TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG-HCM

HOMEWORK 01

Học kỳ 3 – Năm học 2022-2023

MÃ LƯU TRỮ
(do phòng KT-ĐBCL ghi)

Tên học phần:	**Khai thác Dữ liệu Đồ thị**	Mã HP:	**CSC17103**
Thời gian làm bài:	**07 ngày**	Ngày nộp:	**19/06/2023**

Gợi ý: Các bạn sẽ cần đọc các tài liệu tham khảo để làm được bài tập này.

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

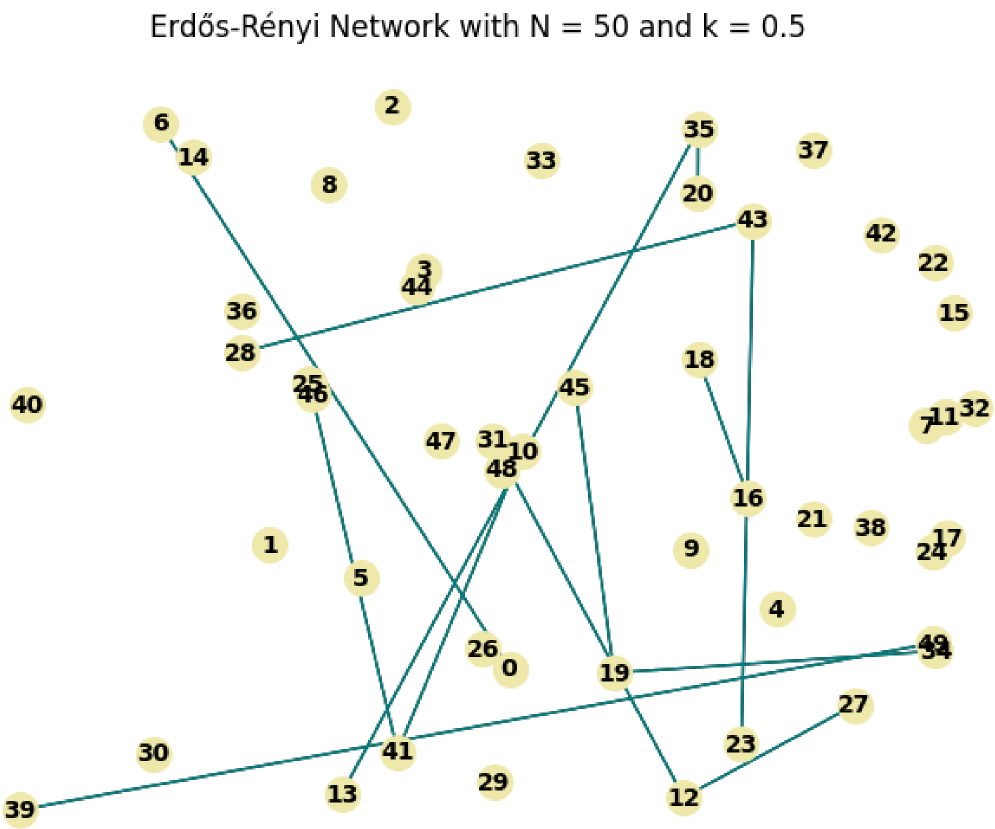
BÀI TRÌNH BÀY

1. Phát sinh mạng Erdős-Rényi:

Hãy trực quan hoá mạng Erdős-Rényi với N = 50 nút và bậc trung bình $\langle k\rangle$ lần lượt là:

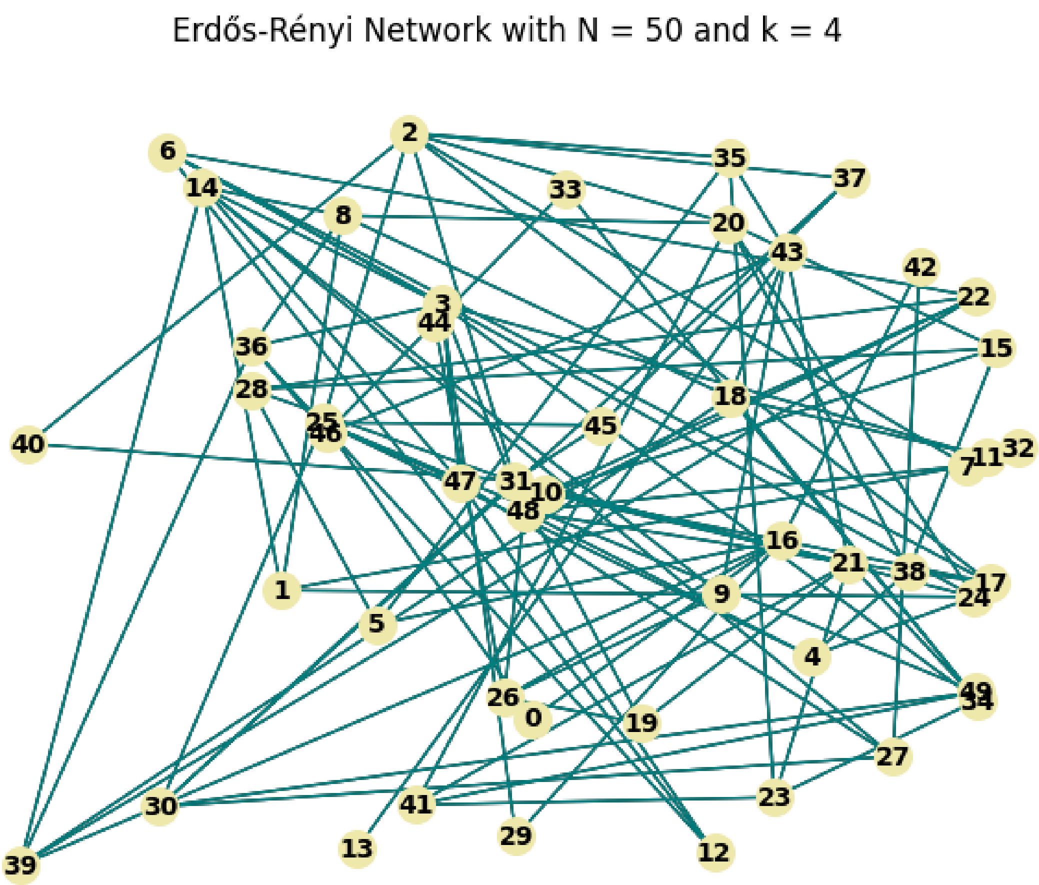
a. $\langle k\rangle$ = 0.5 (Random layout)

```
In [110]: N = 50 # Tổng số Node
k = 0.5 # Bậc trung bình
p = k /(N - 1) # Xác suất có cạnh xuất hiện giữa 2 Node
G = nx.erdos_renyi_graph(N, p) # Tạo ra mạng Erdős-Rényi
pos = nx.random_layout(G, seed=50) # Layout trực quan
nx.draw(G, pos, with_labels=True, node_color='#EEE8AA', node_size=200,
        font_size = 10, font_weight = "bold") # Thay đổi màu sắc, kích cỡ
nx.draw_networkx_edges(G, pos, edge_color='#008080')
plt.title("Erdős-Rényi Network with N = 50 and k = 0.5") # Tên cho đồ thị
plt.show()
```



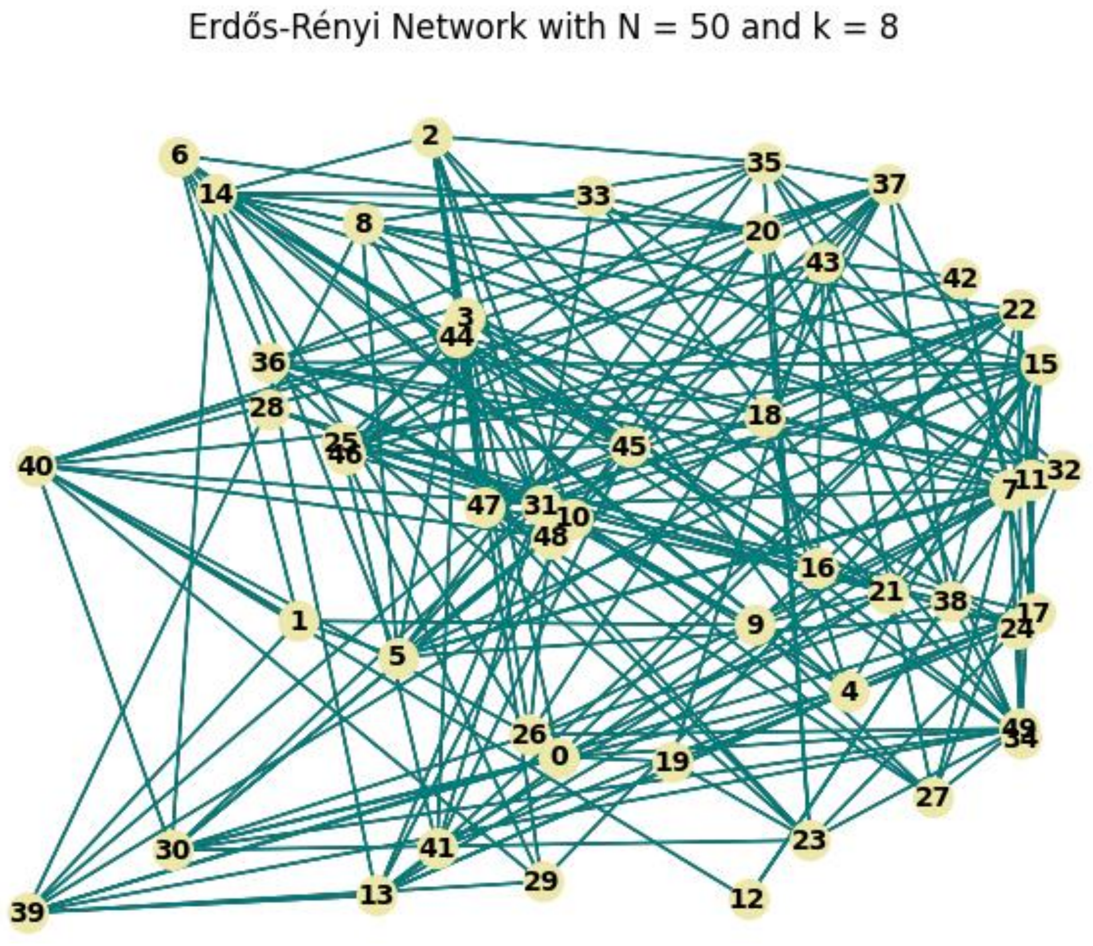
b. ⟨k⟩ = 4 (Random layout)

In [111]:
```python
N = 50 # Tổng số Node
k = 4 # Bậc trung bình
p = k /(N - 1) # Xác suất có cạnh xuất hiện giữa 2 Node
G = nx.erdos_renyi_graph(N, p) # Tạo ra mạng Erdős-Rényi
pos = nx.random_layout(G, seed=50) # Layout trực quan
nx.draw(G, pos, with_labels=True, node_color='#EEE8AA', node_size=200,
        font_size = 10, font_weight = "bold")
nx.draw_networkx_edges(G, pos, edge_color='#008080')
plt.title("Erdős-Rényi Network with N = 50 and k = 4")
plt.show()
```

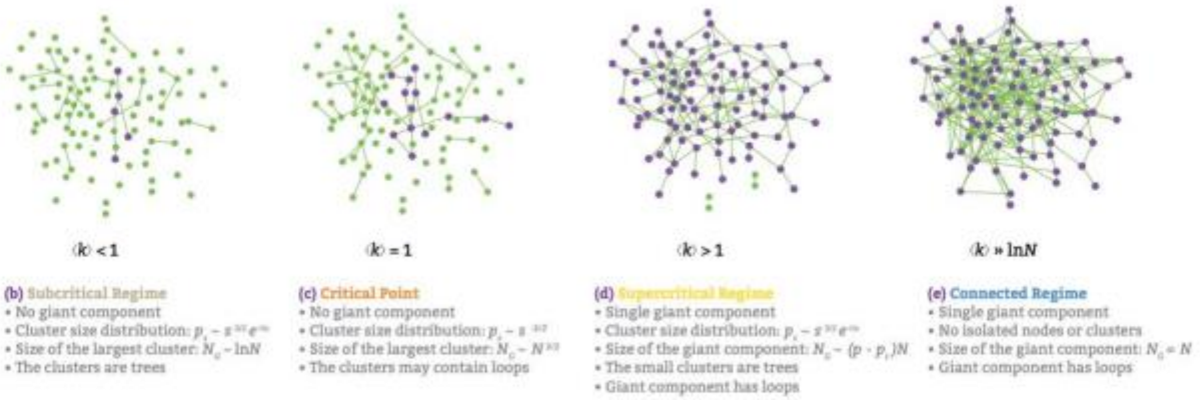


c. ⟨k⟩ = 8 (Random layout)

```
In [112]: N = 50 # Tổng số Node
          k = 8 # Bậc trung bình
          p = k /(N - 1) # Xác suất có cạnh xuất hiện giữa 2 Node
          G = nx.erdos_renyi_graph(N, p) # Tạo ra mạng Erdős-Rényi
          pos = nx.random_layout(G, seed=50) # Layout trực quan
          nx.draw(G, pos, with_labels=True, node_color='#EEE8AA', node_size=200,
                 font_size = 10, font_weight = "bold")
          nx.draw_networkx_edges(G, pos, edge_color='#008080')
          plt.title("Erdős-Rényi Network with N = 50 and k = 8")
          plt.show()
```



2. Mạng Erdős-Rényi:



Xem xét mạng G(N, p) có N = 3000 nút và được kết nối với nhau với xác suất $p = 10^{-3}$
Hãy trả lời các câu hỏi dưới đây.

a. Xác định số lượng liên kết kỳ vọng ⟨L⟩ và bậc trung bình ⟨k⟩ của mạng.

Ta có công thức số lượng liên kết kỳ vọng ⟨L⟩ sau : $\langle L\rangle = p.\frac{N(N-1)}{2}$

Nên: $\langle L\rangle = 10^{-3}.\frac{[3000.(3000-1)]}{2} = \boldsymbol{4498.5}$

Ta cũng có công thức tính bậc trung bình ⟨k⟩ của mạng là $\langle k\rangle = \frac{2L}{N} = p\ .\ (N-1)$

Nên: $\langle k\rangle = 10^{-3}\ .\ (3000 - 1) = \boldsymbol{2.999}$

b. Xác suất có chính xác 50 liên kết trong mạng là bao nhiêu ?

Theo tài liệu tham khảo được cung cấp, ta có công thức:

$$p_L = \binom{\frac{N(N-1)}{2}}{L} . p^L . (1-p)^{0.5\,.N(N-1)-L}$$

Nên: $p_{50} = C^{50}_{0.5(3000-1).3000} . (10^{-3})^{50} . (1- 10^{-3})^{0.5(3000-1).3000\ -\ 50}$

c. Dựa vào hình 1, xác định xem mạng ở chế độ (regime) nào ?

- Với ⟨k⟩ = 2.999 nên mạng sẽ ở chế độ ***Supercritical regime***.

d. Tính xác suất p_c để mạng ở chế độ critical point.

- Để mạng ở chế độ critical point thì ⟨k⟩ = 1
- Lúc đó: $p_c = \frac{1}{N} = \frac{1}{3000}$

e. *Tính số nút N^{cr}, bậc trung bình $\langle k^{cr}\rangle$ và khoảng cách trung bình giữa hai nút được chọn ngẫu nhiên $\langle d\rangle$ để mạng chỉ có một thành phần.*

- Theo công thức thì: $\langle k\rangle = p\ .\ (N^{cr}-1)$, mà để mạng chỉ có một thành phần thì:

  $\langle k\rangle >> \ln(N^{cr})$ (Connected Regime)

  Với: $\langle k\rangle >> \ln(N^{cr})$ (1)

  → $p.(N^{cr}-1) >> \ln(N^{cr})$

  → $N^{cr}-1 >> \ln(N^{cr})\ .\ 1000$

  → $N^{cr} >> 1+ \ln(N^{cr})\ .\ 1000$

  Gọi $N^{cr} = x$, xét phương trình $y_1 = x$ và $y_2 = \ln(x).1000 +1$có đồ thị sau:

- Xét trong đoạn từ 1 đến 10000, bước nhảy là 1000, ta có bảng sau:

y_1	y_2
1	1
1000	6908.75
2000	7601.90
3000	8007.37
4000	8295.05
5000	8518.19
6000	8700.51
7000	8854.67
8000	8988.19
9000	9105.98
10000	**9211.34**

Ta thấy: ở giá trị x = 10000 thì $y_1 > y_2$

- Xét kĩ hơn đoạn 9000 đến 9500 với bước nhảy là 100 có:

y_1	y_2
9000	9105.98
9100	9117.03
9200	**9127.95**
9300	9138.77
9400	9149.46
9500	9160.04

Ta thấy: ở giá trị x = 9200 thì $y_1 > y_2$

- Xét kĩ hơn đoạn 9100 đến 9150 với bước nhảy là 10 có:

y_1	y_2
9100	9117.03
9110	9118.12
9120	**9119.22**
9130	9120.32
9140	9121.41
9150	9122.51

Ta thấy: ở giá trị x = 9120 thì $y_1 > y_2$

- Xét kĩ hơn đoạn 9115 đến 9120 với bước nhảy là 1 có:

y_1	y_2
9115	9118.68
9116	9118.79
9117	9118.89
9118	9119.01

9119	9119.12
9120	**9119.22**

Vậy với $\mathbf{N^{cr}}$ > **9120** thì mạng chỉ có một thành phần.

Lúc đó: $\langle k^{cr} \rangle = p \,.\, (N^{cr} - 1) \gg 10^{-3} \,.\, (9120 - 1) \gg \mathbf{9.119}$

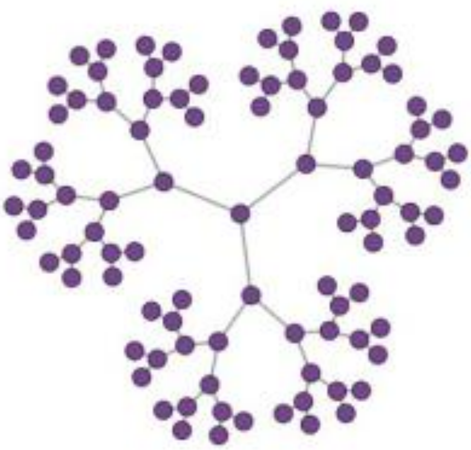
- Khoảng cách trung bình giữa hai nút được chọn ngẫu nhiên $\langle d \rangle$ để mạng chỉ có một thành phần là: $\langle d \rangle \approx \frac{\ln N}{\ln \langle k \rangle}$.

f. Tìm phân bố bậc p_k của mạng này (xấp xỉ với phân bố bậc Poisson).

- Theo lí thuyết ta biết rằng trong mạng Erdős-Rényi, phân phối bậc tuân theo phân phối nhị thức: $p_k = \binom{N-1}{k} \,.\, p^k \,.\, (1-p)^{N-1-k}$, không phải phân phối Poisson.
- Nhưng bằng cách tính xấp xỉ với phân bố bậc Poisson ta có thể tính theo công thức sau: $p_k = e^{-\langle k \rangle} . \frac{\langle k \rangle^k}{k!}$

3. <u>Cây Cayley (Cayley tree):</u>

Cây Cayley là cây đối xứng, được xây dựng bắt đầu từ nút trung tâm bậc k. Mỗi nút ở khoảng cách d tính từ nút trung tâm có bậc k, cho đến khi chúng ta đến các nút ở khoảng cách P có bậc một và được gọi là các lá. Ví dụ, hình 2 là cây Cayley có k = 3 và P = 5.



a. Tính tổng số nút trên cây sau t bước tính từ từ nút trung tâm.

- Như ta có thể thấy ở hình minh họa thì chỉ sau với k = 3, P = 1 thì số node mới được tạo ra là 3. Nhưng sau đó, nếu P > 1 thì chỉ có (k−1) liên kết tách ra để tạo các node mới, 1 liên kết sẽ luôn nối trở lại gốc và sau bước đầu tiên, ta sẽ mở rộng (t−1) lần. nên tổng số node trên cây sau t bước tính từ từ nút trung tâm là:

$$k.\left[\frac{(k-1)^t-1}{(k-1)-1}\right] + 1\text{(node trung tâm)}$$

- Ví dụ với cây Cayley có k = 3 và P = 5 như đề cho thì tổng số nút trên cây sau 5 bước tính từ nút trung tâm là:

$$k.\left[\frac{(k-1)^t-1}{(k-1)-1}\right] + 1 = 3.\left[\frac{(3-1)^5-1}{(3-1)-1}\right] + 1 = 94 \text{ (node)}$$

b. Tính độ phân phối bậc (degree distribution) của mạng.

- Phía bên ngoài cùng có: $k.(k-1)^{(t-1)}$ đỉnh bậc 1

→ Phân phối bậc ở các đỉnh này là: $\frac{\mathbf{k.(k-1)^\wedge(t-1)}}{\textbf{Cayley size}}$

- Còn lại có: $k.\left[\frac{(k-1)^t-1}{(k-1)-1}\right] + 1 - k.(k-1)^{(t-1)}$ đỉnh có bậc k

→ Phân phối bậc ở các đỉnh này là: $\frac{\mathbf{k.\left[\frac{(k-1)^t-1}{(k-1)-1}\right] + 1 - k.(k-1)^\wedge(t-1)}}{\textbf{Cayley size}}$

- Ví dụ như hình minh họa cây Cayley k = 3 và P = 5, sau t bước có:
- $3.(3-1)^{(5-1)} = 48$ đỉnh bậc 1

→ Degree distribution = $\frac{48}{94} = \mathbf{0.511}$

- (94 – 48) = 46 đỉnh bậc 3.

→ Degree distribution = $\frac{46}{94} = \mathbf{0.489}$

c. Tính đường kính d_{max}.

Ta có: Đường kính $d_{max} = 2P$

- Ví dụ như hình minh họa cây Cayley k = 3 và P = 5 thì đường kính $d_{max} = 2 . 5 = 10$.

d. Tìm biểu thức của đường kính d_{max} theo tổng số nút N.

- Ở khoảng cách d ta có: $N(d) = k.\left[\frac{(k-1)^d - 1}{(k-1)-1}\right] + 1$ (nodes)
- Đối với độ sâu P, ta có $0.5 \,.\, d_{max} = P$ nên:

$$N = k.\left[\frac{(k-1)^{0.5.dmax} - 1}{(k-1)-1}\right] + 1$$

→ $$N - 1 = \left(\frac{k}{k-2}\right).\left[(k-1)^{0.5.dmax} - 1\right]$$

→ $$1 + \left[\frac{(N-1).\,(k-2)}{k}\right] = (k-1)^{0.5dmax}$$

→ $$\log\left[1 + \frac{(N-1).\,(k-2)}{k}\right] = \log_{k-1} 0.5\,dmax \qquad \text{(Lấy logarit 2 vế)}$$

→ $$d_{max} = \left[\frac{2}{\log(k-1)}\right]\log\left[1 + \frac{(N-1).\,(k-2)}{k}\right] \approx \left[\frac{2}{\log(k-1)}\right]\log\left[1 + \frac{N.\,(k-2)}{k}\right]$$

→ $$d_{max} \approx \frac{2\log(N)}{\log(k-1)}$$

- Vậy biểu thức của đường kính d_{max} theo tổng số node N là $d_{max} \approx \frac{2\log(N)}{\log(k-1)}$

4. Nghịch lý tình bạn (Friendship Paradox):

Phân phối bậc p_k là xác suất mà một nút được chọn ngẫu nhiên có k hàng xóm. Tuy nhiên, nếu chúng ta chọn ngẫu nhiên một liên kết, xác suất để một nút ở một trong các đầu của nó có bậc k là $q_k = Akp_k$, trong đó A là hệ số chuẩn hóa.

a. Tìm hệ số chuẩn hóa A, giả sử rằng mạng có phân bố bậc theo luật mũ với $2 < \gamma < 3$, với bậc nhỏ nhất k_{min} và bậc lớn nhất k_{max}.

- Ta có bậc của nút là các số nguyên dương, k = 0, 1, 2, ..., nên hình thức rời rạc cung cấp xác suất p_k mà một nút có chính xác k liên kết là: $p_k = C.k^{-\gamma}$
- Với A là hệ số chuẩn hóa thì $\sum_{kmin}^{kmax} q_k = 1$ (theo công thức phân phối) hay:

$$\sum_{kmin}^{kmax} Akp_k = 1$$

$$\rightarrow \sum_{kmin}^{kmax} A.k.C.k^{-\gamma} = 1$$

$$\rightarrow \sum_{kmin}^{kmax} A.C.k^{1-\gamma} = 1$$

$$\rightarrow A.C.\int_{kmin}^{kmax} k^{1-\gamma}\,dk = 1 \ (*)$$

- Mà công thức 4.11 của Network Science by Albert-László Barabási thì:

$$C = \frac{1}{\int_{kmin}^{kmax} k^{-\gamma} dk} = (\gamma - 1)k_{min}^{\gamma-1} \quad (**)$$

- Thế (**) và (*) thì:

$$A \,.\, (\gamma - 1)k_{min}^{\gamma-1} \,.\, \int_{kmin}^{kmax} k^{1-\gamma}\, dk = 1$$

$$\rightarrow \mathbf{A} = \frac{\mathbf{1}}{(\boldsymbol{\gamma}-\mathbf{1})\mathbf{k}_{\mathbf{min}}^{\boldsymbol{\gamma}-\mathbf{1}} \,.\int_{\mathbf{kmin}}^{\mathbf{kmax}} \mathbf{k}^{\mathbf{1}-\boldsymbol{\gamma}}\, \mathbf{dk}}$$

b. *Chọn ngẫu nhiên một nút trong mạng có N = 104, γ= 2.3, k_{min}= 1 và k_{max}= 1000. Tính bậc trung bình của các nút lân cận.*

- Theo câu a, ta biết được: $A = \frac{1}{(\gamma-1)\,.\,k_{min}^{\gamma-1}\,.\,\int_{kmin}^{kmax} k^{1-\gamma}\, dk} = \frac{1}{(2.3-1)\,.1^{2.3-1}\,.\,\int_{1}^{1000} k^{1-2.3}\, dk}$

  $= 0.264$

- Theo công thức 4.20 của Network Science by Albert-László Barabási thì:

$$\langle k \rangle = \int_{kmax}^{kmin} k \,.\, q_k\, dk$$

- Mà $q_k = Akp_k$ nên $\langle k \rangle = \int_{kmin}^{kmax} k.\,Akp_k dk$

$$= \int_{kmin}^{kmax} k \,.\, \frac{1}{(\gamma-1)k_{min}^{\gamma-1}\,.\int_{kmin}^{kmax} k^{1-\gamma}\, dk} \,.\, k \,.\, p_k\, dk$$

$$= \int_{kmin}^{kmax} k \,.\, \frac{1}{(\gamma-1)k_{min}^{\gamma-1}\,.\int_{kmin}^{kmax} k^{1-\gamma}\, dk} \,.\, k \,.\, C \,.\, k^{-\gamma}\, dk$$

$$= \int_{kmin}^{kmax} k \,.\frac{1}{(\gamma-1)k_{min}^{\gamma-1}\,.\int_{kmin}^{kmax} k^{1-\gamma}\, dk} \,.\, k \,.\, (\gamma - 1)k_{min}^{\gamma-1} \,.\, k^{-\gamma} dk$$

$$= \int_{1}^{1000} k \,.\, \frac{1}{(2.3-1)1^{2.3-1}\,.\int_{1}^{1000} k^{1-2.3}\, dk} \,.\, k \,.\, (2.3 - 1).\, 1^{2.3-1} \,.\, k^{-\,2.3}\, dk$$

$$= \int_{1}^{1000} k \,.\, \frac{1}{(1.3).1^{1.3}\,.\int_{1}^{1000} k^{1.3}\, dk} \,.\, k \,.\, (1.3).\, 1^{1.3} \,.\, k^{-\,2.3}\, dk$$

$= 61.23431$

Vậy bậc trung bình của các nút lân cận là **61,234**.

5. Tài liệu tham khảo:

[1]: Erdős-Rényi Network: http://networksciencebook.com/chapter/3#random-network

[2]: Cayley's Tree Theorem: https://slwu89.github.io/src/network_science_notes.html

[3]: Cayley's Tree Theorem: https://www.youtube.com/watch?v=Wi8IvnlMNxs

[4]: Cayley's Tree Theorem: http://networksciencebook.com/chapter/3#random-network

[5]: Friendship Paradox theory: http://networksciencebook.com/chapter/4

[6]: Friendship Paradox theory: https://qubeshub.org/resources/740/download/ModuleFPQ.pdf

[7]: Friendship Paradox theory: https://en.wikipedia.org/wiki/Friendship_paradox

[8]: Friendship Paradox theory: https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0190-8

[9]: Friendship Paradox theory: https://www.youtube.com/watch?v=GEjhO65FYks

Hết