Tên học phần:	**Khai thác Dữ liệu Đồ thị**	Mã HP:	**CSC17103**
Thời gian làm bài:	**07 ngày**	Ngày nộp:	**25/07/2023**

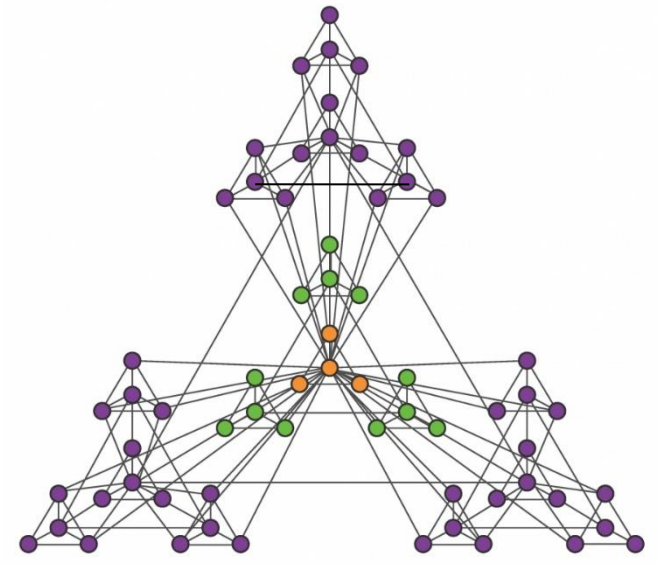
HOMEWORK 03: COMMUNITY DETECTION

Họ tên sinh viên: Võ Văn Hoàng **MSSV**: 20127028

BÀI TRÌNH BÀY

1. **Problem 1. Hierarchical Networks:**

- Calculate the degree exponent of the hierarchical network shown in figure below.



Firgure 01. Hierarchical Networks

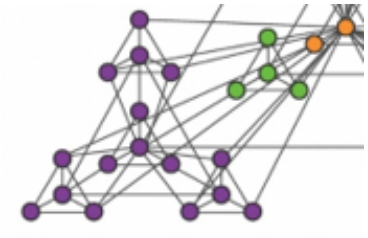
- Ta có: Số mũ bậc của một mạng phân cấp (degree exponent of the hierarchical network) là chỉ số đo lường cách mà số lượng kết nối của một node trong mạng (bậc của node đó) thay đổi khi cấp bậc của cấu trúc phân cấp tăng lên..
- Ta có công thức sau:

$$\gamma = 1 + \frac{\ln(M)}{\ln(M-1)}$$

- Trong đó: γ: là degree exponent

M: là hệ số nhân bản của mô hình mạng phân cấp, là số lượng nút được tạo ra cho mỗi nút ở cấp trước đó của hierarchical network.

- Nhìn vào hình ta có thể thấy được trị số của M = 4, cụ thể như sau:



- Từ 1 node màu vàng có thể tạo ra được 4 node màu xanh lá cây.
- Cũng như từ 1 node màu vàng cũng có thể tạo ra 4 node màu tím.

Do đó: $\gamma = 1 + \frac{\ln(M)}{\ln(M-1)} = 1 + \frac{\ln(4)}{\ln(3)} = 2.262$

Vậy: degree exponent của hierarchical network trên là 2.262

2. Problem 2. Communities on a Circle:

- Consider a one dimensional lattice with N nodes that form a circle, where each node connects to its two neighbors. Partition the line into n_c consecutive clusters of size $N_c=N/n_c$.

a. Calculate the modularity of the obtained partition.

- Ta có công thức tính Modularity của 1 community sau: $M_c = \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2$

- Với: **L_c:** là số lượng cạnh trong community.
  k_c: là tổng số bậc của node trong 1 community.
  L: là tổng số cạnh trong network.

- Tiếp theo ta sẽ tính giá trị E_c, vì trong 1 community gồm N_c nodes, mỗi nodes kết nối với 2 neighbors nên ta có được số cạnh trong community đó là $L_c = N_c - 1$, và vì đây là 1 circle nên L = N (số cạnh bằng với số nodes).
- Ngoài ra, mỗi nodes kết nối với 2 neighbors nên số bậc của 1 node sẽ là 2
  → Tổng số bậc của node trong community $\mathbf{k_c}$ = 2. N_c
- Vậy ta có: $M_c = \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2 = \frac{N_c - 1}{N} - \left(\frac{2.N_c}{2N}\right)^2$ (1)
- Mà theo đề thì: $N_c = \frac{N}{n_c}$, thay vào (1) ta được:

$$M_c = \frac{\frac{N}{n_c} - 1}{N} - \left(\frac{\frac{2.N}{n_c}}{2N}\right)^2 = \frac{N - n_c}{N.n_c} - \left(\frac{1}{n_c}\right)^2 = \frac{1}{n_c} - \frac{1}{N} - \frac{1}{{n_c}^2}$$

■ Ngoài ra, ta lại từ dữ kiện đề bài là partiotion thành n_c cụm liên tiếp nên:

$$M = M_c . n_c = \left(\frac{1}{n_c} - \frac{1}{N} - \frac{1}{{n_c}^2}\right) . n_c = 1 - \frac{n_c}{N} - \frac{1}{n_c}$$

b. According to the Maximum Modularity Hypothesis, the maximum of M_c corresponds to the best partition. Obtain the community size n_c corresponding to the best partition.

■ Theo câu a thì M $= 1 - \frac{n_c}{N} - \frac{1}{n_c}$

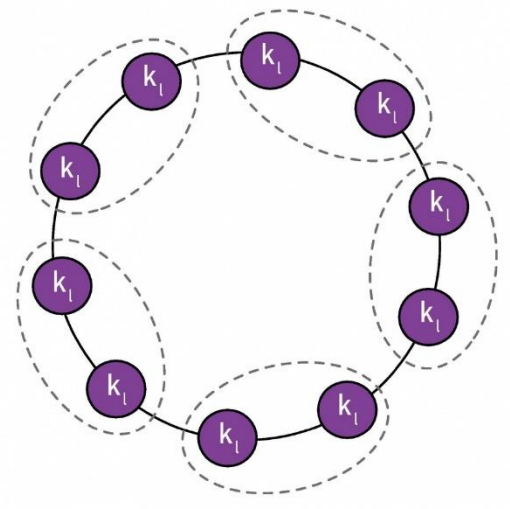
→ $\frac{dM}{dn_c} = \frac{d}{dx}\left(1 - \frac{n_c}{N} - \frac{1}{n_c}\right) = 0 - \frac{1}{N} + \frac{1}{{n_c}^2}$

■ Để biết cực trị ta cho $\frac{dM}{dn_c} = 0 \rightarrow \frac{1}{N} = \frac{1}{{n_c}^2} \rightarrow N = {n_c}^2$ hay $\sqrt{N} = n_c$

Vậy: M đạt cực đại khi $n_c = \sqrt{N}$

3. Problem 3. Modularity Resolution Limit:

- Consider a network consisting of a ring of n_c cliques, each clique having N_c nodes and m(m–1)/2 links. The neighboring cliques are connected by a single link (Figure 2). The network has an obvious community structure, each community corresponding to a clique.



Firgure 02. Modularity

a. Determine the modularity M_{single} of this natural partition, and the modularity M_{pairs} of the partition in which pairs of neighboring cliques are merged into a single community, as indicated by the dotted lines in Figure 2.

- Ta có công thức tổng quát hơn từ câu 2, công thức tính Modularity là:

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

- Vì đồ thị có hình chiếc nhẫn, mỗi clique lại có N_c node với m(m–1)/2 liên kết nên:

$$N_C = m$$

- Hơn nữa từ giả thiết ta cũng biết rằng hình đã cho có n_c cliques nên network có tổng cộng $N = n_c.m$ (nodes) và $L = n_c.m(m-1)/2 + n_c$ (links) và tổng số bậc $k_c = m(m-1) + 2.$
- Cấu trúc natural community của network được biểu diễn bởi partition trong đó mỗi community ứng với 1 clique duy nhất nên:

$$M_{single} = \sum_{c=1}^{n_c} \left[\left\{ \frac{\frac{m(m-1)}{2}}{n_c \left[\frac{m(m-1)}{2} + 1 \right]} \right\} - \left\{ \left(\frac{m(m-1)+2}{2.n_c \left[\frac{m(m-1)}{2} + 1 \right]} \right)^2 \right\} \right]$$

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

- Tương tự với giá trị Modularity M_{pairs} của partition là các cặp neighboring cliques được sát nhập lại thành single community như được khoanh vùng bởi cát nét chấm trên Firgure 2 thì L_c sẽ gồm có $m(m-1)$cặp cạnh neighbor clique và 1 cạnh nối chúng lại với nhau. Ngoài ra, tổng số bậc k_c được tính gồm có $2m(m-1)+2$ bậc từ cặp cạnh neighbor cliques và có thêm 2 bậc từ cặp cạnh nối các community với nhau nên:

$$M_{pairs} = \sum_{c=1}^{n_c} \left[\left\{ \frac{m(m-1)+1}{n_c \left[\frac{m(m-1)}{2} + 1 \right]} \right\} - \left\{ \left(\frac{2m(m-1)+4}{2.n_c \left[\frac{m(m-1)}{2} + 1 \right]} \right)^2 \right\} \right]$$

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

b. Show that only for $n_c < 2L$ will the modularity maximum predict the intuitively correct community partition, where $L = \frac{n_c m(m-1)}{2} + n_c$

- Để chứng minh được điều này, ta dùng phương pháp thế, nếu bất đẳng thức tạo thành luôn đúng thì đó là điều cần phải chứng minh.
- $n_c < 2L$

  → $n_c < 2.\left[\frac{n_c m(m-1)}{2} + n_c\right]$

  → $n_c < n_c.m(m-1) + 2.n_c$

  → $0 < n_c.m(m-1) + n_c$

  → $n_c.[m(m-1) + 1] > 0$

  → $n_c.[m^2 - m + 1] > 0$ (**)

  Ta có: $\begin{cases} n_c \text{ là số clique của network nên sẽ luôn } \geq 1 \\ (m^2 - m + 1) = \left(m - \frac{1}{2}\right)^2 + \frac{3}{4} \geq \frac{3}{4} \end{cases}$

  Do đó (**) luôn luôn là một bất đẳng thức đúng. Vậy khẳng định đề bài đưa ra là hoàn toàn đúng.

c. Discuss the consequences of violating the above inequality.

- Khi $n_c \geq 2L$ trong community c, điều này có nghĩa là community đó có số cạnh nội bộ vượt quá số cạnh ngẫu nhiên dự kiến. Điều này dẫn đến các hệ quả sau:
- Độ phân giải không đủ: community kết nối quá chặt, làm cho thuật toán hợp nhất các sub-communities nhỏ thành các community lớn, dẫn đến thiếu cấu trúc chi tiết. Ngoài ra, các communities nhỏ có ý nghĩa bị hợp nhất vào các communities lớn, do đó nó đã phần nào giấu đi cấu trúc cộng đồng thực tế và làm giảm độ chính xác. Hoặc, community sẽ có thể gặp các sự nhiễu hoặc ngẫu nhiên, xác định các cộng đồng nhỏ không có ý nghĩa hơn không phản ánh cấu trúc thực tế.
- Giá trị Modularity bị làm lệch: Số cạnh nội bộ dư thừa làm tăng giá trị Modularity, tạo ra nhầm lẫn có thể nghiêm trọng về chất lượng cao hơn trong partition.

4. Problem 4. Modularity Maximum:

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

- Show that the maximum value of modularity M cannot exceed one.

L_c: số cạnh trong community c

L: tổng số cạnh trong cả network

k_c: Tổng số bậc của các nodes trong community c.

n_c: là số community trong cả network

- Mục tiêu là ta sẽ đi chứng minh $M \leq 1$
- Ta có: $M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2\right]$
- Trong đẳng thức này ta sẽ chia làm 2 phần nhỏ: **phần 1** là $\frac{L_c}{L}$ và **phần 2** là $\left(\frac{k_c}{2L}\right)^2$
- Phần 1 cho ta biết được tỉ lệ các liên kết kết nối các cặp node thuộc community c.
- Còn phần 2 cho ta biết được tỉ lệ các liên kết mong đợi sẽ tìm thấy bên trong community đó nếu các liên kết được đặt ngẫu nhiên trong .
- Theo cách hiểu này thì nếu phần 1 vượt quá phần 2 thì tập con c trong network đúng là 1 community, vì nó có nhiều liên kết hơn kì vọng bởi sự ngẫu nhiên. Nếu sự chênh lệch này càng lớn thì community sẽ càng được xác định tốt hơn.
- Cũng vì lí do đó nên ta có thể xác nhận rằng với một subgraph S có L_c liên kết nội bộ và tổng bậc k_c là một community nếu:

$$\frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2 > 0 \ (*)$$

- Nếu tất cả các subsets c của partition là các communities, theo ý nghĩa ở (*) thì Modularity của partition là dương (M > 0).
- Mặt khác Modularity là một hàm có giới hạn. Vì mỗi phần tử tổng không thể lớn hơn giá trị của Phần 1, ta có:

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

Vậy: giá trị lớn nhất của modularity M không thể vượt quá 1.

5. Tài liệu tham khảo:

[1]:Chapter 9: Communities: *http://networksciencebook.com/chapter/9#basics*

[2]: Emergence of Scaling in Random Networks – Barabasi: *https://barabasi.com/f/67.pdf*

[3]: Resolution limit in community detection: *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1765466/*

[4]: Community detection in networks: A user guide: http://www.cs.cornell.edu/courses/cs6241/2019sp/readings/Fortunato-2016-guide.pdf

[5]: Modularity and community structure in networks: https://www.pnas.org/doi/10.1073/pnas.0601602103

[6]: On Modularity - NP-Completeness and Beyond: https://publikationen.bibliothek.kit.edu/1000005777/3255

[7]: Modularity (networks): https://en.wikipedia.org/wiki/Modularity_(networks)

[8]: Limits of modularity maximization in community detection: https://www.academia.edu/18321200/Limits_of_modularity_maximization_in_community_detection

[9]: Resolution limit in community detection:

https://www.pnas.org/doi/epdf/10.1073/pnas.0605965104

[10]: Hierarchical network model: https://en.wikipedia.org/wiki/Hierarchical_network_model?fbclid=IwAR3FSeahhqajo_bWW8B7ZKljeX-OOUwPtg4EXKowklmAaWJ0DkTGSXSPDUk

[11]: Quality functions in community detection: https://www.researchgate.net/publication/1892349_Quality_functions_in_community_detection/link/540c9d4b0cf2d8daaacae9c4/download

[12]: A New Metric for Quality of Network Community Structure: https://www.researchgate.net/publication/273062104_A_New_Metric_for_Quality_of_Network_Community_Structure/link/54f62cbd0cf27d8ed71d66ae/download

Hết