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

HOMEWORK 01

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

Erdos-Renyi network với N = 3000 đỉnh, được liên kết với nhau với xác suất $p = 10^{-3}$.

Câu a

Câu hỏi: Số lượng liên kết trung bình (kỳ vọng số lượng liên kết) $\langle L \rangle$ =?

Trả lời:

Ta có công thức tính số lượng liên kết trung bình của Erdos-Renyi network:

$$\langle L \rangle = p \cdot \binom{N}{2} = 10^{-3} \cdot \binom{3000}{2} = 4498.5$$

Vậy số lượng liên kết trung bình của Erdos-Renyi network với N = 3000 và p = 10^{-3} là 4498.5.

Câu b

Câu hỏi: Mạng này đang nằm trong trạng thái nào (regime)?

Trả lời: Các trạng thái của mạng Erdos-Renyi network (topological regimes) bao gồm:

- Subcritical regime (trạng thái dưới ngưỡng): Trạng thái này xảy ra khi xác suất liên kết p nhỏ hơn nhiều so với ngưỡng $\frac{1}{N}$. Trong trạng thái này, mạng chủ yếu bao gồm các thành phần nhỏ và rời rạc.

$$\langle k \rangle < 1 \Rightarrow p < \frac{1}{N}$$

- Critical regime (Trạng thái tới hạn): Trạng thái này xảy ra khi xác suất liên kết p gần bằng với ngưỡng $\frac{1}{N}$. Đây là thời điểm quan trọng khi các thành phần nhỏ bắt đầu kết hợp để tạo ra thành phần khổng lồ.

$$\langle k \rangle = 1 \Rightarrow p = \frac{1}{N}$$

- Supercritical regime (Trạng thái trên ngưỡng): Trạng thái này xảy ra khi xác suất liên kết p lớn hơn so với ngưỡng $\frac{1}{N}$. mạng bắt đầu hình thành một thành phần khổng lồ, chiếm một phần đáng kể số đỉnh trong mạng.

$$\langle k \rangle > 1 \Rightarrow p > \frac{1}{N}$$

- Connected regime (Trạng thái liên thông): Trạng thái này xảy ra khi xác suất liên kết p đủ lớn để đảm bảo rằng hầu hết các đỉnh đều thuộc về một thành phần liên thông duy nhất. Mạng ở trạng thái này thường có một thành phần khổng lồ bao phủ toàn bộ mạng

$$\langle k \rangle > ln(N) \Rightarrow p > \frac{ln(N)}{N}$$

Với $N = 3000$ và $p = 10^{-3}$, ta có:

$$\langle k \rangle = p \cdot (N-1) \approx p \cdot N \rightarrow ln(N) \approx 8 > k = 3 > 1$$

nên mạng này đang ở trạng thái supercritical regime (trạng thái trên ngưỡng).

Câu c

Câu hỏi: Tính xác xuất p_c mà mạng đang ở thời điểm quan trọng (the critical point).

Trả lời:

Xác suất p_c mà mạng đang ở trạng thái quan trọng (critical point) là khi $\langle k \rangle = 1$.

$$\langle k \rangle = p_c \cdot (N-1) = 1 \Rightarrow p_c = \frac{1}{N-1} = \frac{1}{3000-1} \approx 3.33 \times 10^{-4}$$

Vậy xác suất p_c mà mạng đang ở trạng thái quan trọng (critical point) là khoảng 3.33×10^{-4}.

Câu d

Câu hỏi: Cho trước xác suất liên kết $p = 10^{-3}$, tính số lượng đỉnh N^{cr} mà mạng này chỉ có duy nhất một thành phần

Trả lời:

Để mạng chỉ có duy nhất một thành phần liên thông, ta cần xác suất liên kết p đủ lớn để đảm bảo rằng hầu hết các đỉnh đều thuộc về một thành phần liên thông duy nhất (trạng thái connected regime)

$$\langle k \rangle = p \cdot (N-1) > ln(N) \Rightarrow p > \frac{ln(N)}{N-1}$$

Với $p = 10^{-3}$, ta có:

$$p > \frac{ln(N)}{N-1} \Rightarrow 10^{-3} > \frac{ln(N)}{N-1} \Rightarrow N - 1 > \frac{ln(N)}{10^{-3}}$$

$$\Rightarrow N > 1 + \frac{ln(N)}{10^{-3}} \Rightarrow N > 1 + 1000 \cdot ln(N)$$

Để giải phương trình trên, ta sử dụng phương pháp vét cạn (brute-force) để tìm giá trị N thỏa mãn phương trình trên.

```python
import math

# Khởi tạo giá trị N
N = 1

# Tăng dần giá trị N cho đến khi N > 1000 * ln(N) + 1
while True:
    if N > 1000 * math.log(N) + 1:
        break
    N += 1

print(N)
```

Kết quả khi chạy đoạn code trên là $N = 9120$.

Vậy số lượng đỉnh N^{cr} mà mạng này chỉ có duy nhất một thành phần liên thông là 9120.

Câu e

Câu hỏi: Với mạng trong câu (d), tính toán bậc trung bình $\langle K^{cr} \rangle$ và khoảng cách trung bình giữa hai đỉnh ngẫu nhiên bất kỳ $\langle d \rangle$

Trả lời: Bậc trung bình $\langle K^{cr} \rangle$ của mạng khi chỉ có duy nhất một thành phần liên thông được tính bằng công thức:

$$\langle K^{cr} \rangle = p \cdot (N-1) = 10^{-3} \cdot (9120 - 1) = 9.119$$

Khoảng cách trung bình giữa hai đỉnh ngẫu nhiên bất kỳ $\langle d \rangle$ của mạng khi chỉ có duy nhất một thành phần liên thông được tính bằng công thức:

$$\langle d \rangle = \frac{ln(N)}{ln(\langle K^{cr} \rangle)} = \frac{ln(9120)}{ln(9.119)} \approx 4.125$$

Vậy bậc trung bình $\langle K^{cr} \rangle$ của mạng khi chỉ có duy nhất một thành phần liên thông là 9.119 và khoảng cách trung bình giữa hai đỉnh ngẫu nhiên bất kỳ $\langle d \rangle$ là khoảng 4.125.

Câu f

Câu hỏi: Tính toán phân phối bậc p_k của mạng này (xấp xỉ với một phân phối bậc Poission)

Trả lời:

Phân phối bậc p_k của mạng Erdos-Renyi network được xấp xỉ với một phân phối bậc Poission với $\langle k \rangle = p \cdot (N-1)$.

Để tính phân phối bậc p_k của mạng này, ta sử dụng công thức phân phối Poission:

$$p_k = \frac{e^{-\langle k \rangle} \cdot \langle k \rangle^k}{k!}$$

Trong đó, $\langle k \rangle = p \cdot (N-1) = 10^{-3} \cdot (3000-1) = 2.999$.

Vậy phân phối bậc p_k của mạng này là:

$$p_k = \frac{e^{-2.999} \cdot 2.999^k}{k!}$$

2 Problem 2

Dựa trên mô hình G(N,p), hãy phát sinh ba mạng với N = 500 đỉnh, và trung bình bậc

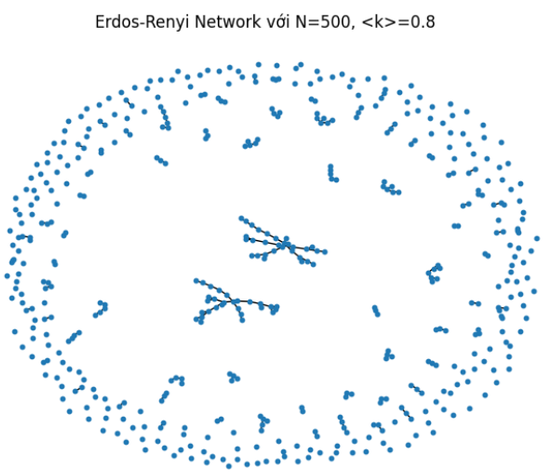
- $\langle k \rangle = 0.8$
- $\langle k \rangle = 1$
- $\langle k \rangle = 8$

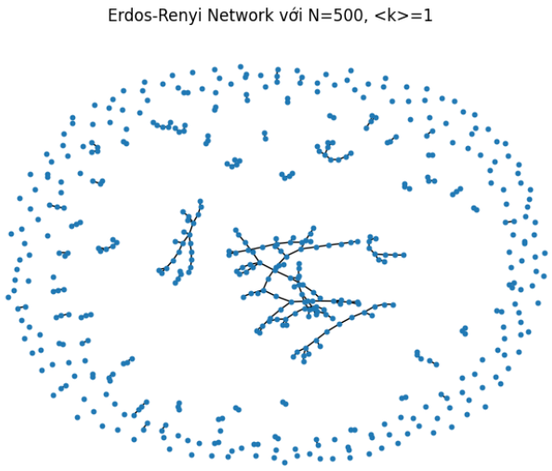
Với lần lượt mỗi trung bình bậc $\langle k \rangle$, em sẽ tính xác suất liên kết p tương ứng với mỗi trung bình bậc đó bằng công thức $p = \frac{\langle k \rangle}{N-1}$ và dùng NetworkX để vẽ mạng.

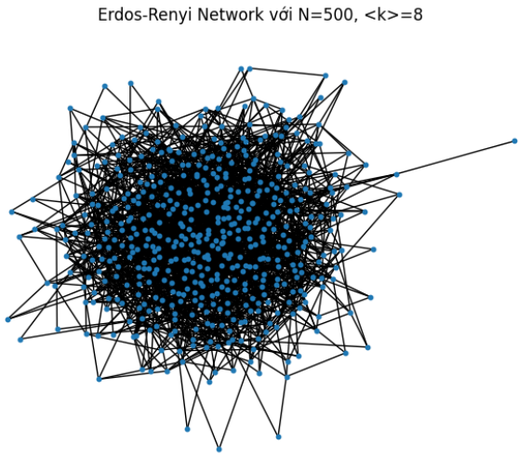
```
import networkx as nx
import matplotlib.pyplot as plt
```

```python

# Hàm vẽ mạng Erdos-Renyi với trung bình bậc k
def draw_er_graph(N, k):
    p = k / (N - 1)
    G = nx.erdos_renyi_graph(N, p)
    nx.draw(G, node_size=10)
    plt.title(f'Erdos-Renyi Network với N={N}, <k>={k}')
    plt.show()

# Số lượng đỉnh
N = 500

# Trung bình bậc k
k_values = [0.8, 1, 8]

# Vẽ mạng với các trung bình bậc khác nhau
for k in k_values:
    draw_er_graph(N, k)
```

Và đây là hình ảnh của ba mạng Erdos-Renyi network với N = 500 và trung bình bậc $\langle k \rangle = 0.8, 1, 8$:







3 Problem 3

(Nghịch lý Tình bạn). Phân phối bậc p_k biểu diễn xác suất mà một đỉnh được chọn ngẫu nhiên có k hàng xóm. Tuy nhiên, nếu ta chọn ngẫu nhiên một liên kết, xác suất một đỉnh mà đỉnh cuối của nó có bậc là k là $q_k = Akp_k$, trong đó A là một nhân tử chuẩn hóa (normalization factor)

Câu a

Câu hỏi: Tìm nhân tử chuẩn hóa A. Giả định rằng, mạng có phân phối bậc luật lũy thừa với $2 < \gamma < 3$, với bậc nhỏ nhất k_{min} và bậc lớn nhất k_{max}

Trả lời

- Xác suất đỉnh cuối của một liên kết có bậc k là $q_k = Akp_k$ với bậc nhỏ nhất k_{min} và bậc lớn nhất k_{max}. Tổng xác suất của tất cả các bậc từ k_{min} đến k_{max} bằng 1 nên ta có:

$$\sum_{k=k_{min}}^{k_{max}} q_k = 1 \rightarrow \sum_{k=k_{min}}^{k_{max}} Akp_k = 1$$

  Nhân tử chuẩn hóa A được tính bằng công thức:

$$A = \frac{1}{\sum_{k=k_{min}}^{k_{max}} kp_k}$$

- Mạng có phân phối bậc luật lũy thừa với $2 < \gamma < 3$ nên $p_k = C \cdot k^{-\gamma}$ với C là hằng số chuẩn hóa. Tổng xác suất của tất cả các bậc từ k_{min} đến k_{max} bằng 1 nên ta lại có:

$$\sum_{k=k_{min}}^{k_{max}} C \cdot k^{-\gamma} = 1 \rightarrow C \cdot \sum_{k=k_{min}}^{k_{max}} k^{-\gamma} = 1$$

  Tính hằng số chuẩn hóa C:

$$C = \frac{1}{\sum_{k=k_{min}}^{k_{max}} k^{-\gamma}} \approx \frac{1}{\int_{k_{min}}^{k_{max}} k^{-\gamma} dk} = \frac{1}{\left[\frac{k^{1-\gamma}}{1-\gamma}\right]_{k_{min}}^{k_{max}}} = \frac{1}{\frac{k_{max}^{1-\gamma} - k_{min}^{1-\gamma}}{1-\gamma}} = \frac{1-\gamma}{k_{max}^{1-\gamma} - k_{min}^{1-\gamma}}$$

- Thế $p_k = C \cdot k^{-\gamma}$ vào công thức tính nhân tử chuẩn hóa A:

$$A = \frac{1}{\sum_{k=k_{min}}^{k_{max}} kp_k} = \frac{1}{\sum_{k=k_{min}}^{k_{max}} k \cdot C \cdot k^{-\gamma}} = \frac{1}{C \cdot \sum_{k=k_{min}}^{k_{max}} k^{1-\gamma}}$$

$$\rightarrow A \approx \frac{1}{C \cdot \int_{k_{min}}^{k_{max}} k^{1-\gamma} dk} = \frac{1}{C \cdot \left[\frac{k^{2-\gamma}}{2-\gamma}\right]_{k_{min}}^{k_{max}}} = \frac{1}{C \cdot \frac{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}}{2-\gamma}}$$

  Thế C vào công thức trên:

$$\rightarrow A = \frac{1}{C} \cdot \frac{2-\gamma}{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}} = \frac{k_{max}^{1-\gamma} - k_{min}^{1-\gamma}}{1-\gamma} \cdot \frac{2-\gamma}{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}}$$

Vậy nhân tử chuẩn hóa A là:

$$A = \frac{k_{max}^{1-\gamma} - k_{min}^{1-\gamma}}{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}} \cdot \frac{2-\gamma}{1-\gamma}$$

Câu b

Đề bài: Trong mô hình thiết lập, q_k cũng là xác suất mà một nút được chọn ngẫu nhiên có một hàng xóm với bậc k. Vậy, bậc trung bình của các hàng xóm của một nút được chọn ngẫu nhiên là bao nhiêu?

Trả lời

Bậc trung bình của các hàng xóm của một nút được chọn ngẫu nhiên được tính bằng công thức:

$$\langle k_{nn} \rangle = \sum_{k=k_{min}}^{k_{max}} k \cdot q_k$$

Thay $q_k = Akp_k$ vào công thức trên:

$$\langle k_{nn} \rangle = \sum_{k=k_{min}}^{k_{max}} k \cdot Akp_k = A \cdot \sum_{k=k_{min}}^{k_{max}} k^2 \cdot p_k$$

$$\langle k_{nn} \rangle = A \cdot \sum_{k=k_{min}}^{k_{max}} k^2 \cdot C \cdot k^{-\gamma} = A \cdot C \cdot \sum_{k=k_{min}}^{k_{max}} k^{2-\gamma}$$

$$\rightarrow \langle k_{nn} \rangle = A \cdot C \cdot \int_{k_{min}}^{k_{max}} k^{2-\gamma} dk = A \cdot C \cdot \left[\frac{k^{3-\gamma}}{3-\gamma} \right]_{k_{min}}^{k_{max}} = A \cdot C \cdot \frac{k_{max}^{3-\gamma} - k_{min}^{3-\gamma}}{3-\gamma}$$

Thay C và A vào công thức trên:

$$\langle k_{nn} \rangle = \frac{k_{max}^{1-\gamma} - k_{min}^{k_max}}{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}} \cdot \frac{2-\gamma}{1-\gamma} \cdot \frac{1-\gamma}{k_{max}^{1-\gamma} - k_{min}^{1-\gamma}} \cdot \frac{k_{max}^{3-\gamma} - k_{min}^{3-\gamma}}{3-\gamma}$$

Vậy bậc trung bình của các hàng xóm của một nút được chọn ngẫu nhiên là:

$$\rightarrow \langle k_{nn} \rangle = \frac{k_{max}^{3-\gamma} - k_{min}^{3-\gamma}}{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}} \cdot \frac{2-\gamma}{3-\gamma}$$

Câu c

Câu hỏi: Tính toán bậc trung bình của các hàng xóm của một nút được chọn ngẫu nhiên trong một mạng với $N = 10^4$, $\gamma = 2.3$, $k_{min} = 1$ và $k_{max} = 1000$. So sánh kết quả với bậc trung bình của mạng $\langle k \rangle$

Trả lời

- Như công thức đã tính ở câu b, bậc trung bình của hàng xóm của một nút chọn ngẫu nhiên trong mạng được tính bằng công thức:

$$\langle k_{nn} \rangle = \frac{k_{max}^{3-\gamma} - k_{min}^{3-\gamma}}{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}} \cdot \frac{2-\gamma}{3-\gamma} = \frac{1000^{3-2.3} - 1^{3-2.3}}{1000^{2-2.3} - 1^{2-2.3}} \cdot \frac{2-2.3}{3-2.3} \approx 61.23$$

- Bậc trung bình của mạng $\langle k \rangle$ được tính bằng công thức:

$$\langle k \rangle = \sum_{k=k_{min}}^{k_{max}} k \cdot p_k = \sum_{k=k_{min}}^{k_{max}} k \cdot C \cdot k^{-\gamma} = C \cdot \sum_{k=k_{min}}^{k_{max}} k^{1-\gamma}$$

$$\rightarrow \langle k \rangle = C \cdot \int_{k_{min}}^{k_{max}} k^{1-\gamma} dk = C \cdot \left[\frac{k^{2-\gamma}}{2-\gamma} \right]_{k_{min}}^{k_{max}} = C \cdot \frac{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}}{2-\gamma}$$

$$\rightarrow \langle k \rangle = \frac{1-\gamma}{k_{max}^{1-\gamma} - k_{min}^{1-\gamma}} \cdot \frac{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}}{2-\gamma} = \frac{k_{max}^{2-\gamma} - k_{min}^{2-\gamma}}{k_{max}^{1-\gamma} - k_{min}^{1-\gamma}} \cdot \frac{1-\gamma}{2-\gamma}$$

$$\rightarrow \langle k \rangle = \frac{1000^{2-2.3} - 1^{2-2.3}}{1000^{1-2.3} - 1^{1-2.3}} \cdot \frac{1-2.3}{2-2.3} \approx 3.79$$

Vậy bậc trung bình của các hàng xóm của một nút chọn ngẫu nhiên trong mạng cao hơn bậc trung bình của mạng ($61.23 > 3.79$).

Câu d

Câu hỏi: Và bây giờ, bạn giải thích "nghịch lý"trong câu (c), rằng là bạn bè của một nút có nhiều bạn bè hơn chính nút đó?

Trả lời

- Nghịch lý tình bạn (friendship paradox) là hiện tượng mà trung bình bạn bè của một người có nhiều bạn bè hơn chính người đó Trong ngữ cảnh của mạng lưới, điều này có nghĩa là bậc trung bình của các hàng xóm của một đỉnh được chọn ngẫu nhiên thường lớn hơn bậc trung bình của đỉnh đó.
- Cụ thể trong câu c ta đã tính toán và nhận thấy bậc trung bình của hàng xóm của một nút chọn ngẫu nhiên trong mạng cao hơn bậc trung bình của mạng.
- Nghịch lý Tình bạn có thể được giải thích dựa trên sự khác biệt giữa hai giá trị này:
  - Chọn ngẫu nhiên một đỉnh: Khi chúng ta chọn ngẫu nhiên một đỉnh trong mạng, xác suất của việc chọn đỉnh có bậc thấp hoặc trung bình là cao hơn vì phân phối bậc của mạng thường tuân theo phân phối lũy thừa (nhiều đỉnh bậc thấp và ít đỉnh bậc cao).
  - Chọn ngẫu nhiên một liên kết: Khi chúng ta chọn ngẫu nhiên một liên kết và xem xét đỉnh mà liên kết này kết nối tới, xác suất chọn đỉnh có bậc cao hơn là lớn hơn. Lý do là vì các đỉnh có bậc cao có nhiều liên kết hơn và do đó xuất hiện nhiều lần hơn trong danh sách các liên kết.
  - Do đó, bậc trung bình của hàng xóm của một đỉnh chọn ngẫu nhiên thường lớn hơn bậc trung bình của một đỉnh chọn ngẫu nhiên trong mạng hay nói cách khác, bạn bè của một đỉnh có nhiều bạn bè hơn chính đỉnh đó.

Tài liệu tham khảo

[1] Barabási, A.-L. (2016). *Network Science*. Retrieved from http://networksciencebook.com/chapter/3

[2] Network Science. (2021). *Network Science - Properties of Random Networks* [Video]. YouTube. Retrieved from https://youtu.be/lXNbJxoSb7o?si=cz9zqEPFqTk1ImFW

[3] Wikipedia. (2024). *Erdős–Rényi model*. Retrieved from https://en.wikipedia.org/wiki/Erd%C5%91s%E2%80%93R%C3%A9nyi_model