cấp 1 (màu cam): Đỉnh gốc và các đỉnh kết nối trực tiếp
- Cấp 2 (màu xanh lá): Các đỉnh kết nối với đỉnh cấp 1
- Cấp 3 (màu tím): Các đỉnh ngoài cùng

1.2 Thống kê số liệu

- Cấp 1:
  - Số đỉnh: $n_1 = 4$
  - Số cạnh nội bộ: $l_1 = 6$

- Cấp 2:
  - Số đỉnh: $n_2 = 16$
  - Số cạnh nội bộ: $l_2 = 24$
  - Tổng số cạnh (bao gồm cả cạnh kết nối với cấp 1): $L_2 = 30$
- Cấp 3:
  - Số đỉnh: $n_3 = 64$
  - Số cạnh nội bộ: $l_3 = 120$
  - Tổng số cạnh (bao gồm cả cạnh kết nối với cấp 2): $L_3 = 126$

1.3 Tính toán lũy thừa bậc

Để tính lũy thừa bậc γ, ta sử dụng công thức:

$$\gamma = 1 + \frac{\ln(n)}{\ln(L)}$$

trong đó n là số đỉnh và L là tổng số cạnh.

- Cấp 1:

$$\begin{aligned}\gamma_1 &= 1 + \frac{\ln(n_1)}{\ln(l_1)} \\ &= 1 + \frac{\ln(4)}{\ln(6)} \\ &\approx 1 + \frac{1.386}{1.792} \\ &\approx 1.774\end{aligned}$$

- Cấp 2:

$$\begin{aligned}\gamma_2 &= 1 + \frac{\ln(n_2)}{\ln(L_2)} \\ &= 1 + \frac{\ln(16)}{\ln(30)} \\ &\approx 1 + \frac{2.773}{3.401} \\ &\approx 1.815\end{aligned}$$

- Cấp 3:

$$\begin{aligned}\gamma_3 &= 1 + \frac{\ln(n_3)}{\ln(L_3)} \\ &= 1 + \frac{\ln(64)}{\ln(126)} \\ &\approx 1 + \frac{4.158}{4.835} \\ &\approx 1.860\end{aligned}$$

2 Problem 2

Câu hỏi: Xem xét một lưới trong một chiều (1-dimensional) gồm N nút. Lưới này hình thành một đường tròn mà trong đó mỗi nút liên kết với hai láng giềng của nó. Phân hoạch đường tròn này thành n_c cụm liên tiếp nhau với kích thước $N_c = \frac{N}{n_c}$

a. Tính toán modularity của phân hoạch thu được.

b. Dựa trên giả thiết cực đại modularity (Maximum Modularity Hypothesis), giá trị lớn nhất của M_c tương ứng với phân hoạch tốt nhất. Hãy đưa ra kích thước cộng đồng n_c tương ứng với phân hoạch tốt nhất

Trả lời

2.1 (a) Tính toán modularity của phân hoạch thu được:

Modularity được định nghĩa bởi công thức:

$$M = \sum_{c=1}^{n_c} \left[\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right]$$

Trong đó:

- L: Tổng số cạnh trong đồ thị
- L_c: Số cạnh trong cụm c
- k_c: Tổng bậc của các nút trong cụm c

Đối với lưới tròn một chiều:

- $L = N$ (mỗi nút có 2 cạnh, nhưng mỗi cạnh được đếm 2 lần)

- $l_c = N_c - 1 = \frac{N}{n_c} - 1$ (số cạnh trong mỗi cụm)
- $k_c = 2N_c = \frac{2N}{n_c}$ (mỗi nút có bậc 2)

Thế vào công thức modularity ta được:

$$M = \sum_{c=1}^{n_c} \left[\frac{N_c - 1}{N} - \left(\frac{2 \times \frac{N}{n_c}}{2 \times N} \right)^2 \right]$$
$$= \sum_{c=1}^{n_c} \left[\frac{N_c - 1}{N} - \left(\frac{1}{n_c} \right)^2 \right]$$

Với n_c cụm giống nhau nên ta có:

$$M = n_c \left[\frac{N_c - 1}{N} - \left(\frac{1}{n_c} \right)^2 \right]$$
$$= n_c \left[\frac{\frac{N}{n_c} - 1}{N} - \left(\frac{1}{n_c} \right)^2 \right]$$
$$= n_c \left[\frac{N - n_c}{N \times n_c} - \frac{1}{n_c^2} \right]$$
$$= \frac{N - n_c}{N} - \frac{1}{n_c}$$
$$= 1 - \frac{n_c}{N} - \frac{1}{n_c}$$

2.2 (b) Xác định kích thước cộng đồng n_c tương ứng với phân hoạch tốt nhất

Trong phần này, chúng ta sẽ tìm kích thước cộng đồng tối ưu để đạt được modularity cao nhất. Với công thức modularity đã tìm được ở phần (a):

$$M = 1 - \frac{n_c}{N} - \frac{1}{n_c}$$

Để tìm giá trị tối ưu của n_c, ta sẽ sử dụng phương pháp tối ưu hóa bằng đạo hàm.

Bước 1: Tìm cực trị của M

Đầu tiên, chúng ta sẽ tính đạo hàm của M theo n_c và đặt nó bằng 0 để tìm điểm cực trị:

- Tính đạo hàm:
$$\frac{dM}{dn_c} = -\frac{1}{N} + \frac{1}{n_c^2} = 0$$

- Giải phương trình này:
$$\begin{aligned} \frac{1}{N} &= \frac{1}{n_c^2} \\ \Leftrightarrow n_c^2 &= N \\ \Leftrightarrow n_c &= \sqrt{N} \end{aligned}$$

Chúng ta đã tìm được một điểm cực trị tại $n_c = \sqrt{N}$. Tuy nhiên, để chắc chắn đây là điểm cực đại, ta cần kiểm tra thêm.

Bước 2: Kiểm tra điều kiện cực đại

Để xác định liệu điểm cực trị này có phải là cực đại hay không, ta tính đạo hàm bậc hai:

- Tính đạo hàm bậc hai:
$$\frac{d^2M}{dn_c^2} = -\frac{2}{n_c^3}$$

- Nhận xét: Tại $n_c = \sqrt{N}$, $\frac{d^2M}{dn_c^2} < 0$, xác nhận đây là điểm cực đại.

Kết quả này cho thấy $n_c = \sqrt{N}$ thực sự là giá trị tối ưu mà chúng ta đang tìm kiếm.

Bước 3: Tính giá trị modularity tối ưu

Bây giờ, chúng ta sẽ tính giá trị modularity tối đa bằng cách thay $n_c = \sqrt{N}$ vào công thức ban đầu:

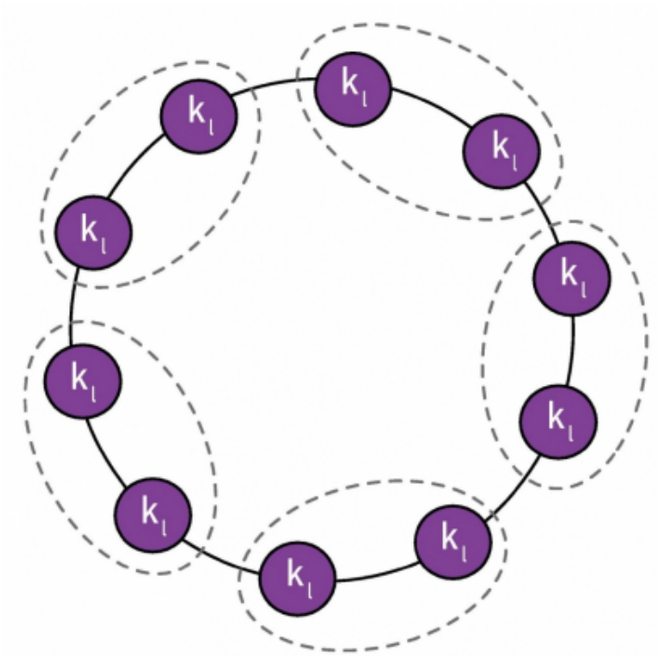
- Thay $n_c = \sqrt{N}$ vào công thức modularity:
$$\begin{aligned} M_c &= 1 - \frac{\sqrt{N}}{N} - \frac{1}{\sqrt{N}} \\ &= 1 - \frac{1}{\sqrt{N}} - \frac{1}{\sqrt{N}} \\ &= 1 - \frac{2}{\sqrt{N}} \end{aligned}$$

Kết luận:

- Kích thước cộng đồng tối ưu: $n_c = \sqrt{N}$
- Giá trị modularity tối ưu: $M_c = 1 - \frac{2}{\sqrt{N}}$

3 Problem 3

Câu hỏi: Xem xét một mạng lưới bao gồm n_c cliques (đồ thị con đầy đủ), mỗi clique có N_c nút và $\frac{m(m-1)}{2}$ liên kết. Các clique lân cận được nối bằng một cạnh đơn. Hình dưới đây minh họa cấu trúc này. Mạng này có cấu trúc cộng đồng rõ ràng, với mỗi cộng đồng tương ứng với một clique.



(a) Xác định modularity M_{single} của phân hoạch tự nhiên này, và modularity M_{pairs} của phân hoạch mà trong đó các cặp lân cận clique được gộp vào một cộng đồng đơn (như được mô tả trong hình trên bởi các đường đứt khúc).

(b) Chứng minh rằng chỉ khi $n_c < 2L$ thì môđun tối đa sẽ dự đoán phân vùng cộng đồng chính xác về mặt trực quan, trong đó $L = nc\frac{m(m-1)}{2} + nc$

(c) Đưa ra quan điểm khi vi phạm bất đẳng thức trên.

Trả lời

3.1 (a) Tính toán Modularity cho hai phương án phân hoạch

Trong phần này, chúng ta sẽ tính toán modularity cho hai cách phân hoạch: M_{single} (mỗi clique là một cộng đồng) và M_{pairs} (các cặp clique lân cận được gộp thành một cộng đồng).

1. Tính toán M_{single}:

Xác định các thông số cần thiết:

- Số đỉnh trong mỗi cộng đồng: m
- Số liên kết nội bộ mỗi cộng đồng: $l_c = \frac{m(m-1)}{2}$
- Tổng số liên kết trong mạng: $L = n_c\frac{m(m-1)}{2} + n_c = \frac{n_c(m(m-1)+2)}{2}$
- Tổng bậc của các đỉnh trong cộng đồng: $k_c = m(m-1) + 2$

Áp dụng công thức modularity:

$$\begin{aligned}
M_{single} &= n_c \left(\frac{l_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right) \\
&= n_c \left(\frac{m(m-1)}{2L} - \left(\frac{m(m-1)+2}{2L} \right)^2 \right) \\
&= n_c \left(\frac{m(m-1)}{2L} - \frac{(m(m-1)+2)^2}{4L^2} \right) \\
&= n_c \left(\frac{m(m-1)}{2L} - \frac{m^2(m-1)^2 + 4m(m-1) + 4}{4L^2} \right)
\end{aligned}$$

Thay $L = \frac{n_c(m(m-1)+2)}{2}$:

$$\begin{aligned}
M_{single} &= n_c \left(\frac{m(m-1)}{n_c(m(m-1)+2)} - \frac{(m(m-1)+2)^2}{n_c^2(m(m-1)+2)^2} \right) \\
&= \frac{m(m-1)}{m(m-1)+2} - \frac{1}{n_c}
\end{aligned}$$

2. Tính toán M_{pairs}:

Xác định các thông số cho phân hoạch cặp:

- Số đỉnh trong mỗi cộng đồng kép: $2m$
- Số liên kết nội bộ mỗi cộng đồng kép: $l_c = m(m-1)+1$
- Tổng bậc của các đỉnh trong cộng đồng kép: $k_c = 2m(m-1)+4$

Áp dụng công thức modularity:

$$\begin{aligned} M_{pairs} &= \frac{n_c}{2}\left(\frac{l_c}{L} - \left(\frac{k_c}{2L}\right)^2\right) \\ &= \frac{n_c}{2}\left(\frac{m(m-1)+1}{L} - \left(\frac{2m(m-1)+4}{2L}\right)^2\right) \\ &= \frac{n_c}{2}\left(\frac{m(m-1)+1}{L} - \frac{(2m(m-1)+4)^2}{4L^2}\right) \\ &= \frac{n_c}{2}\left(\frac{m(m-1)+1}{L} - \frac{4m^2(m-1)^2+16m(m-1)+16}{4L^2}\right) \end{aligned}$$

Thay $L = \frac{n_c(m(m-1)+2)}{2}$ và rút gọn:

$$\begin{aligned} M_{pairs} &= \frac{n_c}{2}\left(\frac{2(m(m-1)+1)}{n_c(m(m-1)+2)} - \frac{4(m(m-1)+2)^2}{n_c^2(m(m-1)+2)^2}\right) \\ &= \frac{m(m-1)+1}{m(m-1)+2} - \frac{2}{n_c} \end{aligned}$$

Kết luận: Các biểu thức cuối cùng cho M_{single} và M_{pairs}:

$$M_{single} = \frac{m(m-1)}{m(m-1)+2} - \frac{1}{n_c}$$

$$M_{pairs} = \frac{m(m-1)+1}{m(m-1)+2} - \frac{2}{n_c}$$

3.2 (b) Chứng minh điều kiện cho phân hoạch tối ưu

Để phân hoạch tự nhiên (mỗi clique là một cộng đồng) là tối ưu, chúng ta cần chứng minh rằng $M_{single} > M_{pairs}$.

$$M_{single} > M_{pairs}$$

$$\Leftrightarrow \frac{m(m-1)}{m(m-1)+2} - \frac{1}{n_c} > \frac{m(m-1)+1}{m(m-1)+2} - \frac{2}{n_c}$$

- Đưa về cùng mẫu số và rút gọn

$$\Leftrightarrow \frac{m(m-1)}{m(m-1)+2} - \frac{m(m-1)+1}{m(m-1)+2} > \frac{1}{n_c} - \frac{2}{n_c}$$

$$\Leftrightarrow \frac{-1}{m(m-1)+2} > -\frac{1}{n_c}$$

- Đảo dấu bất đẳng thức

$$\Leftrightarrow \frac{1}{m(m-1)+2} < \frac{1}{n_c}$$

- Lấy nghịch đảo hai vế

$$\Leftrightarrow m(m-1)+2 > n_c$$

- Nhân cả hai vế với n_c

$$\Leftrightarrow n_c(m(m-1)+2) > n_c^2$$

- Thay $L = n_c\frac{m(m-1)}{2} + n_c = \frac{n_c(m(m-1)+2)}{2}$

$$\Leftrightarrow 2L > n_c^2$$

- Lấy căn bậc hai hai vế

$$\Leftrightarrow \sqrt{2L} > n_c$$

Do đó, điều kiện cuối cùng là:

$$n_c < \sqrt{2L}$$

3.3 (c) Phân tích khi vi phạm điều kiện

Khi $n_c \geq \sqrt{2L}$, phân hoạch theo cặp clique có thể có modularity cao hơn phân hoạch tự nhiên. Điều này có những ý nghĩa quan trọng:

1. **Mất mát thông tin chi tiết:** Phương pháp phát hiện cộng đồng có thể bỏ qua các cấu trúc cộng đồng tinh vi, gộp các cộng đồng nhỏ thành cộng đồng lớn hơn.

2. **Độ phân giải của phân tích:** Khi mạng trở nên lớn và phức tạp, có thể cần điều chỉnh phương pháp để phát hiện cộng đồng ở các quy mô khác nhau.

3. **Giới hạn của modularity:** Kết quả này cho thấy việc tối đa hóa modularity có thể không luôn dẫn đến phân hoạch cộng đồng trực quan nhất.

4. **Cân nhắc bối cảnh:** Trong thực tế, việc phân tích cộng đồng nên kết hợp cả các phương pháp định lượng (như modularity) và hiểu biết về bối cảnh cụ thể của mạng.

Kết luận: Kết quả này nhấn mạnh tầm quan trọng của việc cẩn thận trong việc áp dụng các phương pháp phát hiện cộng đồng tự động. Trong nhiều trường hợp, sự kết hợp giữa phân tích định lượng và đánh giá định tính có thể cung cấp cái nhìn toàn diện hơn về cấu trúc cộng đồng thực sự trong mạng.

4 Problem 4

Câu hỏi: Chứng minh rằng giá trị cực đại của modularity M, được định nghĩa bởi

$$M = \sum_{c=1}^{n_c} \left(\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right),$$

thì không vượt quá 1.

Trả lời

Trong công thức:

- n_c: số lượng cộng đồng
- L_c: số liên kết nội bộ trong cộng đồng c
- L: tổng số liên kết trong đồ thị
- k_c: tổng bậc của các đỉnh trong cộng đồng c

Chứng minh:

1. **Tách biểu thức:** Chia M thành hai phần:

$$M = \underbrace{\sum_{c=1}^{n_c} \frac{L_c}{L}}_{\text{Phần 1}} - \underbrace{\sum_{c=1}^{n_c} \left(\frac{k_c}{2L}\right)^2}_{\text{Phần 2}}$$

2. **Phân tích Phần 1:**

- Tổng số liên kết nội bộ không vượt quá tổng số liên kết: $\sum_{c=1}^{n_c} L_c \leq L$
- Do đó: $\sum_{c=1}^{n_c} \frac{L_c}{L} \leq 1$

3. **Phân tích Phần 2:** Sử dụng bất đẳng thức Cauchy-Schwarz

$$\left(\sum_{c=1}^{n_c} k_c\right)^2 \leq n_c \sum_{c=1}^{n_c} k_c^2$$

Vì $\sum_{c=1}^{n_c} k_c = 2L$,

$$\Leftrightarrow 4L^2 \leq n_c \sum_{c=1}^{n_c} k_c^2$$

$$\Leftrightarrow \sum_{c=1}^{n_c} k_c^2 \geq \frac{4L^2}{n_c}$$

Do đó:

$$\sum_{c=1}^{n_c} \left(\frac{k_c}{2L}\right)^2 \geq \frac{1}{n_c}$$

4. **Kết hợp hai phần:**

$$M \leq 1 - \frac{1}{n_c} < 1$$

Kết luận: Giá trị cực đại của modularity M luôn nhỏ hơn 1.

Tài liệu tham khảo

[1] Network Science Book. *Network Science*. Available online: http://networksciencebook.com. (Accessed: August 2024).

[2] Slide của thầy Lê Ngọc Thành, *Topic 07 - Community Detection*, Trường Đại học Khoa học Tự nhiên, 2024