BIỂU DIỄN ĐỒ THỊ

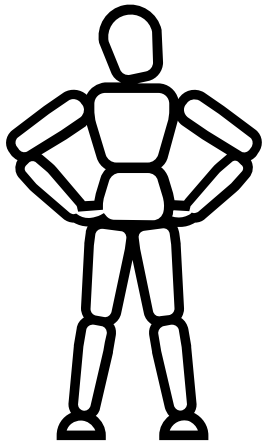
Biên soạn: **ThS. Nguyễn Thị Anh Thư**

Email: thunta@uit.edu.vn

NỘI DUNG

1. **Thuộc tính của mạng (*Network Properties*)**
2. Mô hình đồ thị ngẫu nhiên (*Random Graph Model*)
3. Xây dựng mạng (*Network Construction*)
4. Mô típ mạng (*Network Motifs*)
5. Lan truyền thông tin (*Diffusion Process – SIR*)

1. THUỘC TÍNH CỦA MẠNG



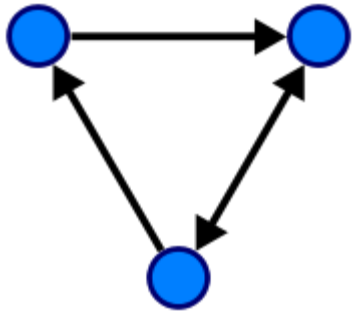
Làm thế nào để đo lường một mạng?

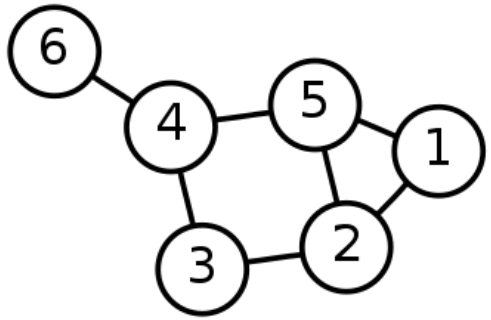
1. THUỘC TÍNH CỦA MẠNG

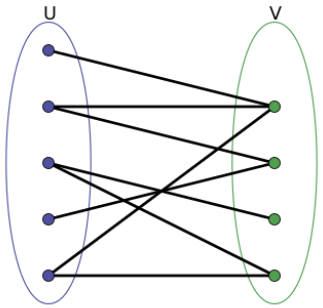
a. Phân phối độ đo (*Degree distribution*): $P(k)$
b. Độ dài đường đi (*Path length*): h
c. Hệ số phân cụm (*Clustering coefficient*): C
d. Thành phần kết nối (*Connected components*): s

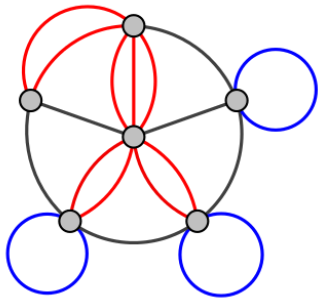
1. THUỘC TÍNH CỦA MẠNG

- **Định nghĩa "Mạng"**
  - Mạng là một mô hình mạng (*đồ thị*) $G = (V, E)$ được cấu tạo bởi các đỉnh và các cạnh:
    - Các đỉnh là tập các đối tượng V.
    - Các cạnh là tập các liên kết E.
    - Cạnh $e=(u,v)$ là cạnh nối 2 đỉnh u và v.
    - Bậc $\deg(v)$ của một đỉnh v là số cạnh liên thuộc với v (*trong đó, khuyên được tính hai lần*).









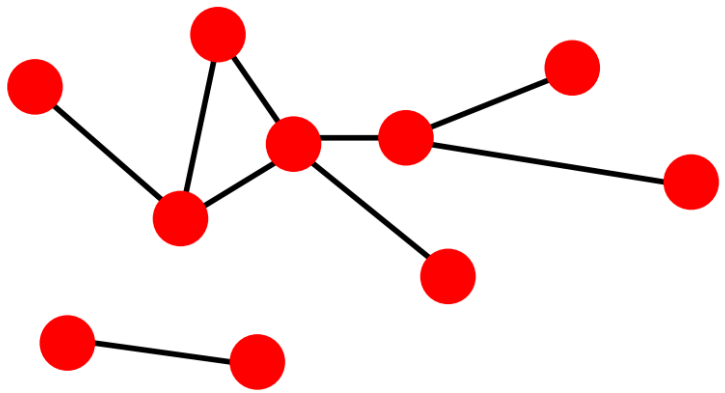
1. THUỘC TÍNH CỦA MẠNG

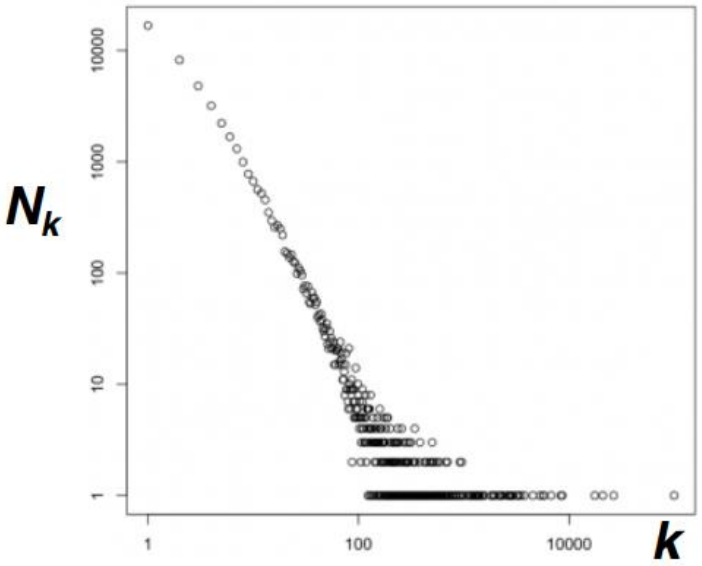
a. Phân phối độ đo (*Degree distribution*): $P(k)$

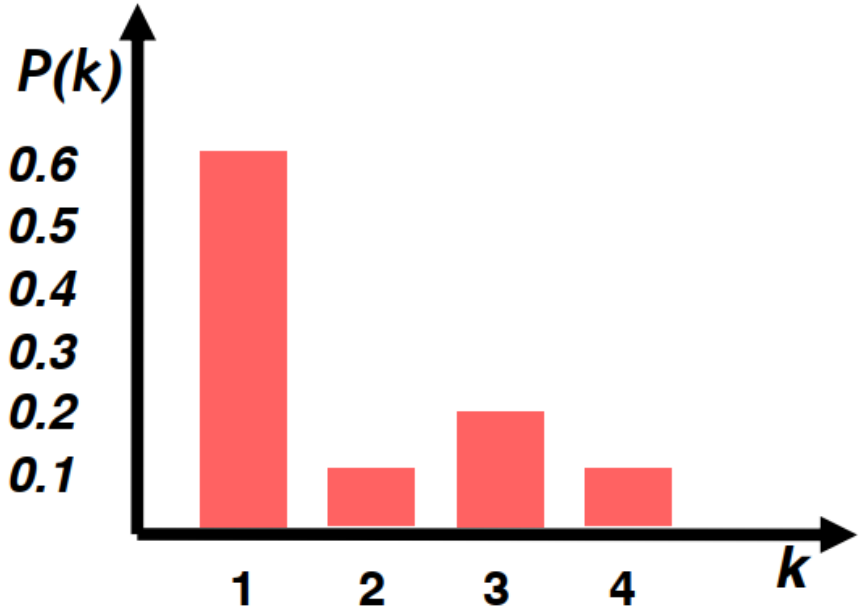
- Xác suất một nút được chọn ngẫu nhiên có bậc k.

$$P(k) = \frac{N_k}{N}$$

- Trong đó:
  - N_k: Số đỉnh trong đồ thị có bậc k.
  - N: Tổng số đỉnh trong đồ thị.





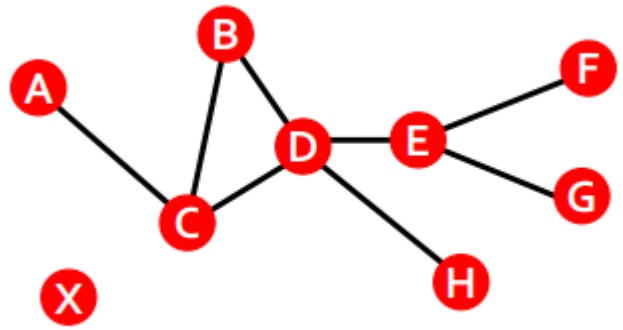


1. THUỘC TÍNH CỦA MẠNG

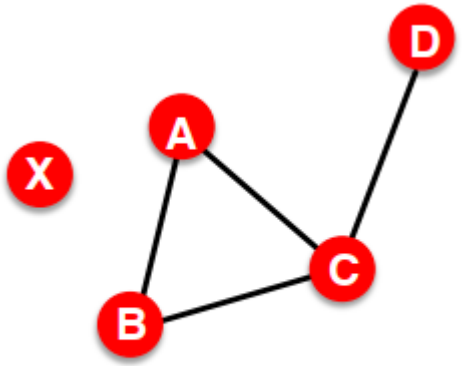
- **Đường đi trong đồ thị**
  - Đường đi là một chuỗi các nút trong đó mỗi nút được liên kết với nút tiếp theo.

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

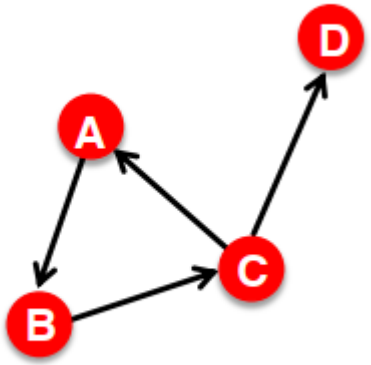
  - Đường đi có thể tự cắt nhau và đi qua cùng một cạnh nhiều lần.
    - Ví dụ: ACBDCDEG



1. THUỘC TÍNH CỦA MẠNG



$h_{B,D} = 2$
$h_{A,X} = \infty$



$h_{B,C} = 1, h_{C,B} = 2$

b. Độ dài đường đi – Khoảng cách (*đường đi ngắn nhất, đường trắc địa*) giữa một cặp nút được xác định là số cạnh dọc theo đường đi ngắn nhất nối các nút.

- Nếu hai nút không được kết nối, khoảng cách thường được xác định là vô hạn.

- Trong đồ thị có hướng, các đường đi cần tuân theo hướng của các mũi tên.
  - Hệ quả: Khoảng cách là không đối xứng.

$$h_{B,C} \neq h_{C,B}$$

1. THUỘC TÍNH CỦA MẠNG

- **Độ dài đường đi trung bình**

$$\bar{h} = \frac{1}{2E_{max}} \sum_{i,j \neq i} h_{ij}$$

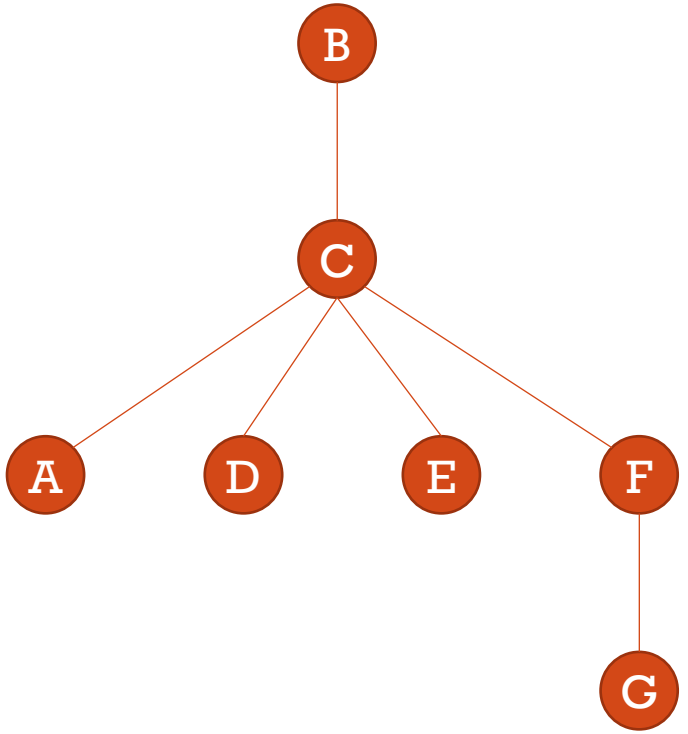
- Với:
  - h_{ij}: là khoảng cách giữa đỉnh i và đỉnh j.
  - E_{max}: số cạnh tối đa có thể tạo được giữa n đỉnh (tổng số các cặp đỉnh).

$$E_{max} = \frac{n(n-1)}{2}$$

- ***Đôi khi, chúng ta chỉ tính giá trị trung bình trên các cặp nút được kết nối. Nghĩa là, chúng ta sẽ bỏ qua những khoảng cách có độ dài vô hạn.***

1. THUỘC TÍNH CỦA MẠNG

- **Đường kính của đồ thị**
- **Đường kính của đồ thị là khoảng cách lớn nhất giữa các cặp đỉnh.**
- Giải pháp để tìm đường kính của đồ thị là tìm tất cả các đường đi (*khoảng cách*) và sau đó tìm đường đi lớn nhất.
- **Ví dụ**:
  - Đường kính của đồ thị bên là 3.
  - BC → CF → FG



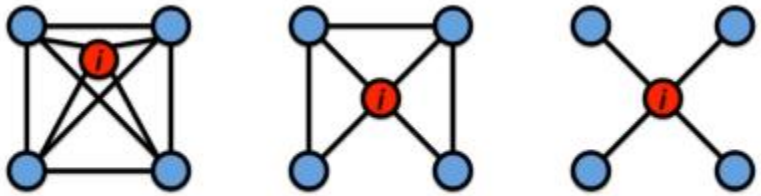
1. THUỘC TÍNH CỦA MẠNG

c. Hệ số phân cụm (*Clustering coefficient*)

- ***Phần nào của những người hàng xóm của tôi được kết nối với nhau?***
- Đỉnh i có bậc đỉnh là k_i.
- Ta có:

$$C_i \in [0,1], C_i = \frac{2e_i}{k_i(k_i - 1)}$$

  - e_i: là số cạnh giữa các đỉnh liền kề (hàng xóm) của đỉnh i.



$C_i = 1$ $C_i = 1/2$ $C_i = 0$

- Hệ số phân cụm trung bình:

$$C = \frac{1}{N}\sum_{i}^{N} C_i$$

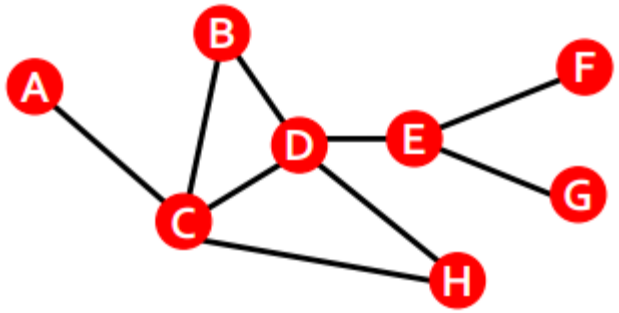
1. THUỘC TÍNH CỦA MẠNG

c. Hệ số phân cụm (*Clustering coefficient*)

- ***Phần nào của những người hàng xóm của tôi được kết nối với nhau?***
- Đỉnh i có bậc đỉnh là k_i.
- Ta có:

$$C_i \in [0,1], C_i = \frac{2e_i}{k_i(k_i - 1)}$$

  - e_i: là số cạnh giữa các đỉnh liền kề (hàng xóm) của đỉnh i.
- Ví dụ: Cho đồ thị như hình sau. **Hãy tính C_B, C_D và C.**



$k_B=2,\ e_B=1,\ C_B=2/2 = 1$

$k_D=4,\ e_D=2,\ C_D=4/12 = 1/3$

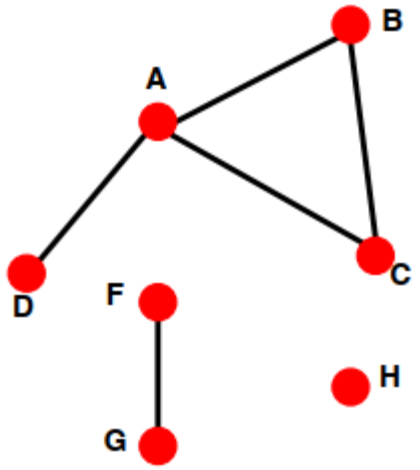
Avg. clustering: C=0.33

1. THUỘC TÍNH CỦA MẠNG

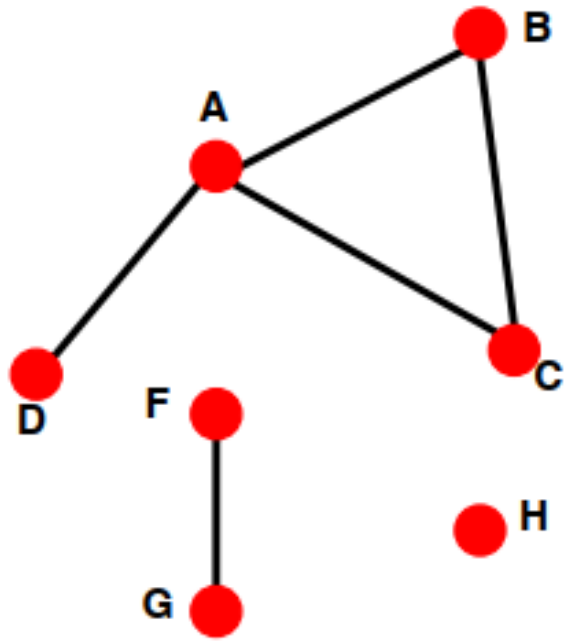
d. Thành phần kết nối (Connected components)

- **Đồ thị thành phần**: Tập hợp trong đó luôn tồn tại đường đi giữa 2 đỉnh bất kỳ.
- Kích thước của thành phần được kết nối lớn nhất
  - Tập hợp lớn nhất trong đó hai đỉnh bất kỳ có thể được nối với nhau bằng một đường dẫn.

⇒ **Đồ thị thành phần lớn nhất (Largest component) = Thành phần khổng lồ (Giant component)**



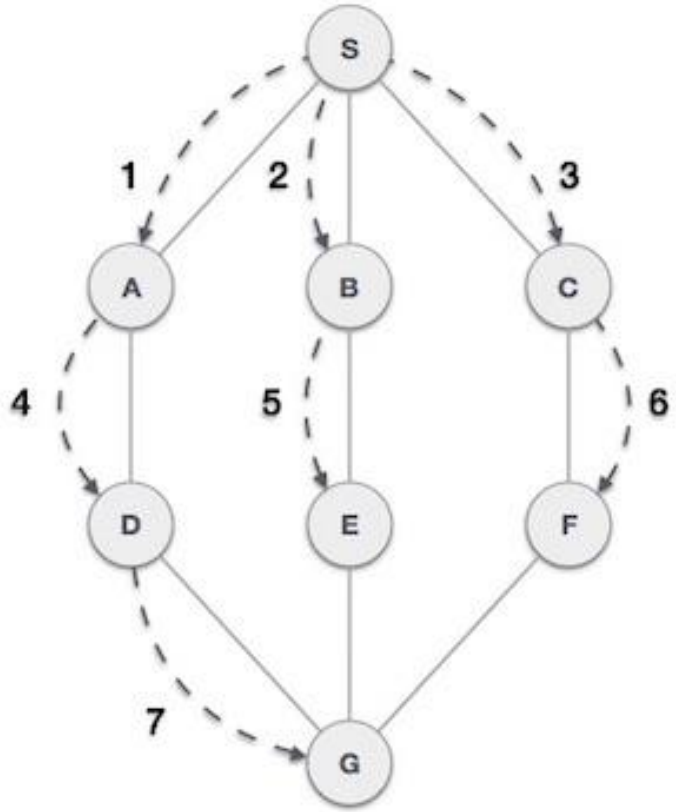
1. THUỘC TÍNH CỦA MẠNG



d. Thành phần kết nối (Connected components)

- **Cách tìm các thành phần được kết nối:**
  - Bắt đầu từ nút ngẫu nhiên và thực hiện ***Tìm kiếm đầu tiên theo chiều rộng (BFS)***
  - Gắn nhãn các nút mà BFS đã truy cập
  - Nếu tất cả các nút được truy cập, mạng đã được kết nối
  - Nếu không, hãy tìm một nút không được truy cập và lặp lại BFS

1. THUỘC TÍNH CỦA MẠNG

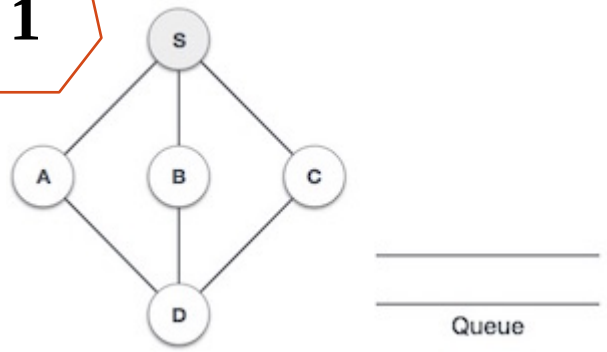


- **BREADTH FIRST SEARCH**
  - **Giải thuật tìm kiếm theo chiều rộng (Breadth First Search – viết tắt là BFS)** duyệt qua một đồ thị theo chiều rộng và ***sử dụng hàng đợi (queue)*** để ghi nhớ đỉnh liền kề để bắt đầu việc tìm kiếm khi không gặp được đỉnh liền kề trong bất kỳ vòng lặp nào.
  - ***Giải thuật này tuân theo qui tắc***:
    - **Qui tắc 1**: Duyệt tiếp tới đỉnh liền kề mà chưa được duyệt. Đánh dấu đỉnh mà đã được duyệt. Hiển thị đỉnh đó và đẩy vào trong một hàng đợi (queue)..
    - **Qui tắc 2**: Nếu không tìm thấy đỉnh liền kề, thì xóa đỉnh đầu tiên trong hàng đợi.
    - **Qui tắc 3**: Lặp lại Qui tắc 1 và 2 cho tới khi hàng đợi là trống.

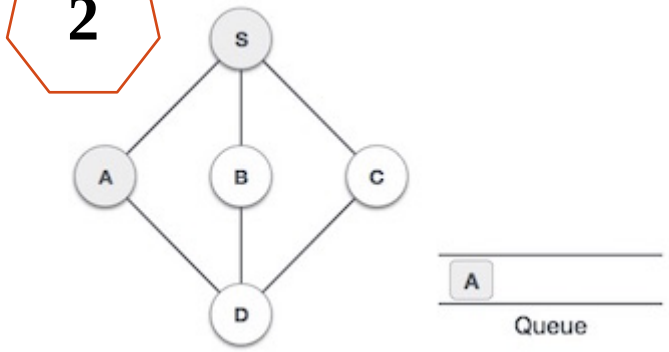
1. THUỘC TÍNH CỦA MẠNG

- **BREADTH FIRST SEARCH**

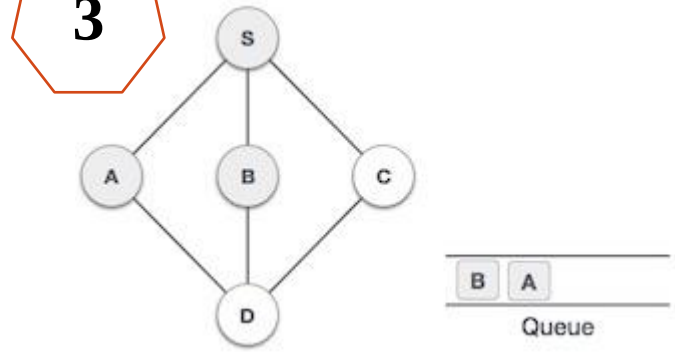
1



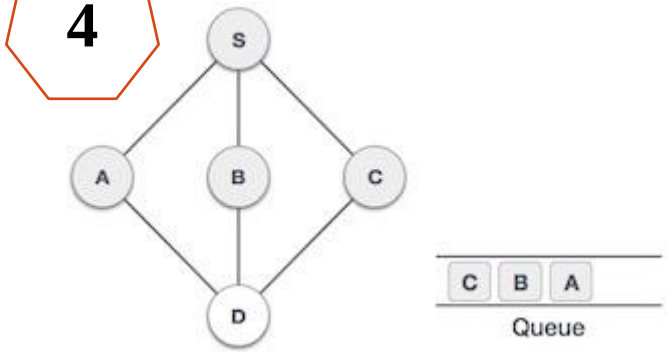
2



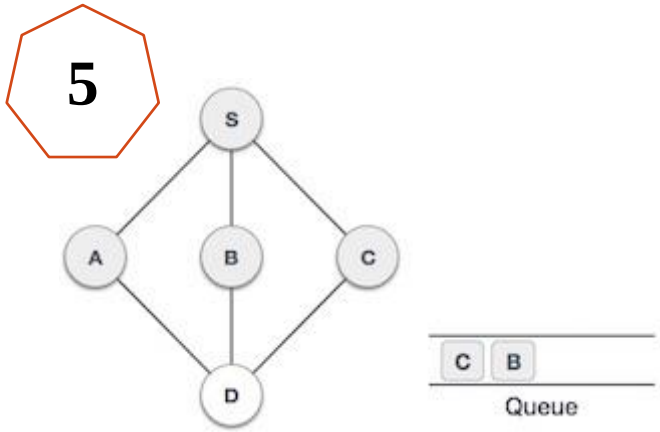
3



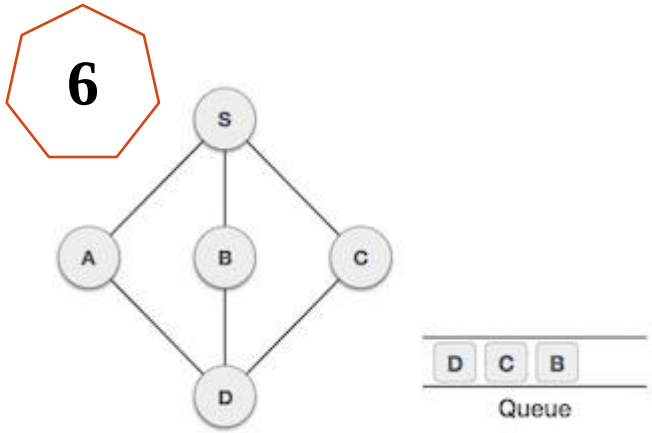
4



5



6



1. THUỘC TÍNH CỦA MẠNG

a. Phân phối độ đo (*Degree distribution*): $P(k)$
b. Độ dài đường đi (*Path length*): h
c. Hệ số phân cụm (*Clustering coefficient*): C
d. Thành phần kết nối (*Connected components*): s

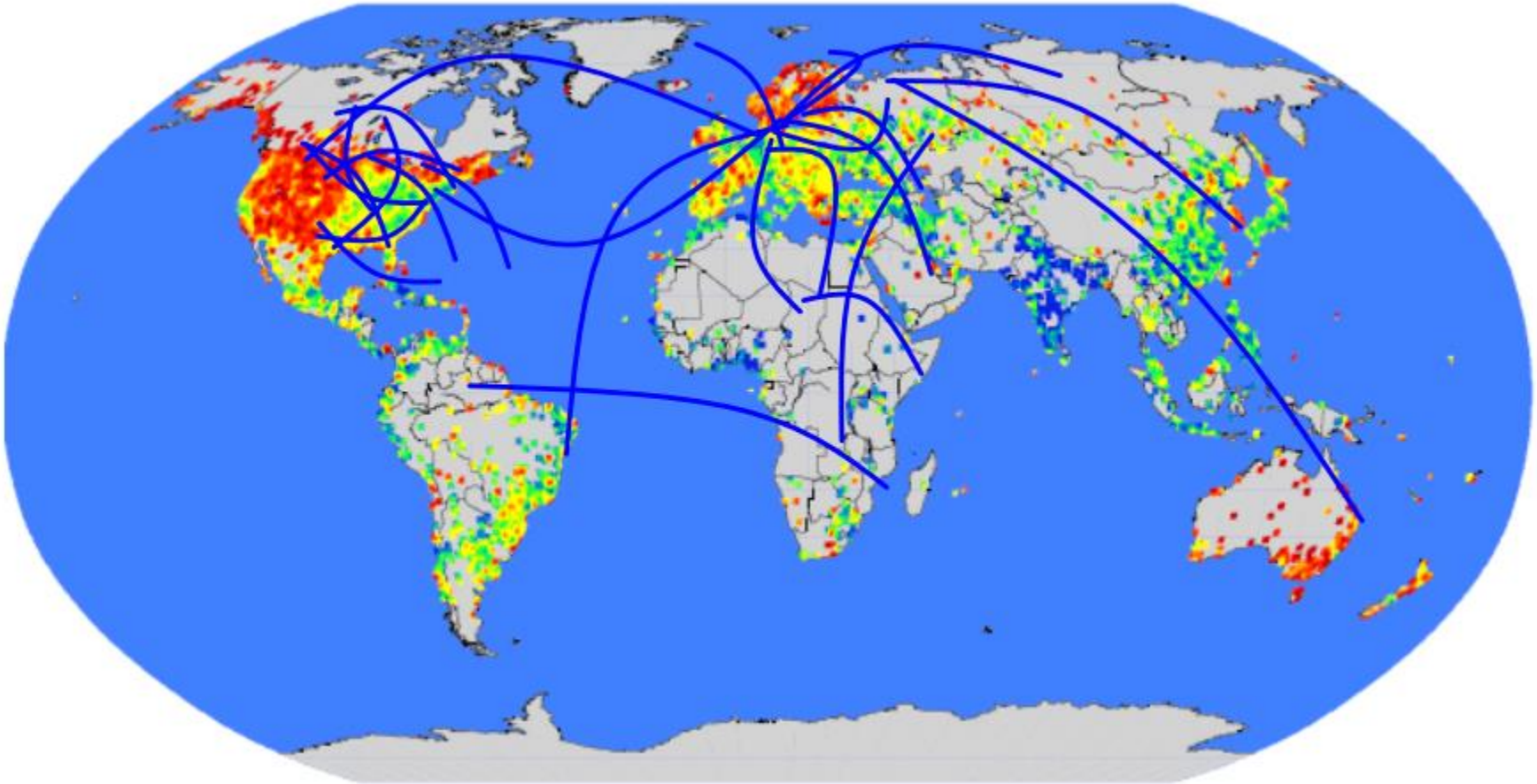
1. THUỘC TÍNH CỦA MẠNG

- **Hãy tính các thuộc tính $P(k)$, h, C và s của một mạng trong thực tế!**
  - Phân phối độ đo (*Degree distribution*): $P(k)$
  - Độ dài đường đi (*Path length*): h
  - Hệ số phân cụm (*Clustering coefficient*): C
  - Thành phần kết nối (*Connected components*): s

1. THUỘC TÍNH CỦA MẠNG

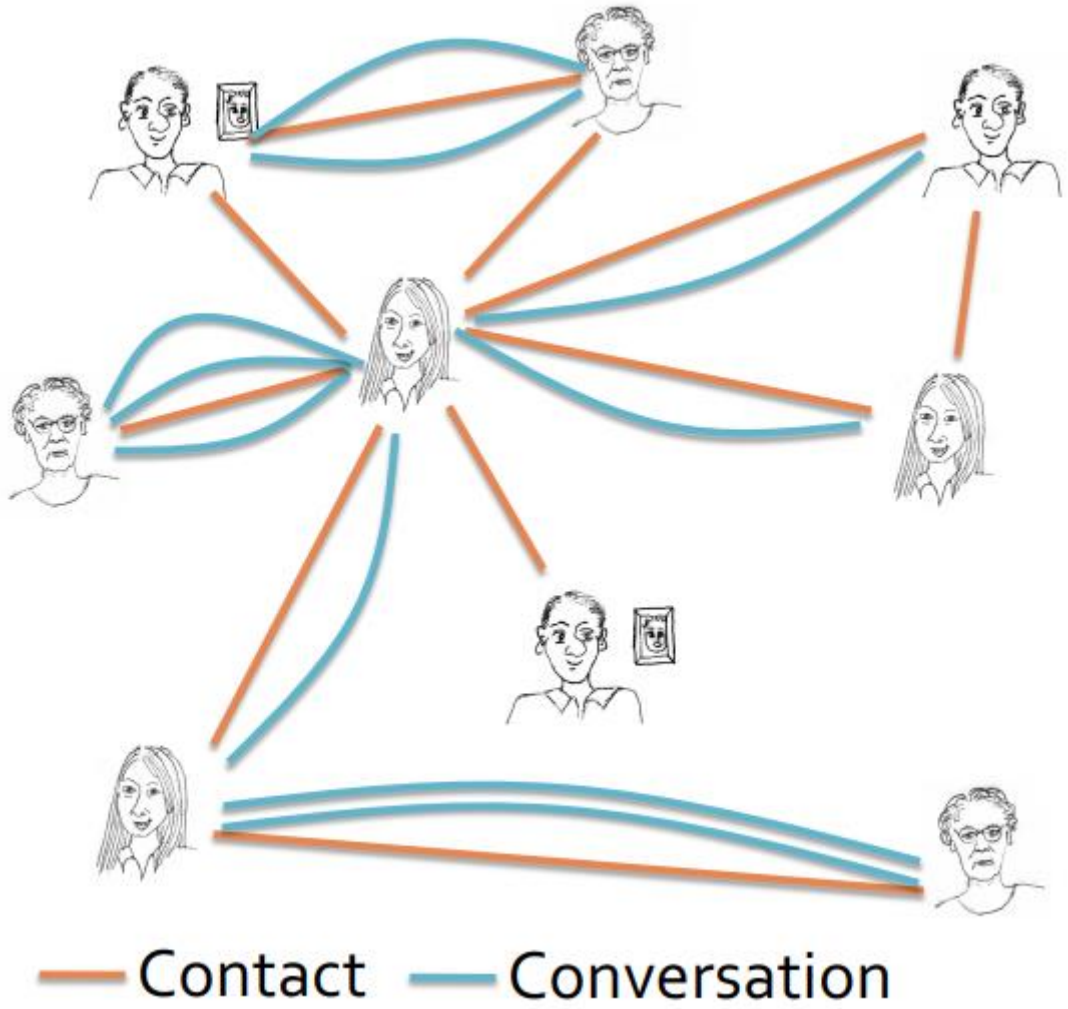
- **MSN Messenger** (*hay còn được gọi là Windows Live Messenger*): Dịch vụ chat của Microsoft.
  - Trong 1 tháng hoạt động
    - 245 triệu người dùng đã đăng nhập
    - 180 triệu người dùng tham gia vào các cuộc trò chuyện
    - Hơn 30 tỷ cuộc trò chuyện
    - Hơn 255 tỷ tin nhắn đã trao đổi

1. THUỘC TÍNH CỦA MẠNG



Mạng: 180 triệu người dùng, 1.3 tỷ cạnh.

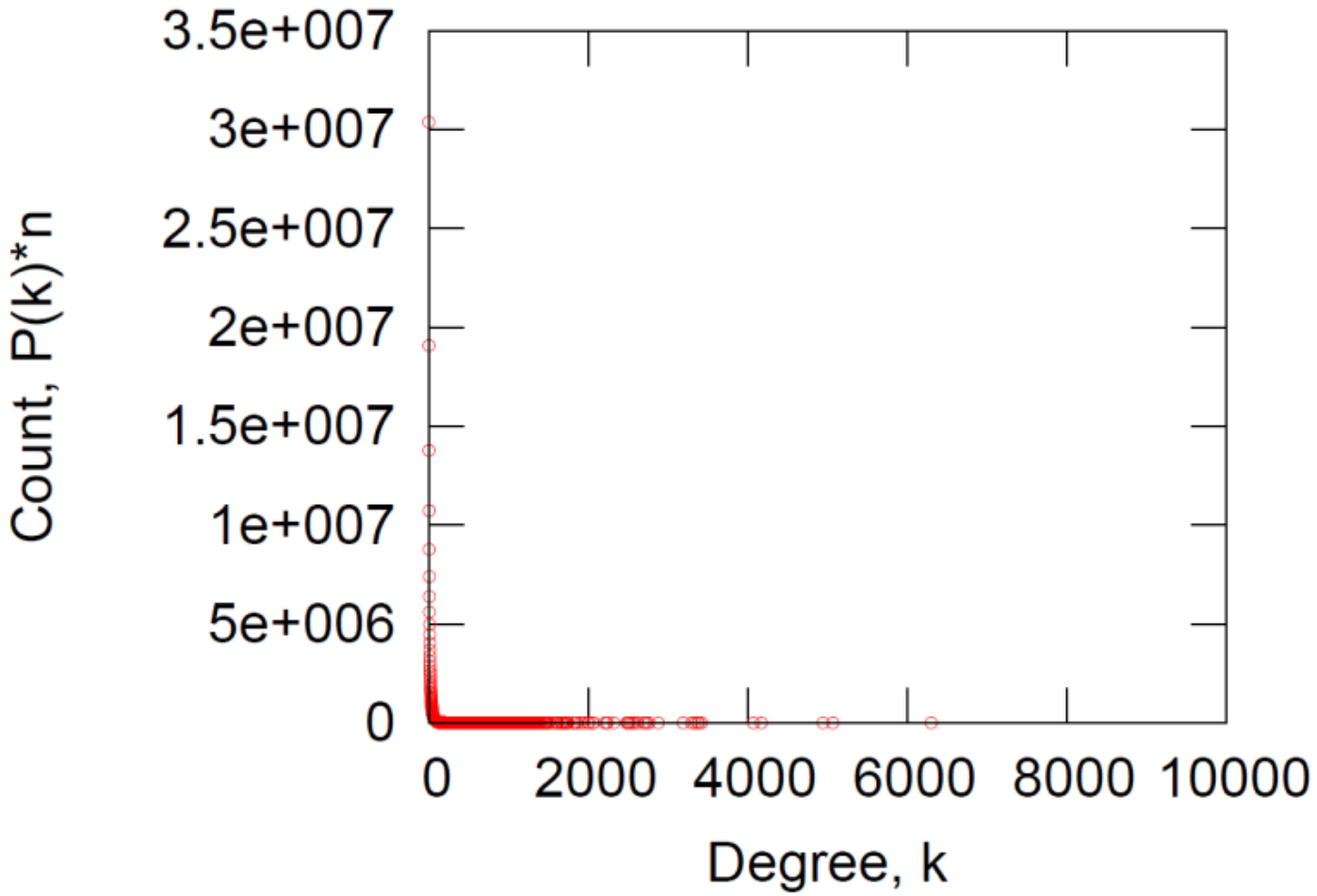
1. THUỘC TÍNH CỦA MẠNG



- **Mạng biểu diễn dữ liệu chat của MSN là đồ thị vô hướng.**
  - ***Đỉnh***: người dùng.
  - ***Cạnh***: thể hiện mối quan hệ *Contact* (có tiếp xúc) hoặc *Conversation* (có cuộc hội thoại giữa 2 người dùng).
- **Thông tin đồ thị MSN:**
  - Cạnh (u, v) tồn tại nếu người dùng u và v đã trao đổi ít nhất 1 tin nhắn.
  - $N = 180$ triệu người
  - $E = 1{,}3$ tỷ cạnh

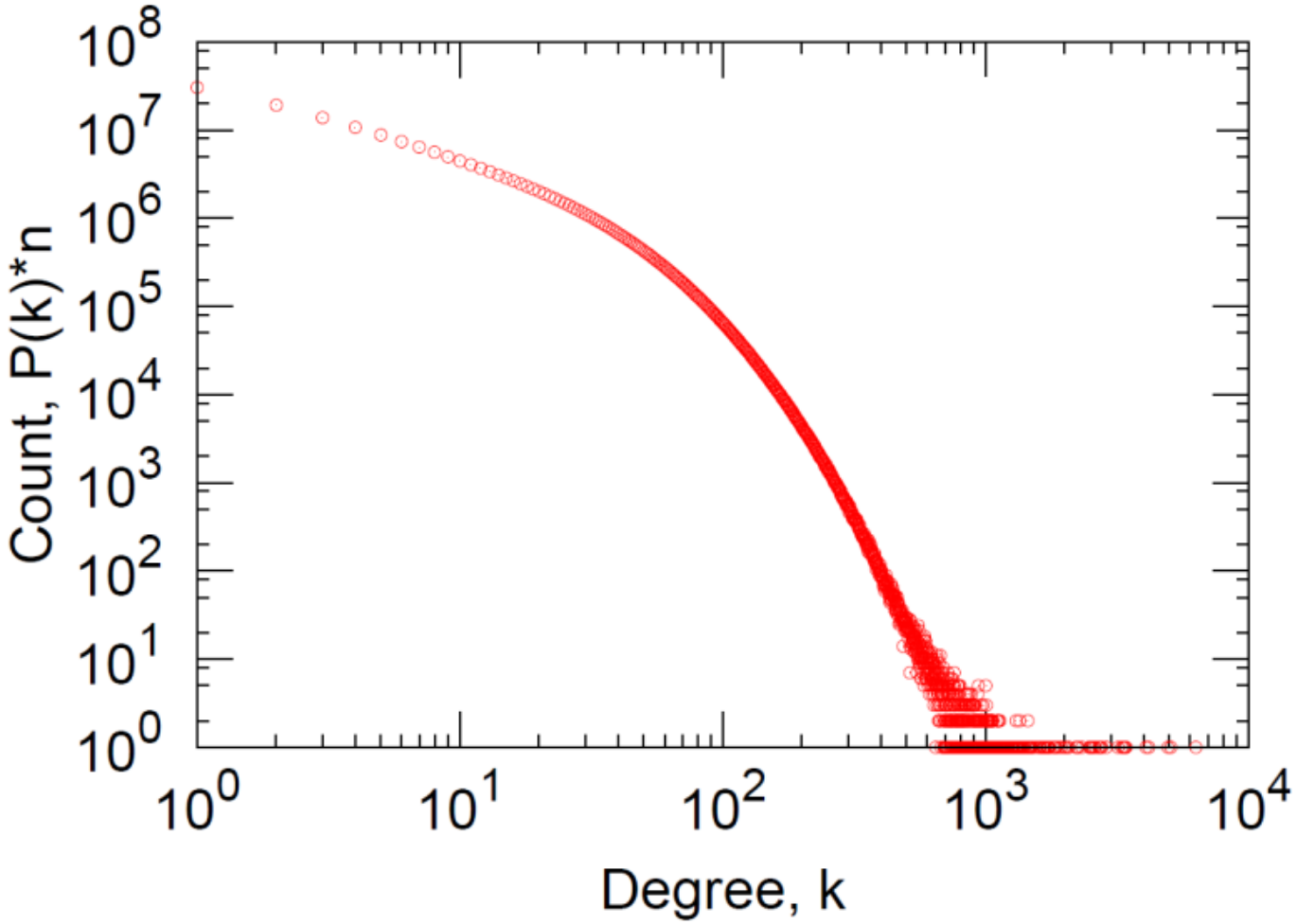
1. THUỘC TÍNH CỦA MẠNG

- **Phân phối độ đo (*Degree distribution*)**

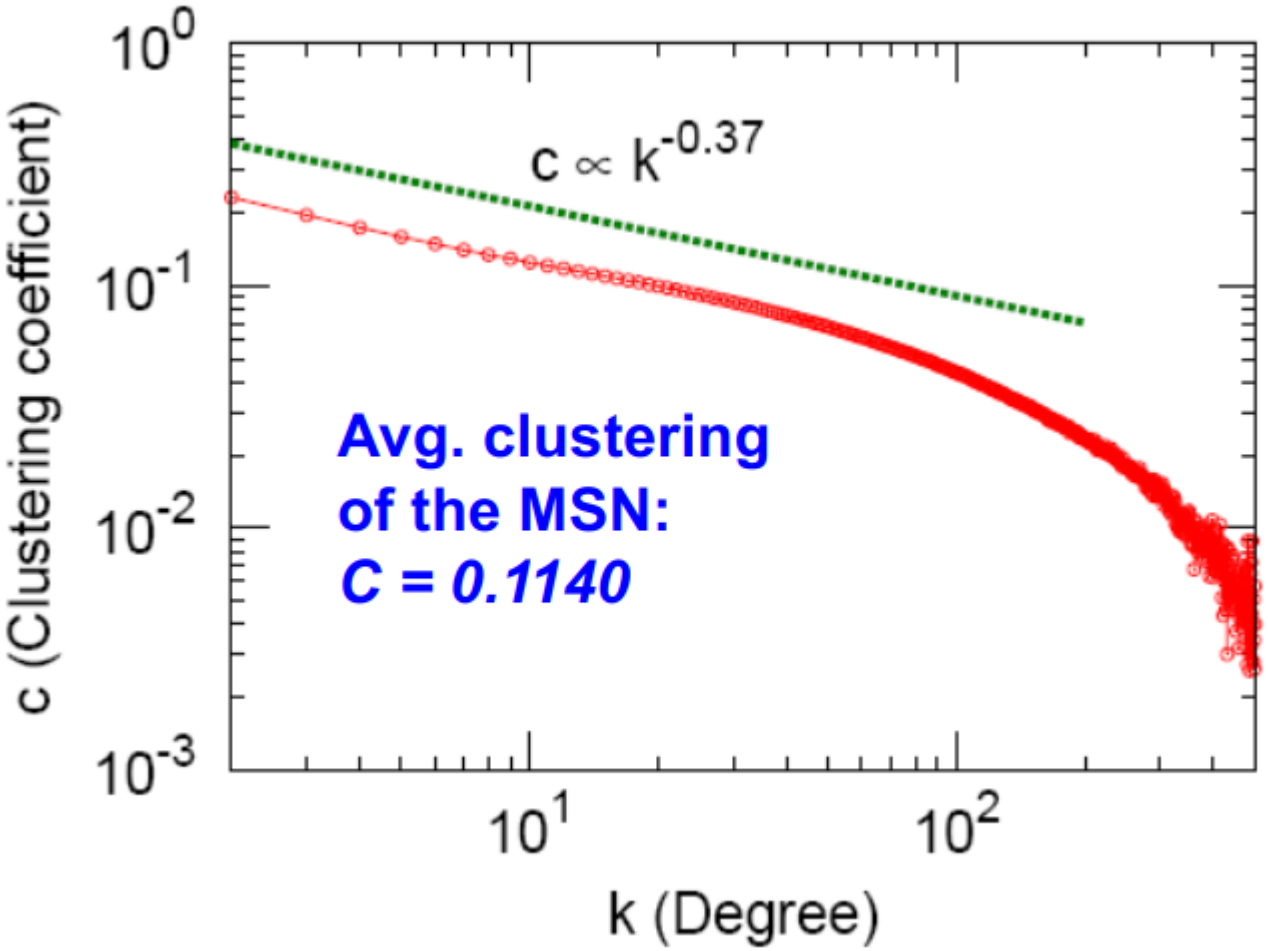


1. THUỘC TÍNH CỦA MẠNG

- **Phân phối độ đo (*Degree distribution*)**



1. THUỘC TÍNH CỦA MẠNG

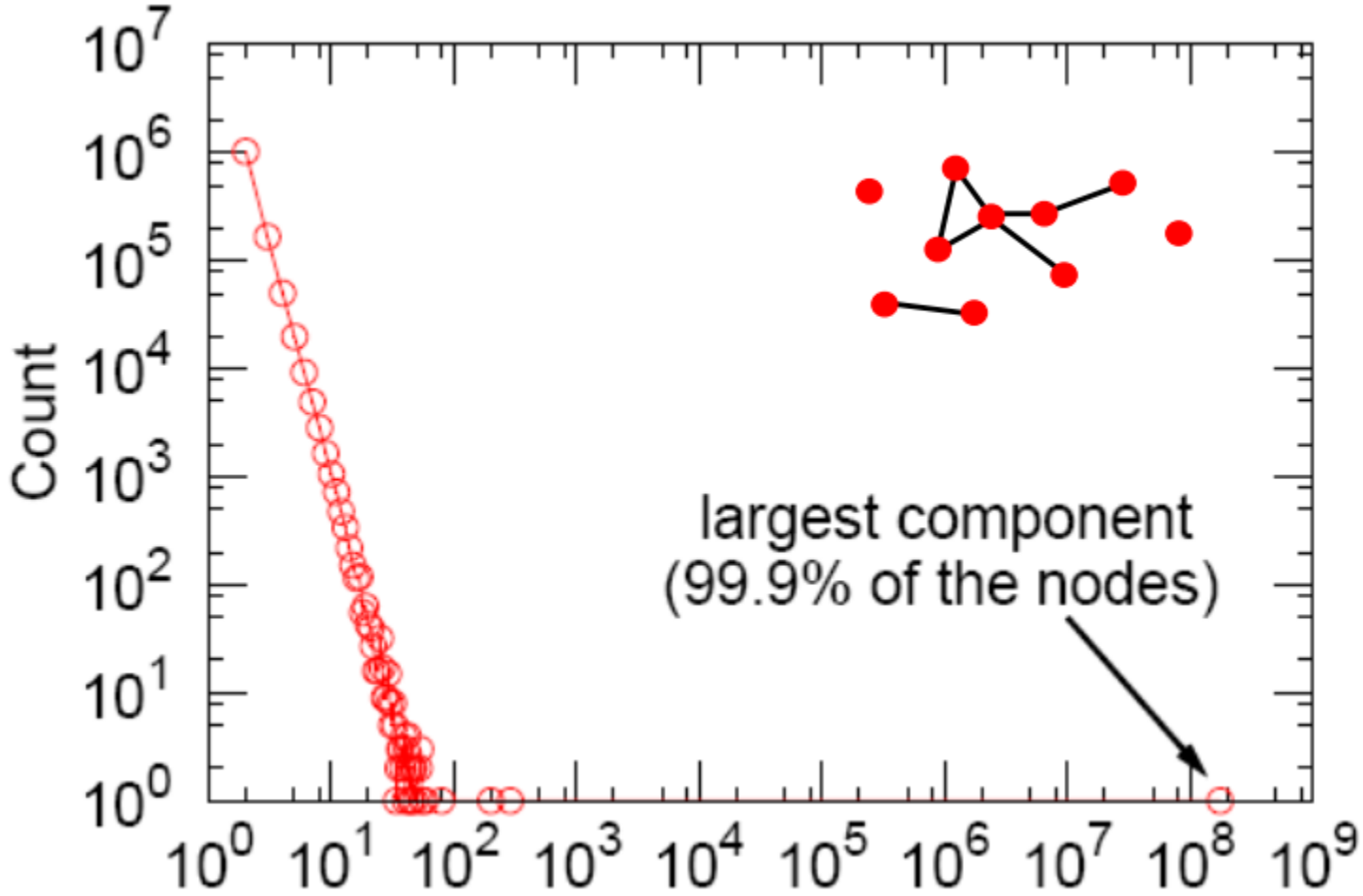


- **Hệ số phân cụm (*Clustering coefficient*)**
  - C_k là hệ số phân cụm trung bình C_i của đỉnh i có bậc k.

$$C_k = \frac{1}{N_k} \sum_{i:k_i=k} C_i$$

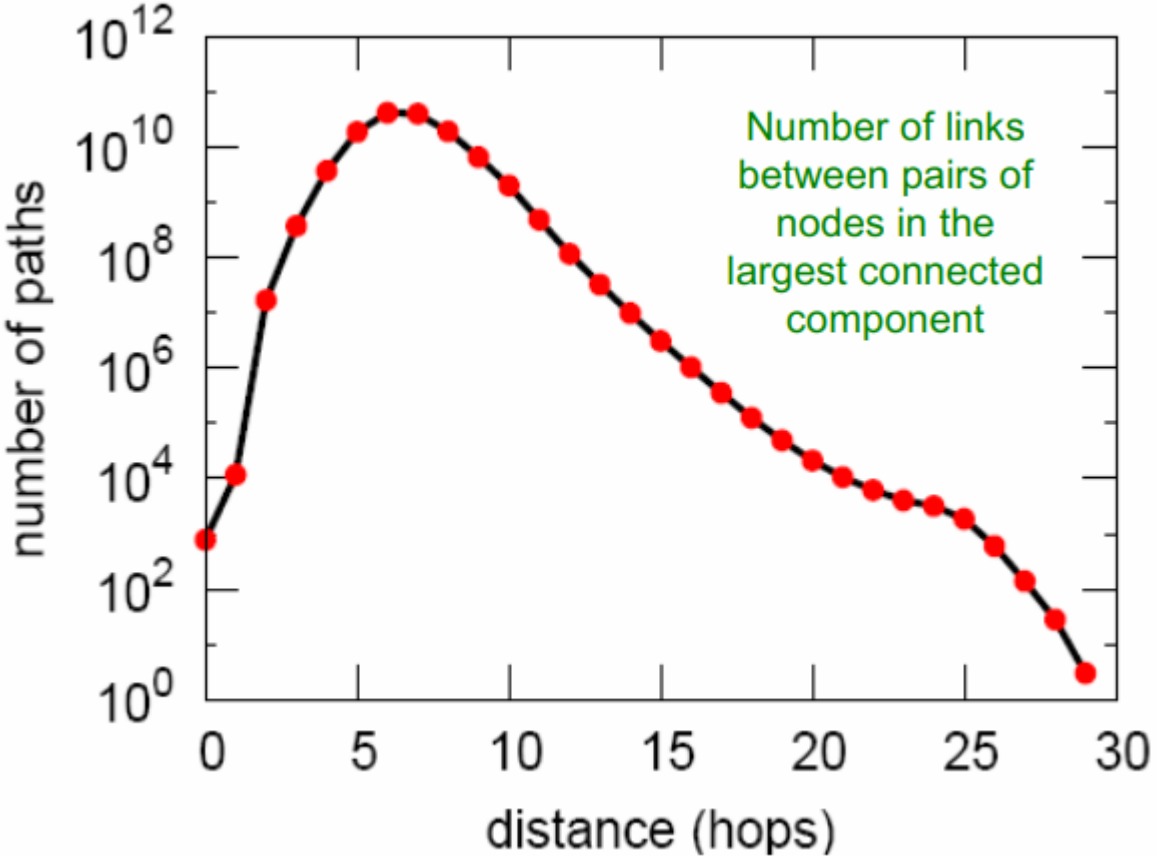
1. THUỘC TÍNH CỦA MẠNG

- **Thành phần kết nối (*Connected components*)**



1. THUỘC TÍNH CỦA MẠNG

- **Độ dài đường đi (*Path length*) – Khoảng cách**
  - Độ dài đường đi trung bình: 6.6
  - 90% số đỉnh được duyệt trong 8 bước nhảy



Steps	#Nodes
0	1
1	10
2	78
3	3,96
4	8,648
5	3,299,252
6	28,395,849
7	79,059,497
8	52,995,778
9	10,321,008
10	1,955,007
11	518,410
12	149,945
13	44,616
14	13,740
15	4,476
16	1,542
17	536
18	167
19	71
20	29
21	16
22	10
23	3
24	2
25	3

nodes as we do BFS out of a random node

1. THUỘC TÍNH CỦA MẠNG

- Phân phối độ đo (*Degree distribution*): Bị lệch nhiều
  Độ đo trung bình là 14.4
- Độ dài đường đi (*Path length*): 6.6
- Hệ số phân cụm (*Clustering coefficient*): 0.11
- Thành phần kết nối (*Connected components*): Đồ thị thành phần khổng lồ

Những giá trị này có "được mong đợi" không?
Các đặc điểm này có"đáng ngạc nhiên" không?
Để trả lời điều này, chúng ta cần một "null-model"!

NỘI DUNG

1. Thuộc tính của mạng (*Network Properties*)
2. **Mô hình đồ thị ngẫu nhiên (*Random Graph Model*)**
3. Xây dựng mạng (*Network Construction*)
4. Mô típ mạng (*Network Motifs*)
5. Lan truyền thông tin (*Diffusion Process – SIR*)

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

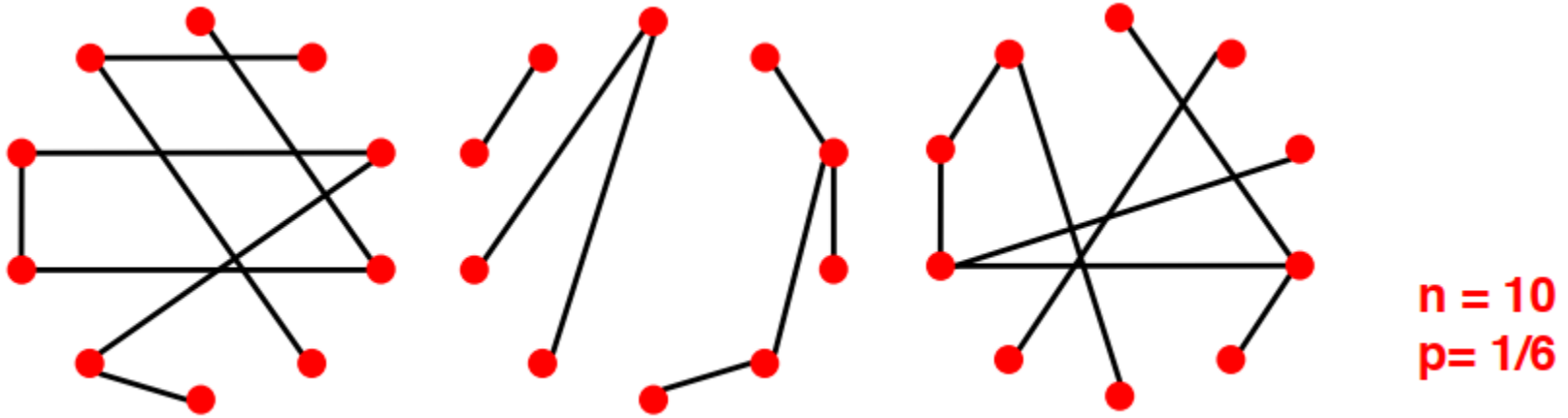
Mô hình đồ thị đơn giản nhất

- **Đồ thị ngẫu nhiên [Erdös-Renyi, ‘60]**
  - $G_{n,p}$: Đồ thị vô hướng có n đỉnh và mỗi cạnh (u, v) xuất hiện một cách độc lập với xác suất p.

Mô hình đồ thị ngẫu nhiên sẽ tạo ra loại mạng nào?

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

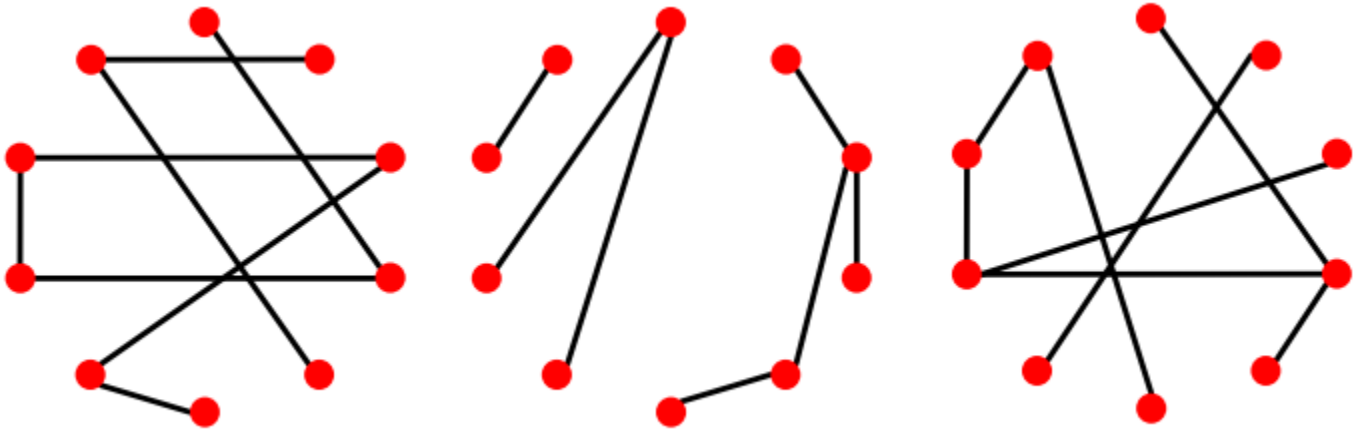
- **Đồ thị ngẫu nhiên [Erdös-Renyi, ‘60]**
  - $G_{n,p}$: Đồ thị vô hướng có n đỉnh và mỗi cạnh (u, v) xuất hiện một cách độc lập với xác suất p.
  - Đồ thị $G_{n,p}$ là kết quả của một quá trình ngẫu nhiên.
  - n và p không xác định duy nhất một đồ thị.
  - **Mô hình đồ thị ngẫu nhiên $G_{n,p}$ tạo ra một tập các đồ thị có cùng n và p.**



2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

- **Đồ thị ngẫu nhiên [Erdös-Renyi, ‘60]**
  - **Mô hình đồ thị ngẫu nhiên $G_{n,p}$ tạo ra một tập các đồ thị có cùng n và p.**

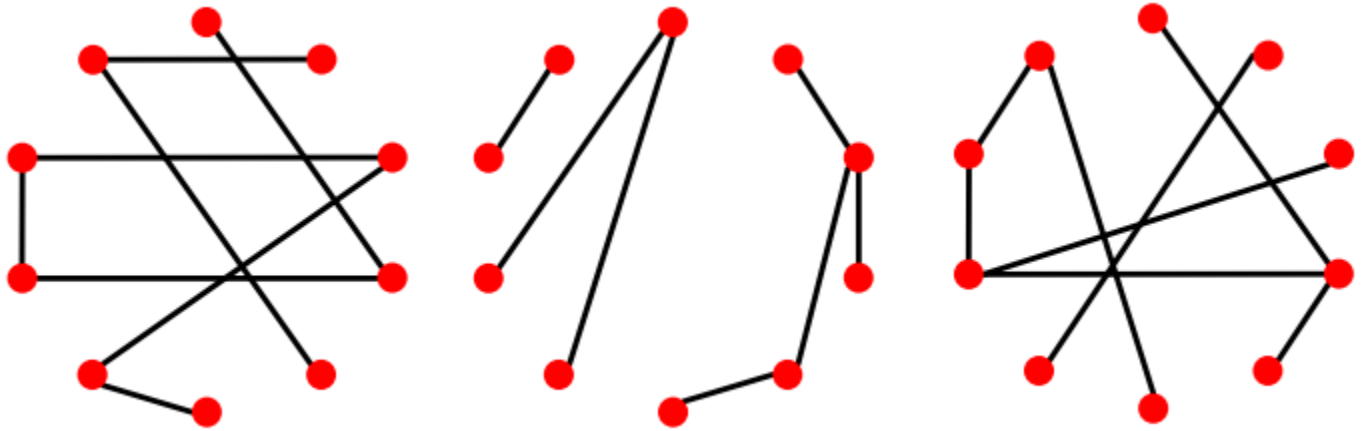
Số cạnh dự kiến trong $G_{n,p}$ là bao nhiêu?



n = 10
p= 1/6

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

- **Đồ thị ngẫu nhiên [Erdös-Renyi, '60]**
  - **Mô hình đồ thị ngẫu nhiên $G_{n,p}$ tạo ra một tập các đồ thị có cùng n và p.**
  - Số cạnh dự kiến trong $G_{n,p}$ là $\mathrm{m} = \binom{n}{2} \times p$.
  - Ví dụ: Số cạnh dự kiến trong $G_{10,\frac{1}{6}}$ là $\mathrm{m} = \binom{10}{2} \times \frac{1}{6} = 7.5$



n = 10
p= 1/6

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

- **Đồ thị ngẫu nhiên [Erdös-Renyi, '60]**
  - **Mô hình đồ thị ngẫu nhiên $G_{n,p}$ tạo ra một tập các đồ thị có cùng n và p.**
  - Số cạnh dự kiến trong $G_{n,p}$ là $\mathrm{m} = \binom{n}{2} \times p$.
  - Số phần tử trong tập đồ thị của mô hình $G_{n,p}$ là $\binom{C_n^2}{m}$.
  - p đóng vai trò là một trọng số và $p \in [0,1]$.
  - Khi p tăng từ 0 đến 1 thì mô hình ngày càng có nhiều khả năng bao gồm:
    - Các đồ thị có nhiều cạnh hơn.
    - Tập các đồ thị có ít đồ thị hơn.

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Thuộc tính của $G_{n,p}$

a. Phân phối độ đo (*Degree distribution*): $P(k)$
b. Độ dài đường đi (*Path length*): h
c. Hệ số phân cụm (*Clustering coefficient*): C

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Phân phối độ đo (*Degree distribution*)

- $P(k)$ của $G_{n,p}$ là một nhị thức.

$$\boldsymbol{P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}}$$

- Trong đó:
  - $\binom{n-1}{k}$: Chọn k đỉnh trong số $n-1$ đỉnh.
  - p^k: Xác suất có k cạnh.
  - $(1-p)^{n-1-k}$: Xác suất của phần còn lại $n-1-k$ cạnh bị thiếu.

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Hệ số phân cụm (*Clustering coefficient*)

- Hệ số phân cụm đỉnh i có bậc đỉnh là k_i trong đồ thị $G = (V, E)$là:

$$C_i \in [0,1], C_i = \frac{2e_i}{k_i(k_i - 1)}$$

  - Với: e_i là số cạnh giữa các đỉnh liền kề (hàng xóm) của đỉnh i.
- Các cạnh trong $G_{n,p}$ xuất hiện một cách độc lập với xác suất p.
- Xác suất xuất hiện một cạnh e_i trong $G_{n,p}$ là:

$$E[e_i] = p\frac{k_i(k_i - 1)}{2}$$

  - Với:
    - p: Xác suất một cạnh độc lập xuất hiện.
    - $\frac{k_i(k_i-1)}{2}$: Số cạnh có thể tạo ra từ các đỉnh liền kề với đỉnh i.

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Hệ số phân cụm (*Clustering coefficient*)

- Hệ số phân cụm:

$$E(C) = \frac{pk_i(k_i - 1)}{k_i(k_i - 1)} = p = \frac{\bar{k}}{n - 1} \approx \frac{\bar{k}}{n}$$

  - Với:
    - **Bậc đỉnh trung bình (Average degree)** của đồ thị vô hướng:

$$\bar{k} = \frac{1}{n}\sum_{i=1}^{n} k_i = \frac{2m}{n}$$

    - $p = \frac{m}{\frac{n(n-1)}{2}} = \frac{2m}{n(n-1)}$: Xác suất một cạnh xuất hiện độc lập trong đồ thị G có n đỉnh và m cạnh.

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Hệ số phân cụm (*Clustering coefficient*)

- Hệ số phân cụm:

$$E(C) = p = \frac{\overline{k}}{n}$$

- Hệ số phân cụm của một đồ thị ngẫu nhiên là nhỏ.
- **Nếu kích thước đồ thị tăng với bậc đỉnh trung bình (*dùng để đo số cạnh so với số đỉnh*) không đổi thì *C* sẽ giảm theo kích thước đồ thị.**

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Thuộc tính của $G_{n,p}$

a. Phân phối độ đo (*Degree distribution*): $\boldsymbol{P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}}$

b. Độ dài đường đi (*Path length*): h

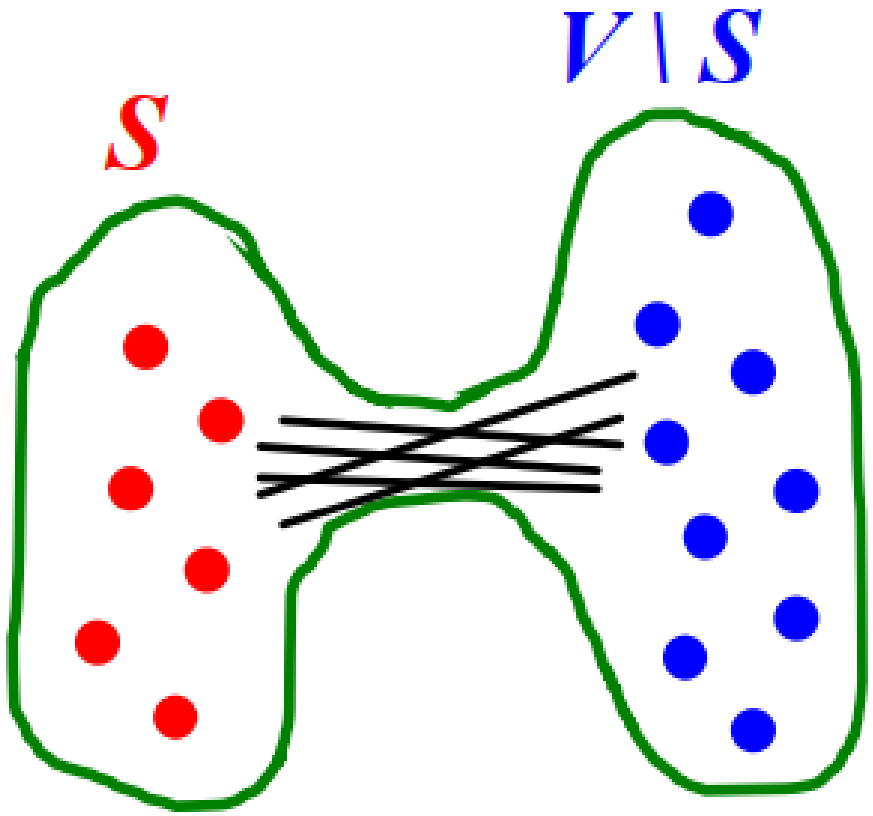
c. Hệ số phân cụm (*Clustering coefficient*): $\boldsymbol{C = p = \frac{\bar{k}}{n}}$

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Độ mở rộng (*Expansion*)

- Đồ thị $G(V, E)$ có độ mở rộng α:
  - Nếu
    $\forall S \subseteq V: \#\ edges\ leaving\ S \geq \alpha . \min(|S|, |V \backslash S|)$
  - Ta có:

$$\alpha = \min_{S \subseteq V} \frac{\#\ edges\ leaving\ S}{\min(|S|, |V \backslash S|)}$$

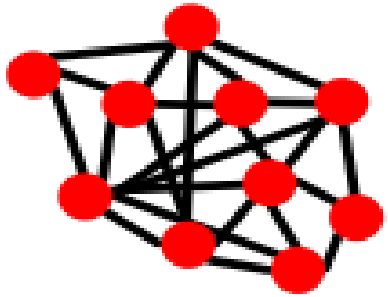


2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

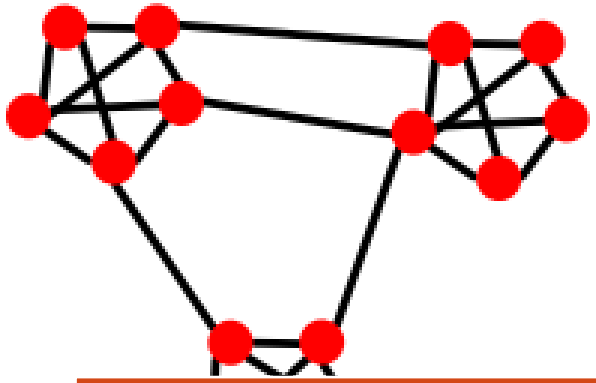
Độ mở rộng (*Expansion*)

- Là thước đo "sự mạnh mẽ" của đồ thị.
- Để ngắt l đỉnh, chúng ta cần cắt $\geq \alpha . l$ cạnh.

Độ mở rộng yếu



Độ mở rộng mạnh

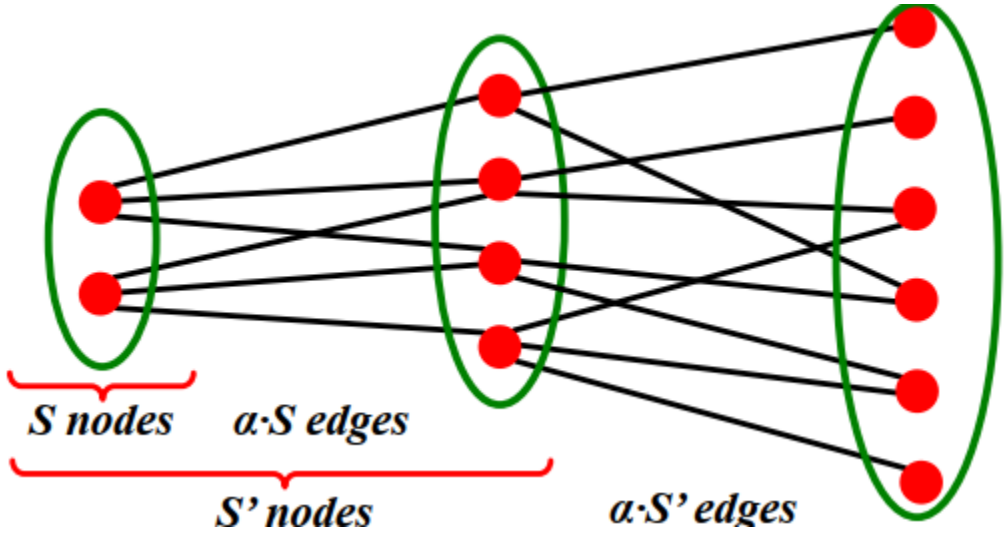


Mạng xã hội (*gồm nhiều cộng đồng*)

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Độ dài đường đi (*Path length*)

- Trong đồ thị có n đỉnh với độ mở rộng α cho tất cả các cặp đỉnh có độ dài đường đi là $O((\log n)/\alpha)$.
- Đồ thị ngẫu nhiên $G_{n,p}$:
  - **Độ dài đường đi (khoảng cách): $O(\log n)$.**
  - Đồ thị ngẫu nhiên có khả năng mở rộng tốt nên cần một số bước logarit để BFS có thể truy cập tất cả các nút.



2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Thuộc tính của $G_{n,p}$

a. Phân phối độ đo (*Degree distribution*): $\boldsymbol{P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}}$

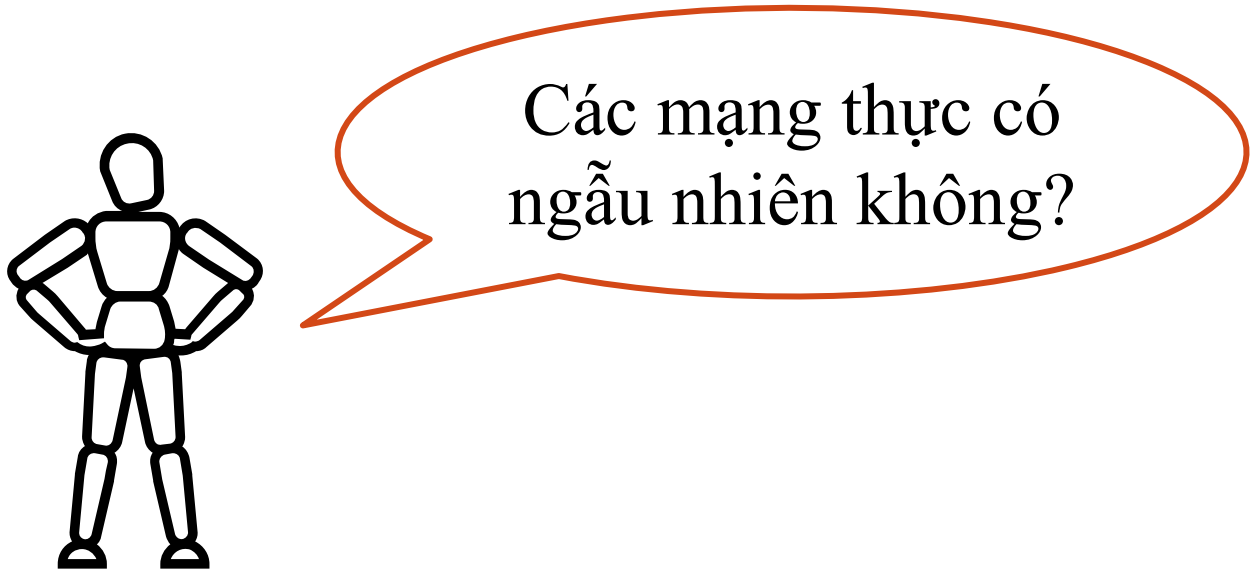
b. Độ dài đường đi (*Path length*): $\boldsymbol{O(\log n)}$

c. Hệ số phân cụm (*Clustering coefficient*): $\boldsymbol{C = p = \frac{\bar{k}}{n}}$

2. MÔ HÌNH ĐỒ THỊ NGẪU NHIÊN

Các vấn đề với mô hình đồ thị ngẫu nhiên:

- Phân phối độ đo khác nhau.
- Không có cấu trúc cục bộ (hệ số phân cụm quá thấp).



- **Câu trả lời đơn giản là KHÔNG.**

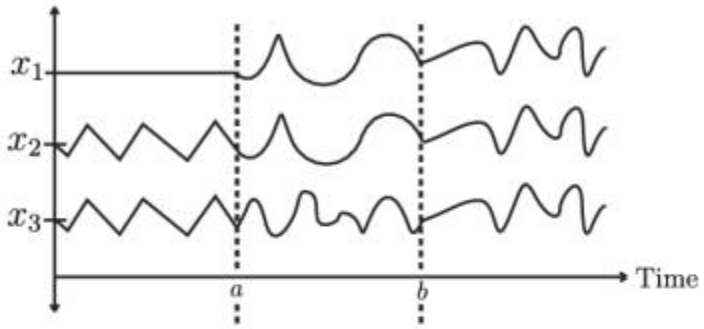
NỘI DUNG

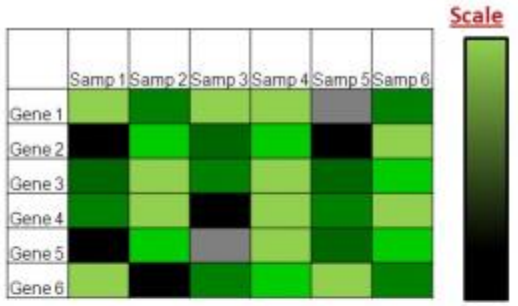
1. Thuộc tính của mạng (*Network Properties*)
2. Mô hình đồ thị ngẫu nhiên (*Random Graph Model*)
3. **Xây dựng mạng (*Network Construction*)**
4. Mô típ mạng (*Network Motifs*)
5. Lan truyền thông tin (*Diffusion Process – SIR*)

3. XÂY DỰNG MẠNG

Dòng dữ liệu (raw data) thường không phải là mạng.

- Ma trận đặc trưng, bảng quan hệ, chuỗi thời gian, kho tài liệu, tập dữ liệu hình ảnh, …



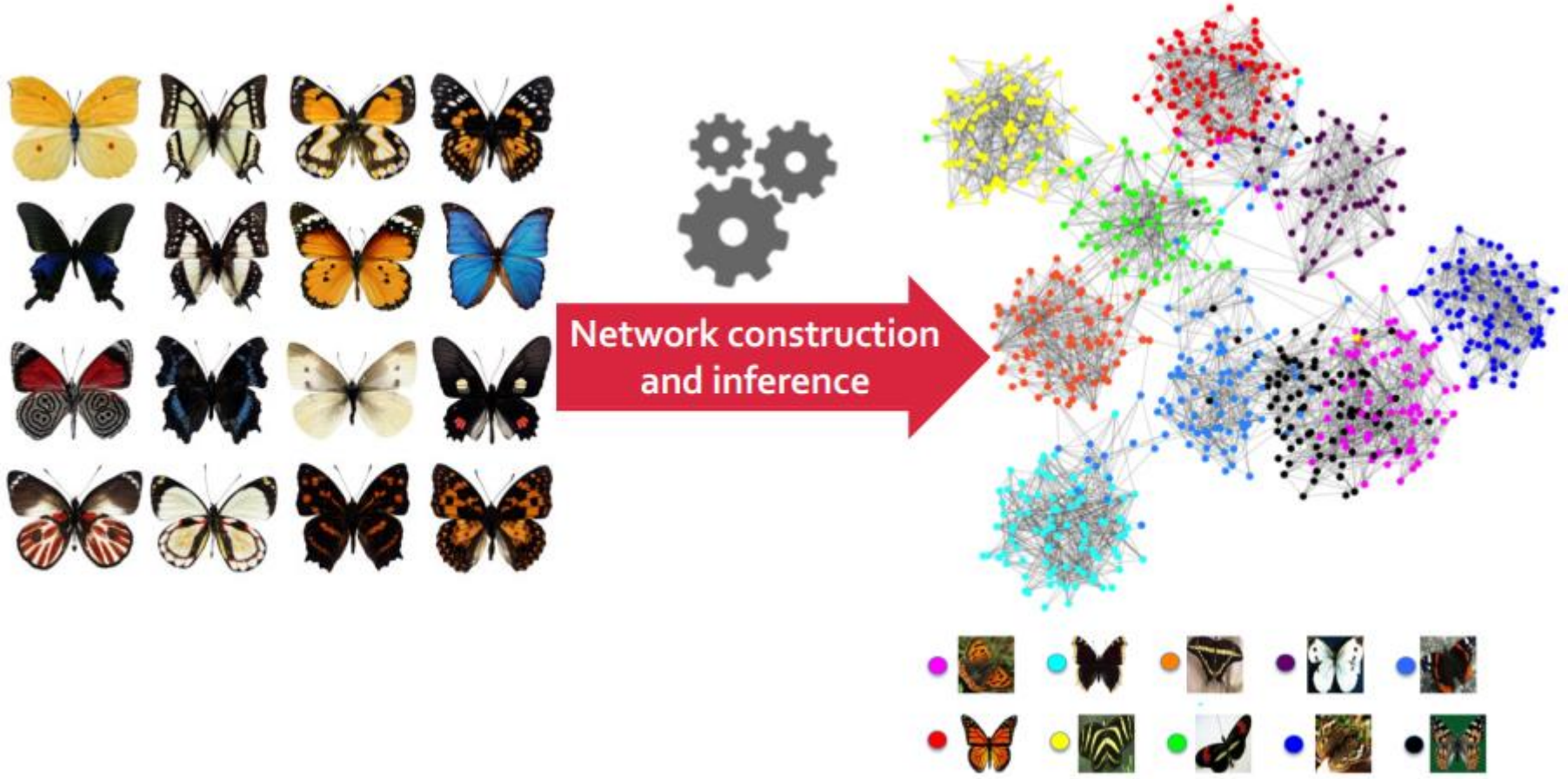


	1	2	3	4	5	6	7	8	9
man	1	0	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0	0
boy	0	0	1	0	0	0	0	0	0
girl	0	0	0	1	0	0	0	0	0
prince	0	0	0	0	1	0	0	0	0
princess	0	0	0	0	0	1	0	0	0
queen	0	0	0	0	0	0	1	0	0
king	0	0	0	0	0	0	0	1	0
monarch	0	0	0	0	0	0	0	0	1

3. XÂY DỰNG MẠNG

Làm thế nào để xây dựng mạng?

- Cách suy luận và xây dựng mạng từ dữ liệu thô (raw data)?



3. XÂY DỰNG MẠNG

3. XÂY DỰNG MẠNG

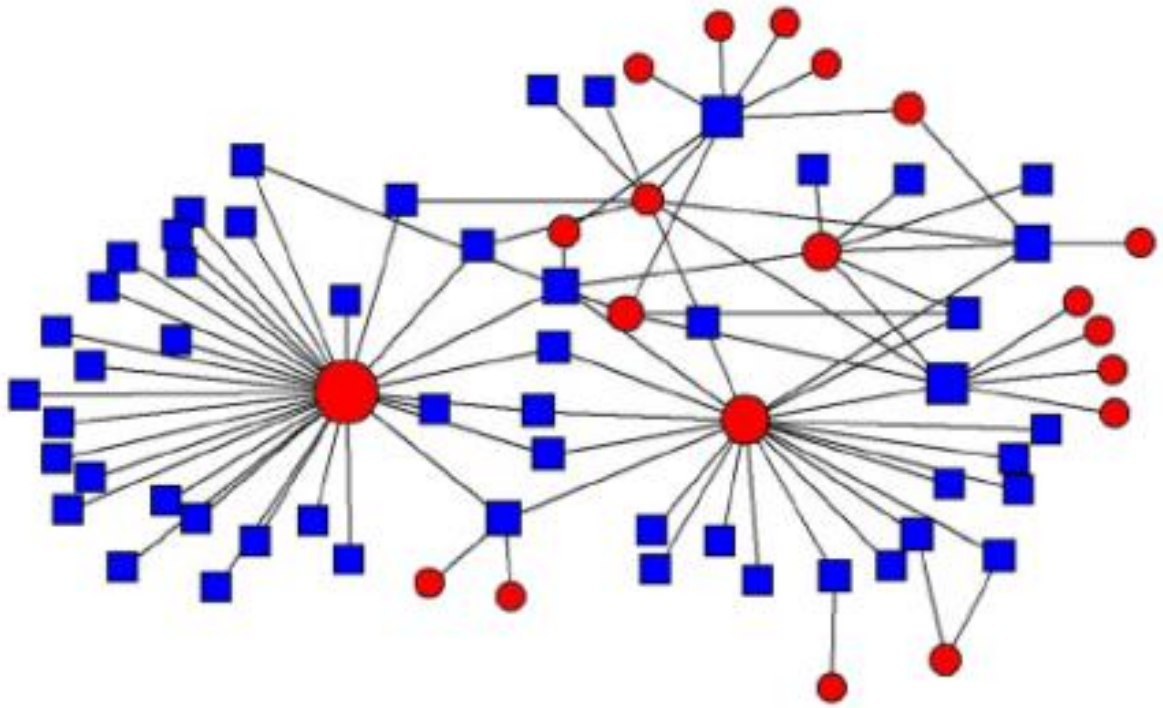
Chuyển đổi mạng

- Hầu hết thời gian, khi tạo một mạng, **tất cả các nút đại diện cho các đối tượng cùng loại**:
  - Con người trong mạng xã hội, điểm dừng xe buýt trong mạng lưới tuyến đường, gen trong mạng lưới gen, …
- **Mạng đa phân (*mạng lưỡng phân và mạng k-phân*) có nhiều loại đỉnh, nơi các cạnh chỉ đi từ loại này sang loại khác.**
  - Mạng sinh viên lưỡng phân: Sinh viên ⇔ Dự án nghiên cứu
  - Mạng điện ảnh 3-phân: Diễn viên ⇔ Phim ⇔ Công ty điện ảnh

3. XÂY DỰNG MẠNG

Mạng xã hội lưỡng phân

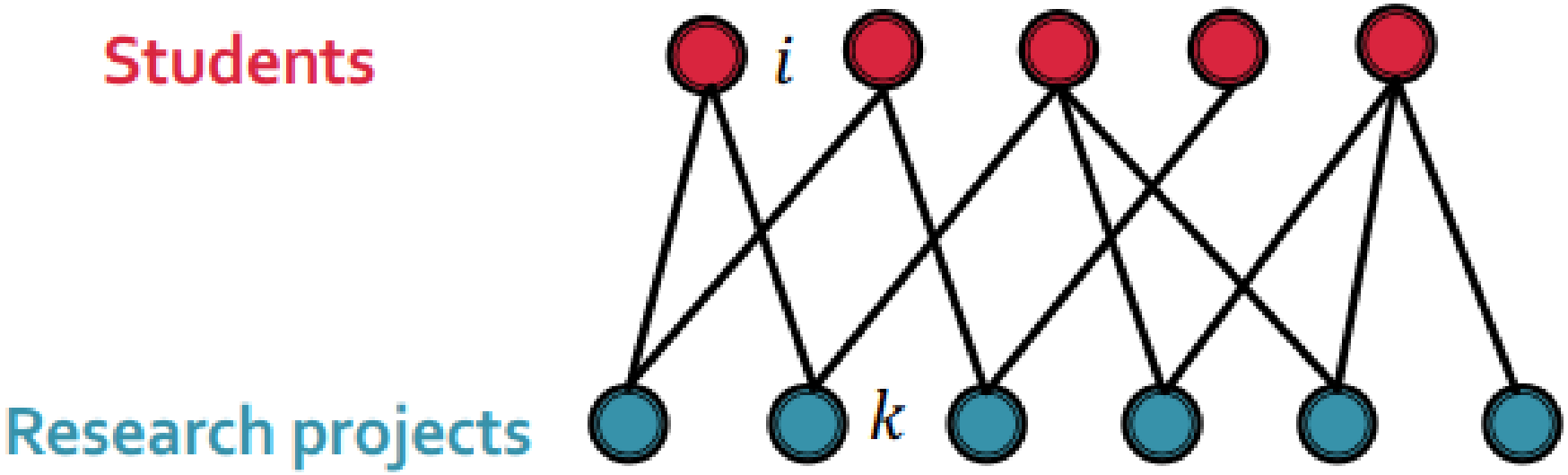
- Hình vuông màu xanh đại diện cho mọi người và hình tròn màu đỏ đại diện cho tổ chức.



3. XÂY DỰNG MẠNG

Mạng xã hội lưỡng phân

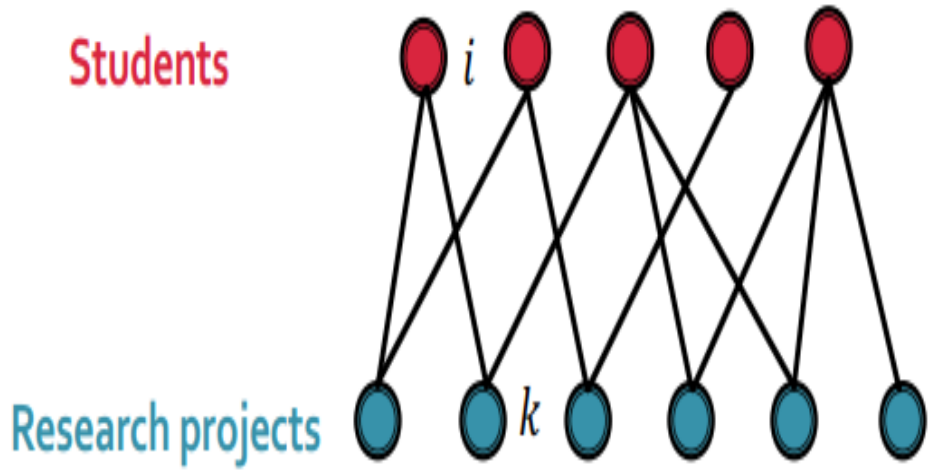
- Ví dụ: Mạng sinh viên – dự án lưỡng phân
  - **Cạnh**: Sinh viên *i* tham gia vào dự án nghiên cứu *k*.



3. XÂY DỰNG MẠNG

Mạng sinh viên – dự án lưỡng phân

- Hai mạng đơn trong mạng sinh viên – dự án lưỡng phân:
  - **Mạng sinh viên**: Sinh viên được liên kết nếu họ làm việc cùng nhau trong một hoặc nhiều dự án.
  - **Mạng dự án**: Các dự án nghiên cứu được liên kết nếu một hoặc nhiều sinh viên làm việc trên cả hai dự án.

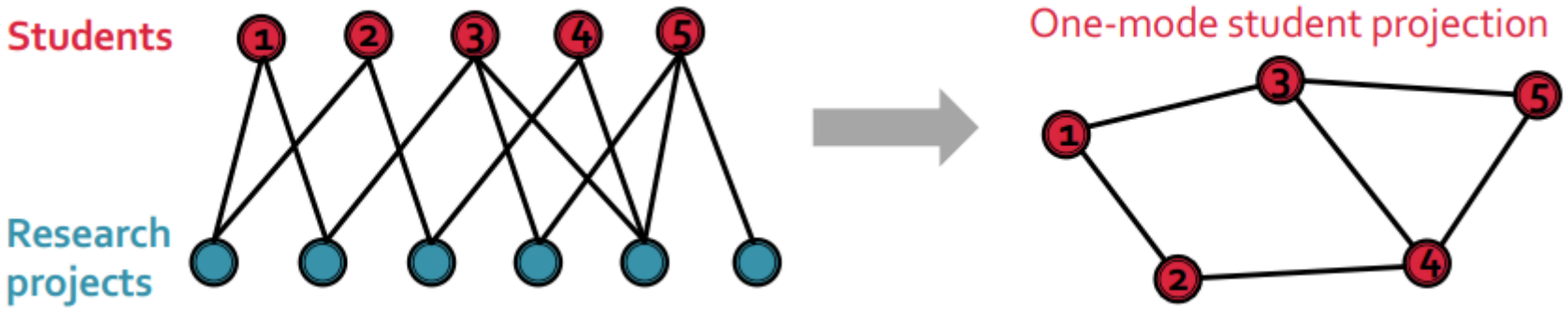


$\Rightarrow$ ***Mạng k-phân có k phép chiếu mạng đơn***.

3. XÂY DỰNG MẠNG

Mạng sinh viên – dự án lưỡng phân

- Xét những sinh viên **3**, **4** và **5** được kết nối **trong một tam giác**.
  - Tam giác có thể là kết quả của:
    - TH1: Từng cặp sinh viên làm việc chung trong những dự án khác nhau.
    - TH2: Cả 3 sinh viên làm việc chung trong một dự án.
- **Các phép chiếu mạng loại bỏ một số thông tin.**
  - Không thể phân biệt giữa TH1 và TH2 chỉ bằng cách nhìn vào hình chiếu.



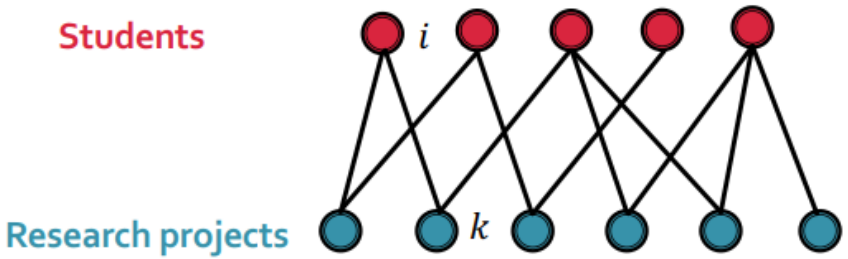
3. XÂY DỰNG MẠNG

Mạng sinh viên – dự án lưỡng phân

- Phép chiếu lên mạng sinh viên:
  - Sinh viên i và j làm việc cùng nhau tương đương với **độ dài đường đi giữa đỉnh i và j trong mạng lưỡng phân là 2**.
- Gọi **C là ma trận kề** của mạng sinh viên – dự án:

$$C_{ik} = \begin{cases} 1 \\ 0 \end{cases}$$

  - 1: Nếu i làm việc trong dự án k.
  - 0: Trái lại.



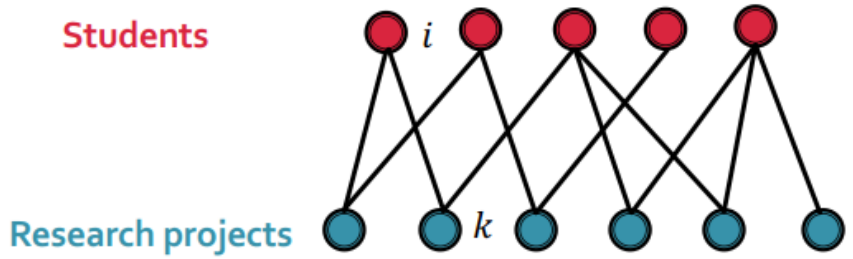
- **C là một ma trận nhị phân không đối xứng $n \times m$**
  - n là tổng số sinh viên.
  - m là tổng số dự án.

3. XÂY DỰNG MẠNG

Mạng sinh viên – dự án lưỡng phân

- **Ý tưởng**: Sử dụng C để xây dựng nhiều mạng đơn khác nhau.
- **<u>Mạng sinh viên có trọng số</u>**:

$$B_{ij} = \begin{cases} w_{ij} \\ 0 \end{cases}$$



  - w_{ij}: Tổng số dự án mà i và j cùng tham gia.
  - 0: Trái lại.
- $B_{ij} = \sum_{k=1}^{m} C_{ik}C_{jk}$: **Số lượng đường đi có độ dài là 2** kết nối sinh viên i và j trong mạng lưỡng phân.
- $B = CC^T$ và B_{ii} đại diện cho tổng số dự án mà sinh viên i tham gia.

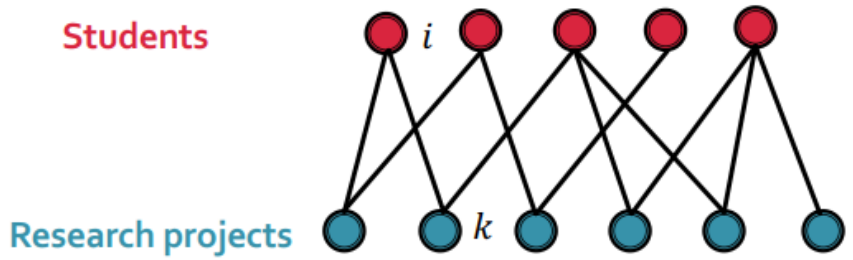
3. XÂY DỰNG MẠNG

Mạng sinh viên – dự án lưỡng phân

- **Ý tưởng**: Sử dụng C để xây dựng nhiều mạng đơn khác nhau.
- Tương tự, ta có **mạng dự án có trọng số**:

$$D_{kl} = \begin{cases} w_{kl} \\ 0 \end{cases}$$

  - w_{kl}: Tổng số sinh viên tham gia vào k và l.
  - 0: Trái lại.



- $D_{kl} = \sum_{i=1}^{n} C_{ik} C_{il}$: **Số lượng đường đi có độ dài là 2** kết nối dự án k và l trong mạng lưỡng phân.
- $D = C^T C$ và D_{kk} đại diện cho tổng số sinh viên tham gia vào dự án k.

3. XÂY DỰNG MẠNG

- **Dựa vào *B* và *D* có thể xây dựng 2 mạng đơn khác nhau.**
- **Xây dựng mạng dựa vào các độ đo tương tự của đỉnh:**
  - **Số hàng xóm chung (Common neighbors)**: Số đỉnh chung cùng liền kề với hai đỉnh.
    - **Mạng sinh viên**: i và j sẽ liên kết nếu họ làm việc cùng nhau trong r dự án hoặc hơn ($B_{ij} \geq r$).
    - **Mạng dự án**: k và l sẽ liên kết nếu r sinh viên hoặc hơn cùng tham gia vào cả 2 dự án ($D_{kl} \geq r$).
  - **Độ đo Jaccard (Jaccard index)**: Tỷ lệ phần chung trong tập hợp đỉnh liền kề hoàn chỉnh của hai đỉnh.
    - **Mạng sinh viên**: i và j sẽ liên kết nếu họ làm việc cùng nhau trong ít nhất p phần trong những dự án của họ ($B_{ij}/(B_{ii} + B_{jj} - B_{ij}) \geq p$).
    - **Mạng dự án**: k và l sẽ liên kết nếu có ít nhất p sinh viên cùng làm việc trong một bộ phận của hai dự án ($D_{kl}/(D_{kk} + D_{ll} - D_{kl}) \geq p$).

3. XÂY DỰNG MẠNG

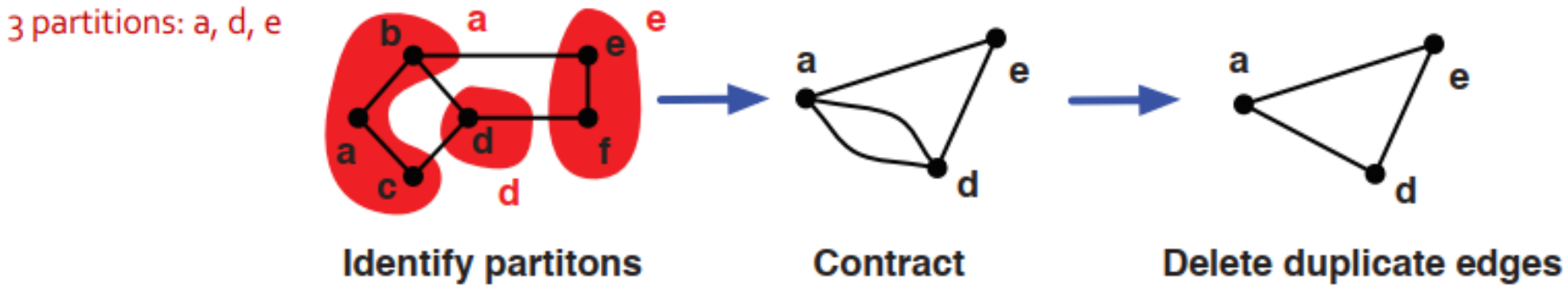
Rút gọn đồ thị

- Tính toán các thuộc tính của mạng bằng phương pháp tính toán song song theo nguyên tắc chia để trị.
- Ý tưởng:
  - Rút gọn đồ thị thành đồ thị nhỏ hơn.
  - Đệ quy trên đồ thị nhỏ hơn.
  - Sử dụng kết quả từ quá trình đệ quy với đồ thị ban đầu để tính kết quả mong muốn.
- **Làm thế nào để rút gọn một đồ thị?**

3. XÂY DỰNG MẠNG

Thuật toán rút gọn đồ thị

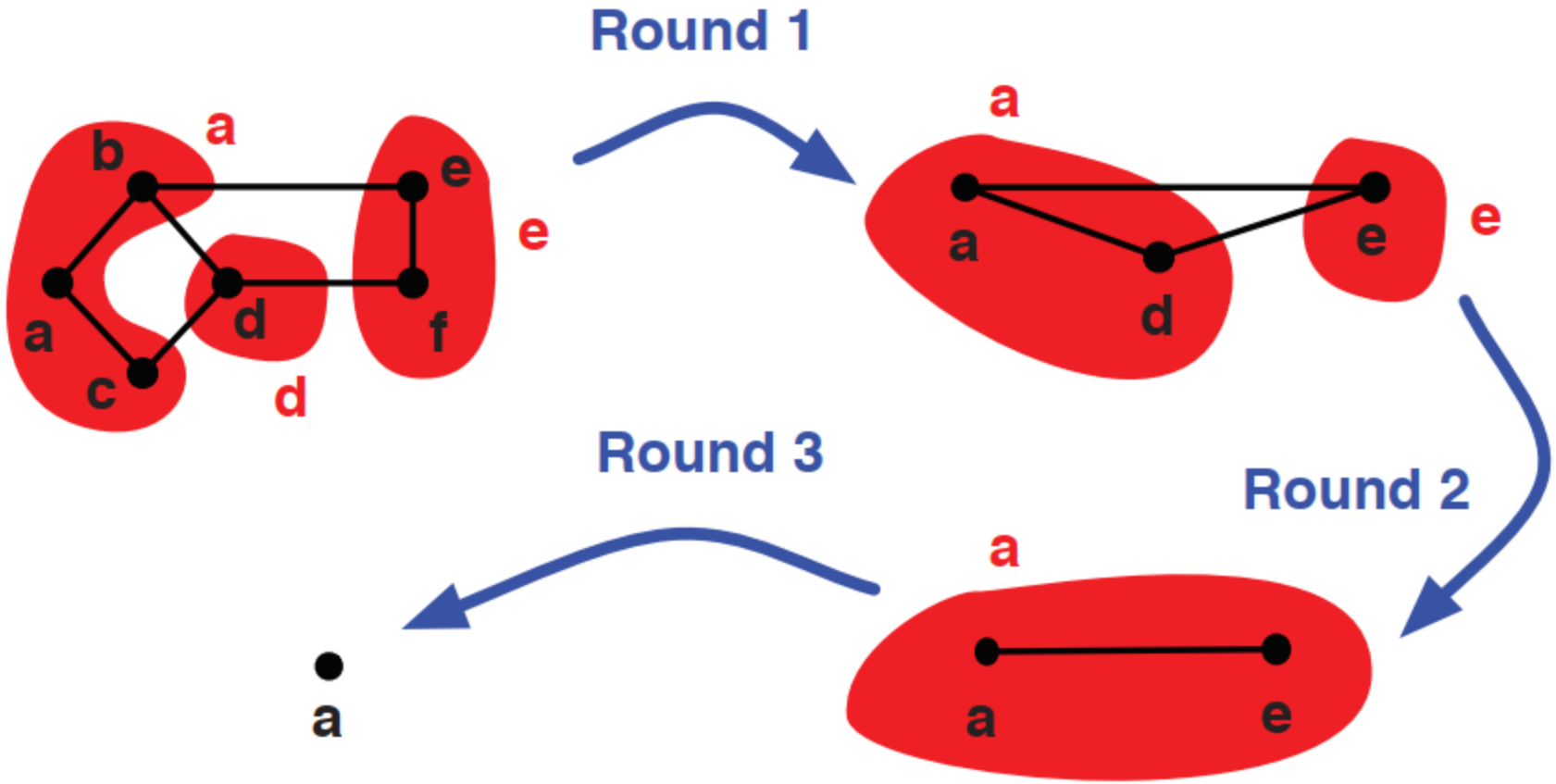
- Bắt đầu với **dữ liệu input là đồ thị G**:
  1. Chọn một phân vùng đỉnh của G để rút gọn. Những phần này rời rạc và chúng bao gồm tất cả các đỉnh trong G.
  2. Rút gọn mỗi phân vùng thành một đỉnh duy nhất (*một siêu đỉnh*).
  3. Hủy các cạnh bên trong một phân vùng.
  4. Định tuyến lại các cạnh đến các siêu đỉnh tương ứng.
  5. Gán G thành đồ thị nhỏ hơn mới và lặp lại.
- Ví dụ: Một vòng lặp của quá trình rút gọn đồ thị



3. XÂY DỰNG MẠNG

Rút gọn đồ thị

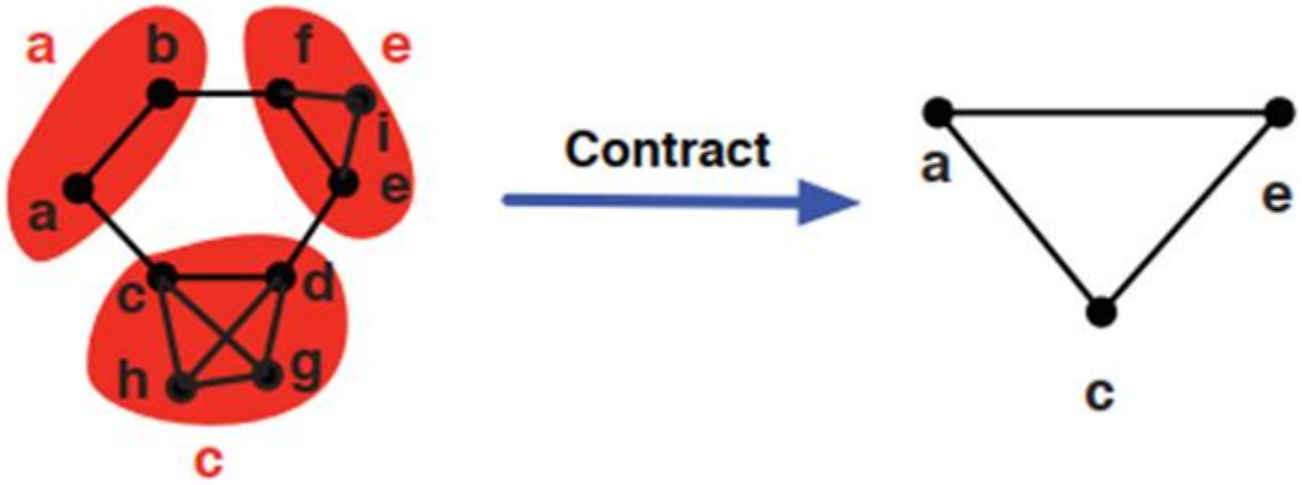
- Rút gọn một đồ thị thành một nút đơn trong 3 vòng lặp:



3. XÂY DỰNG MẠNG

Rút gọn đồ thị

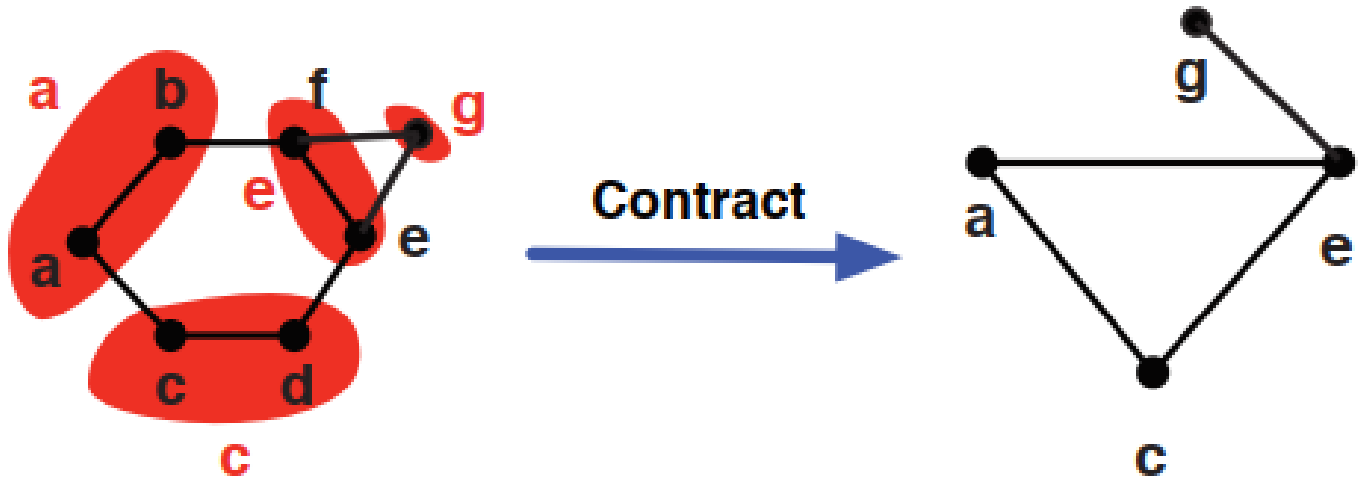
- Các phân vùng phải rời rạc và bao gồm tất cả các đỉnh trong G.
- Ba kiểu phân vùng đỉnh:
  - Mỗi phân vùng là **một nhóm (cực đại) các đỉnh**.



3. XÂY DỰNG MẠNG

Rút gọn đồ thị

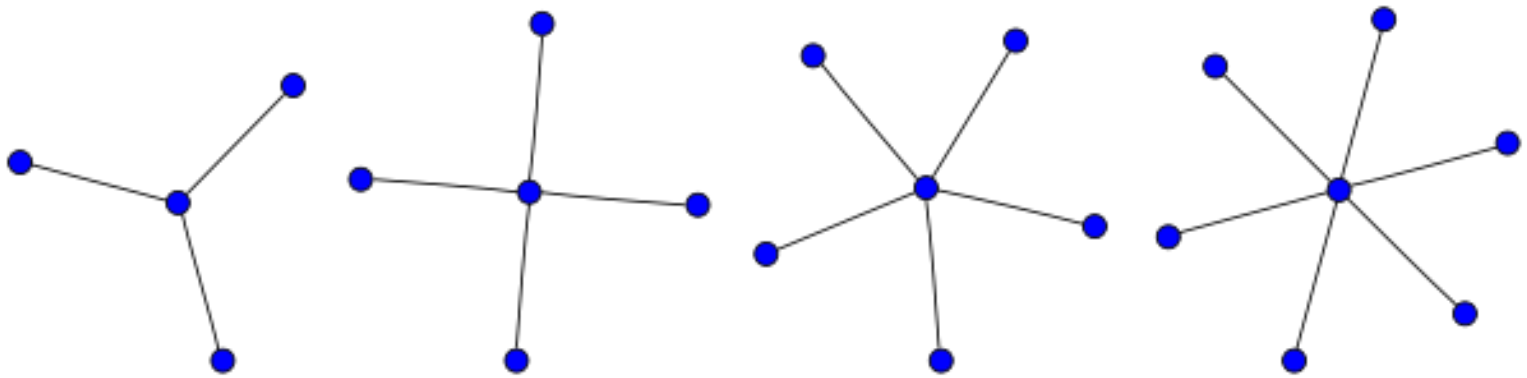
- Các phân vùng phải rời rạc và bao gồm tất cả các đỉnh trong G.
- Ba kiểu phân vùng đỉnh:
  - Mỗi phân vùng là **một đỉnh đơn hoặc hai đỉnh kết nối với nhau**.



3. XÂY DỰNG MẠNG

Rút gọn đồ thị

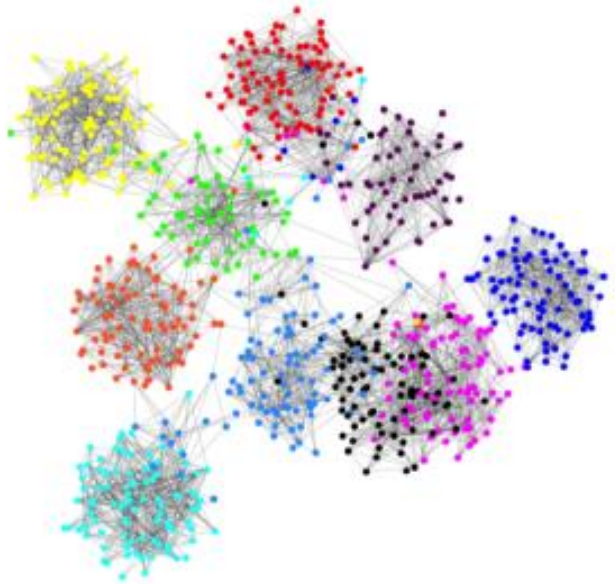
- Các phân vùng phải rời rạc và bao gồm tất cả các đỉnh trong G.
- Ba kiểu phân vùng đỉnh:
  - Mỗi phân vùng là **một ngôi sao của các đỉnh**, …



3. XÂY DỰNG MẠNG

Đồ thị K-Nearest Neighbor

- **Một tập các đối tượng V là đồ thị có hướng với tập đỉnh V.**
- Mỗi **cạnh trong đồ thị K-NN thể hiện sự tương đồng** giữa các đối tượng với một số độ đo tương tự đã cho:
  - **Độ đo Cosine** cho dữ liệu text.
  - l_2 **khoảng cách của đặc trưng CNN-derived** cho dữ liệu hình ảnh.

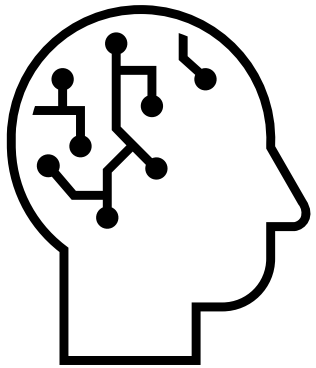


3. XÂY DỰNG MẠNG

Xây dựng đồ thị K-Nearest Neighbor là một hoạt động quan trọng:

- **Hệ thống khuyến nghị**: Kết nối người dùng với các mẫu xếp hạng sản phẩm tương tự. Sau đó, đưa ra đề xuất dựa trên các "hàng xóm" trên biểu đồ của người dùng.
- **Hệ thống truy xuất tài liệu**: Kết nối các tài liệu có nội dung tương tự, trả lời nhanh các truy vấn đầu vào của người dùng.
- Các vấn đề khác trong **gom cụm, trực quan hóa, truy xuất thông tin, khai thác dữ liệu, …**
- **Phương pháp K-NNGs giúp chúng ta có thể sử dụng các phương thức khai thác mạng trên tập dữ liệu không có cấu trúc đồ thị rõ ràng.**

3. XÂY DỰNG MẠNG



Giải mã mạng và
suy luận

3. XÂY DỰNG MẠNG

- **Mạng thể hiện sự phụ thuộc giữa các đối tượng**:
  - Đồng tác giả giữa các nhà khoa học.
  - Tình bạn giữa mọi người.
  - Đối tượng nào ăn đối tượng nào trong lưới thức ăn.
  - Liên kết giữa các gốc phân tử.
  - Mối quan hệ điều hòa giữa các gen.
- Phụ thuộc gián tiếp xảy ra do tác động bắc cầu của mối tương quan.
- **Vấn đề**: Làm thế nào để tách phần phụ thuộc trực tiếp khỏi phần phụ thuộc gián tiếp?

3. XÂY DỰNG MẠNG

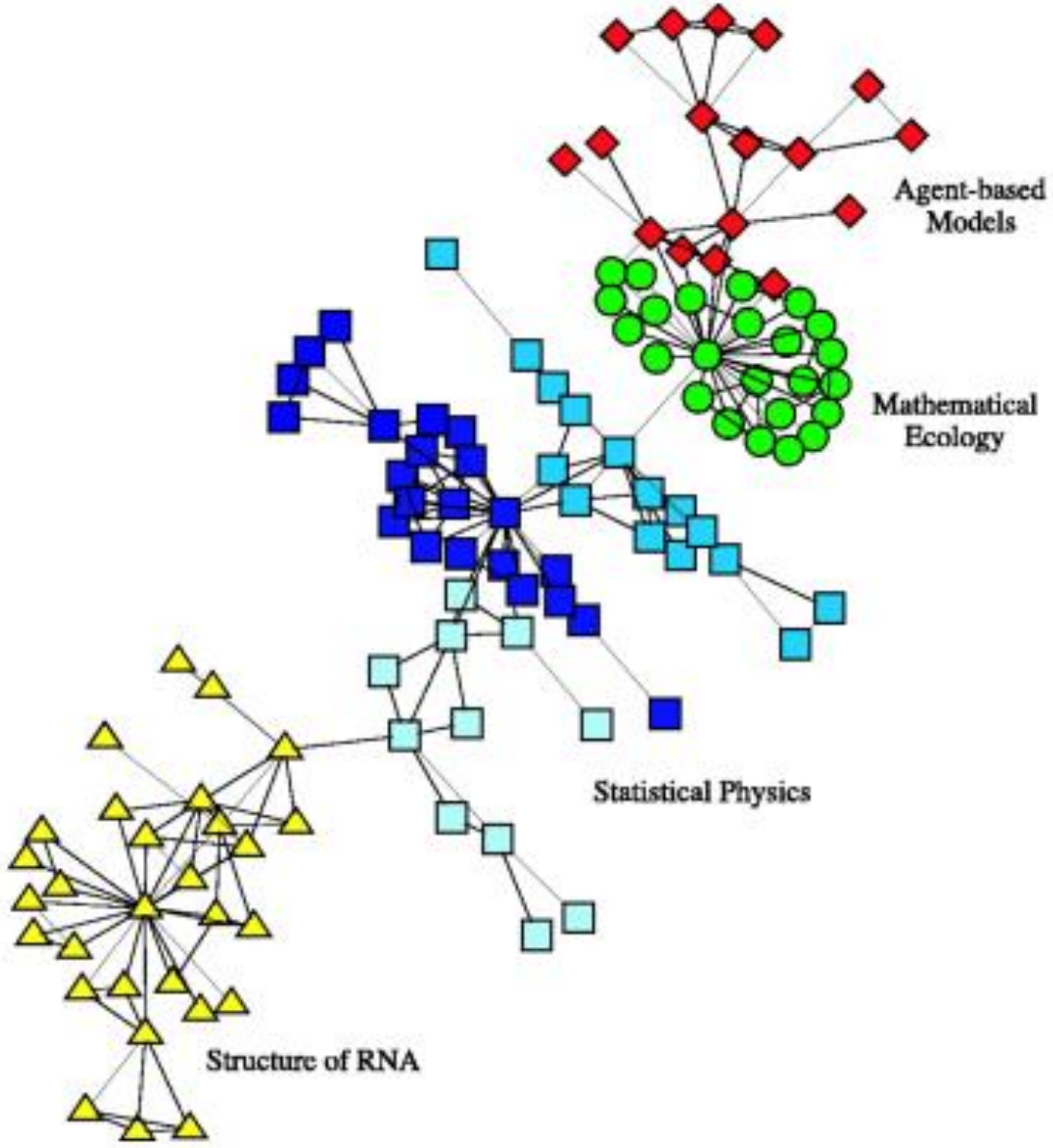
Ứng dụng: Mạng đồng tác giả.

- **Mục tiêu**: Phân biệt sự hợp tác **mạnh** và **yếu** giữa các nhà khoa học.
- Các điểm mạnh ràng buộc cộng tác **phụ thuộc vào chi tiết xuất bản**, chẳng hạn như:
  - Tổng số bài báo mỗi cặp nhà khoa học đã hợp tác.
  - Số lần đồng tác giả trên mỗi bài báo.

3. XÂY DỰNG MẠNG

Ứng dụng: Mạng đồng tác giả.

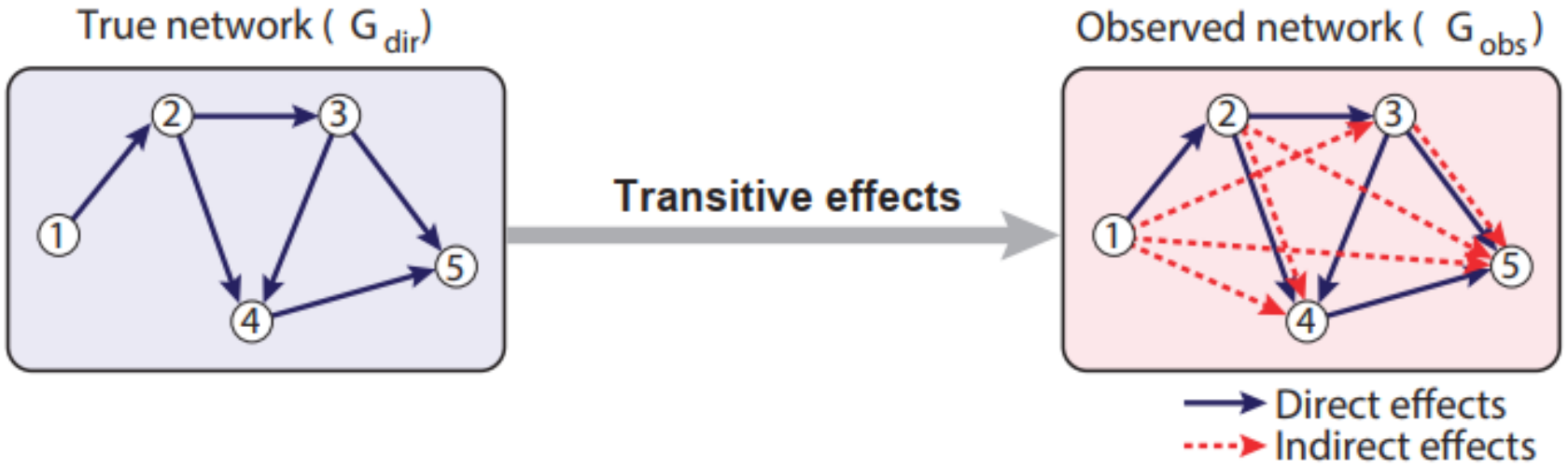
- **Mục tiêu**: Phân biệt sự hợp tác **mạnh** và **yếu** giữa các nhà khoa học.
- Độ bền của mối quan hệ rất quan trọng đối với:
  - Giới thiệu bạn bè và đồng nghiệp.
  - Nhận biết xung đột lợi ích.
  - Đánh giá đóng góp của các tác giả cho các nhóm.



3. XÂY DỰNG MẠNG

Mạng quan sát (Observed network)

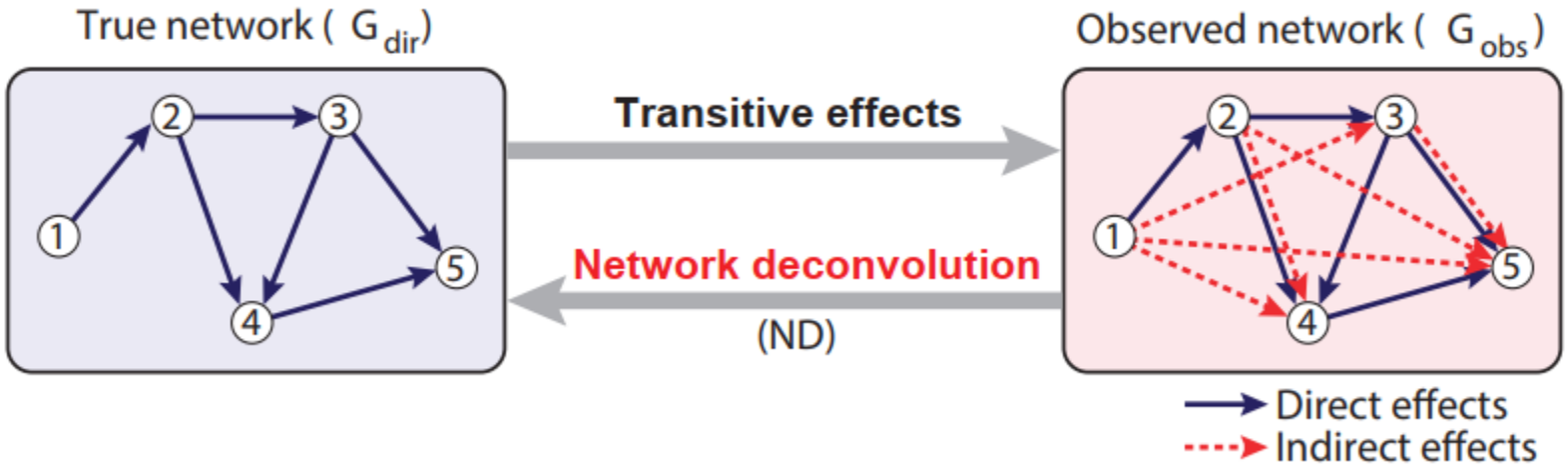
- **Kết hợp sự phụ thuộc trực tiếp và phụ thuộc gián tiếp.**
- Các cạnh gián tiếp có thể là do các tương tác bậc cao (Ví dụ: 1 → 4).
- Mỗi cạnh có thể chứa cả thành phần trực tiếp và gián tiếp (Ví dụ: 2 → 4).



3. XÂY DỰNG MẠNG

Giải mã mạng

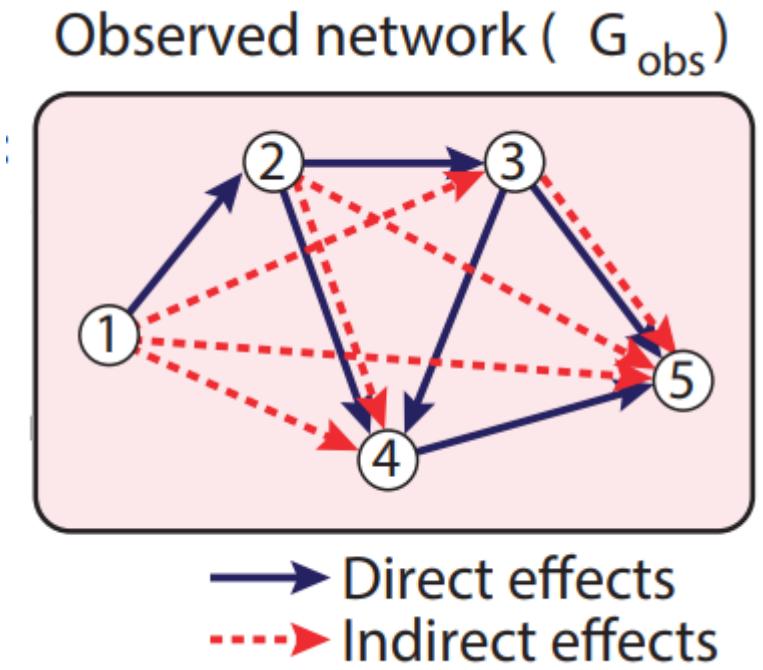
- **Mục tiêu**: Đảo ngược tác động của luồng thông tin bắc cầu trên tất cả các con đường gián tiếp.
- Khôi phục mạng trực tiếp thực (các cạnh màu xanh G_{dir}) dựa trên mạng quan sát (kết hợp của các cạnh màu xanh và màu đỏ G_{obs}).



3. XÂY DỰNG MẠNG

Giải mã mạng

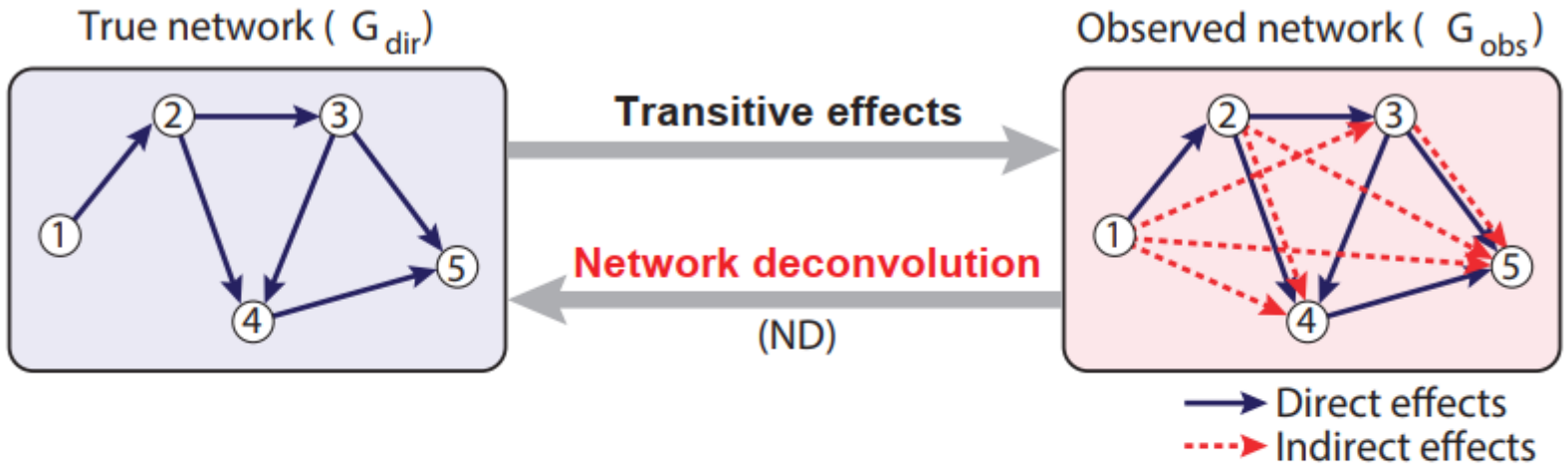
- Các cạnh trực tiếp trong một mạng có thể dẫn đến các mối quan hệ gián tiếp:
  - Luồng thông tin bắc cầu.
- Các phụ thuộc gián tiếp có thể là kết hợp:
  - Cả phụ thuộc trực tiếp và gián tiếp (Ví dụ: 2 → 4).
  - Nhiều phụ thuộc gián tiếp cùng các đường dẫn khác nhau (Ví dụ: 2 → 3 → 5, 2 → 4 → 5).



3. XÂY DỰNG MẠNG

Giải mã mạng

- Sử dụng biểu thức dạng đóng cho G_{obs} để khôi phục mạng trực tiếp thực G_{dir}.



Indirect effects Series closed form

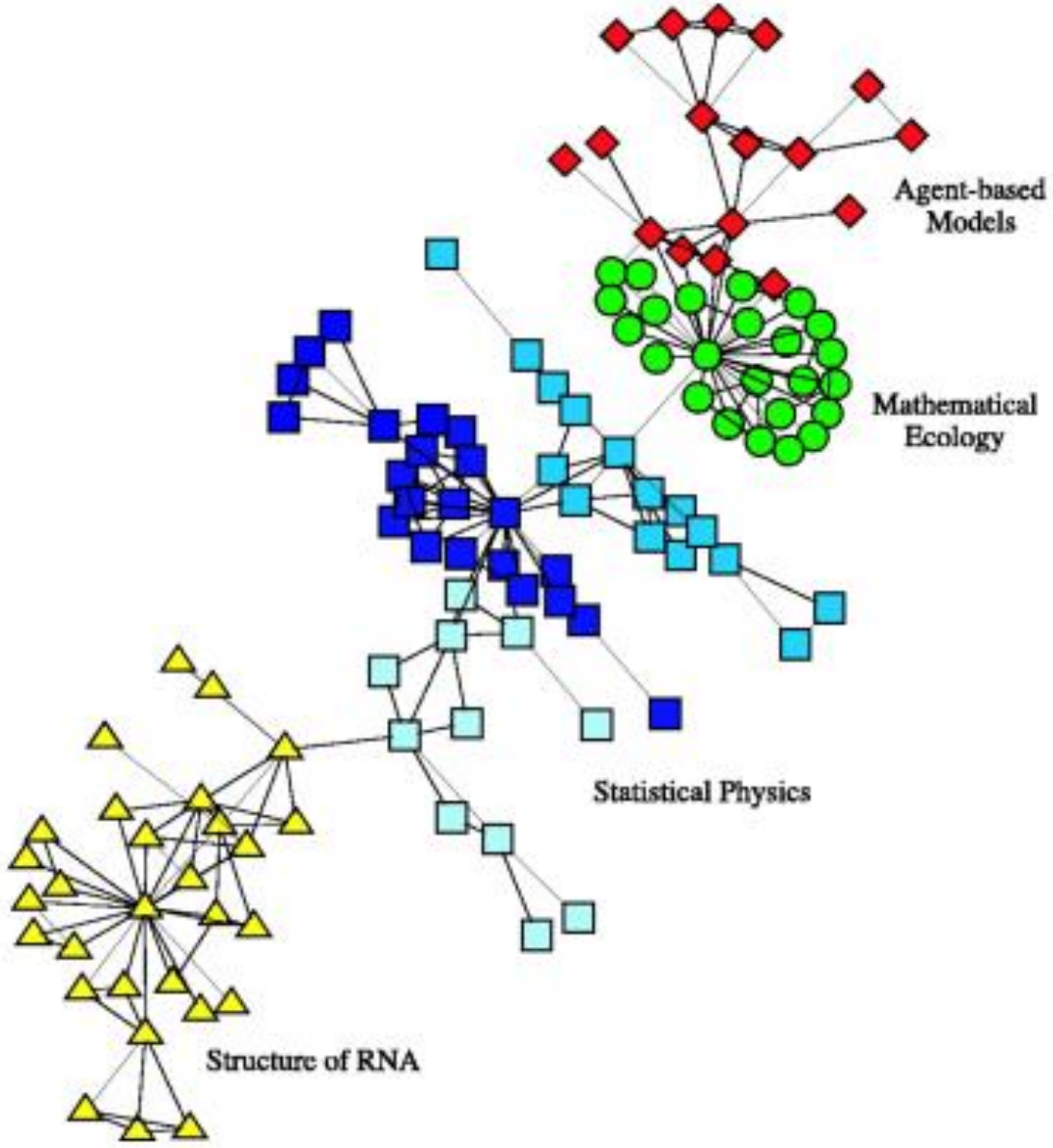
Transitive closure: $G_{obs} = G_{dir} + G_{dir}^2 + G_{dir}^3 + \ldots = G_{dir}(I - G_{dir})^{-1}$

Network deconvolution: $G_{dir} = G_{obs}(I + G_{obs})^{-1}$

3. XÂY DỰNG MẠNG

Ứng dụng: Mạng đồng tác giả.

- **Mục tiêu**: Phân biệt sự hợp tác **mạnh** và **yếu** giữa các nhà khoa học.
- Độ bền của mối quan hệ rất quan trọng đối với:
  - Giới thiệu bạn bè và đồng nghiệp.
  - Nhận biết xung đột lợi ích.
  - Đánh giá đóng góp của các tác giả cho các nhóm.



3. XÂY DỰNG MẠNG

- **Dữ liệu**: Mạng không trọng số của các nhà khoa học làm việc trong lĩnh vực khoa học.
  - Hai tác giả được liên kết nếu họ đồng tác giả ít nhất một bài báo.
- **Thiết lập**: Áp dụng ND trên mạng lưới đồng tác giả.
  - ND trả về một mạng có trọng số có:
    - Đảo ngược tác động của luồng thông tin bắc cầu.
    - Trọng số thể hiện mức độ hợp tác suy ra từ các tương tác trực tiếp.
  - Đầu ra: **Xếp hạng các cạnh đồng tác giả theo trọng số** do ND ấn định.
- **Dữ liệu thật cơ bản**:
  - **Các điểm mạnh cộng tác thực sự được tính bằng cách tính tổng số bài báo có đồng tác giả và giảm trọng số của mỗi bài báo theo số lượng đồng tác giả bổ sung.**
  - Tính toán mối tương quan giữa trọng số do ND chỉ định và sức mạnh cộng tác thực sự.

3. XÂY DỰNG MẠNG

3. XÂY DỰNG MẠNG

Chuyển đổi mạng

Xây dựng mạng từ dữ liệu thô.

Rút gọn đồ thị phức tạp.

K-NNGs

Khai thác mạng trên dữ liệu không có cấu trúc rõ ràng.

Giải mã mạng

Tạo ra dữ liệu thật cơ bản (giảm tác động của luồng thông tin bắc cầu).

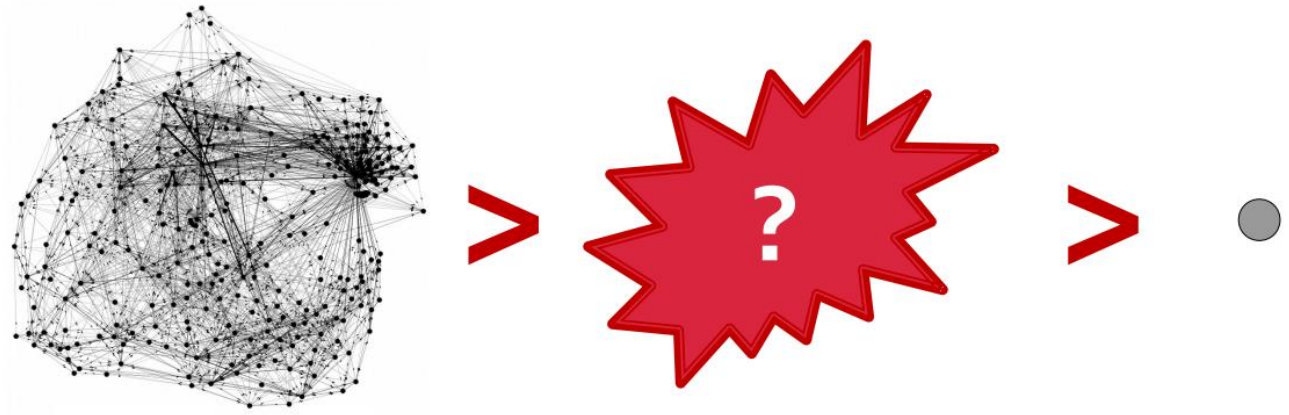
NỘI DUNG

1. Thuộc tính của mạng (*Network Properties*)
2. Mô hình đồ thị ngẫu nhiên (*Random Graph Model*)
3. Xây dựng mạng (*Network Construction*)
4. **Mô típ mạng (*Network Motifs*)**
5. Lan truyền thông tin (*Diffusion Process – SIR*)

4. MÔ TÍP MẠNG

Chỉ số mạng

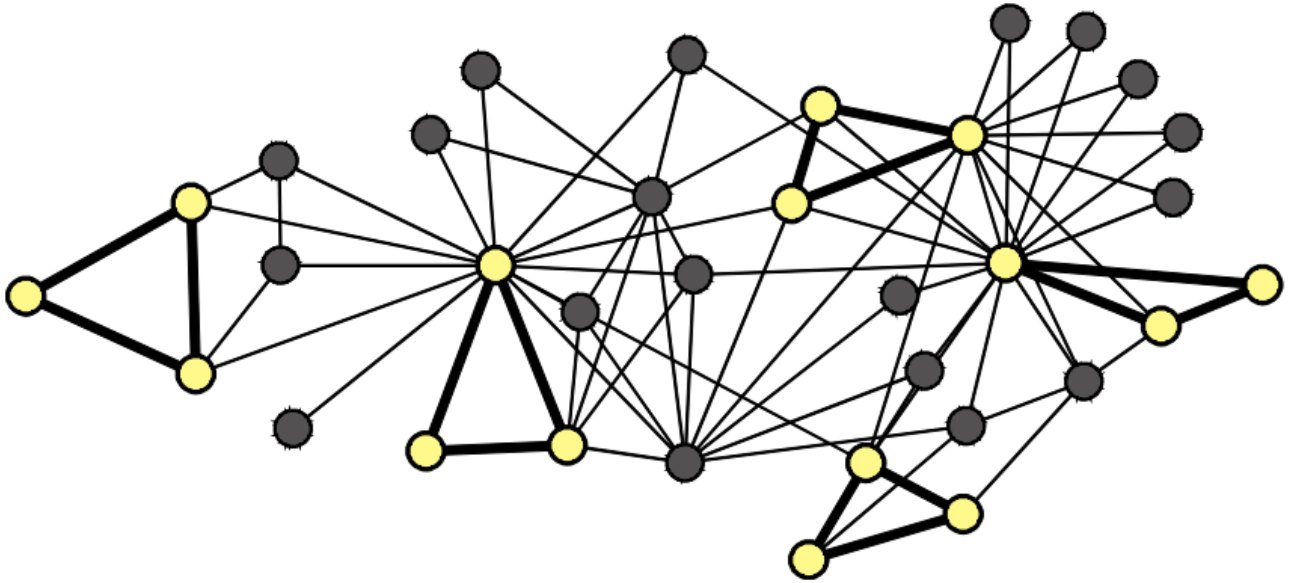
- **Nhiều chỉ số về đỉnh:**
  - Bậc của đỉnh, độ đo PageRank, chỉ số phân cụm của đỉnh, …
- **Nhiều chỉ số về mạng:**
  - Đường kính, số cụm, kích thước của thành phần khổng lồ, …
- Ý nghĩa giữa các chỉ số này là gì?
  - **Mô tả đặc tính theo tỷ lệ trung bình của mạng:**

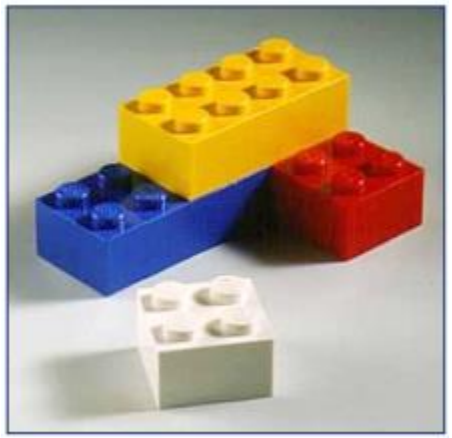


4. MÔ TÍP MẠNG

Mạng con

- **Mạng con**, hoặc **đồ thị con**, là các khối xây dựng nên mạng.
- Mạng con có sức mạnh để mô tả và phân biệt các mạng.





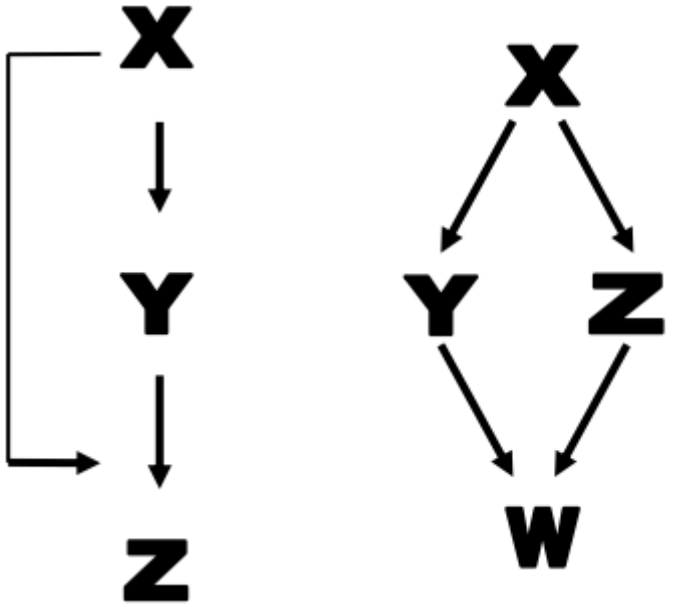
4. MÔ TÍP MẠNG

- **Mô típ mạng**: "lặp đi lặp lại, các mô hình kết nối quan trọng".
- Cách xác định mô típ mạng:
  - **Mô hình**: đồ thị con.
  - **Định kỳ**: được tìm thấy nhiều lần với tần suất cao.
  - **Kết nối quan trọng**: xuất hiện thường xuyên hơn dự kiến.

4. MÔ TÍP MẠNG

Tại sao chúng ta cần tìm hiểu mô típ mạng?

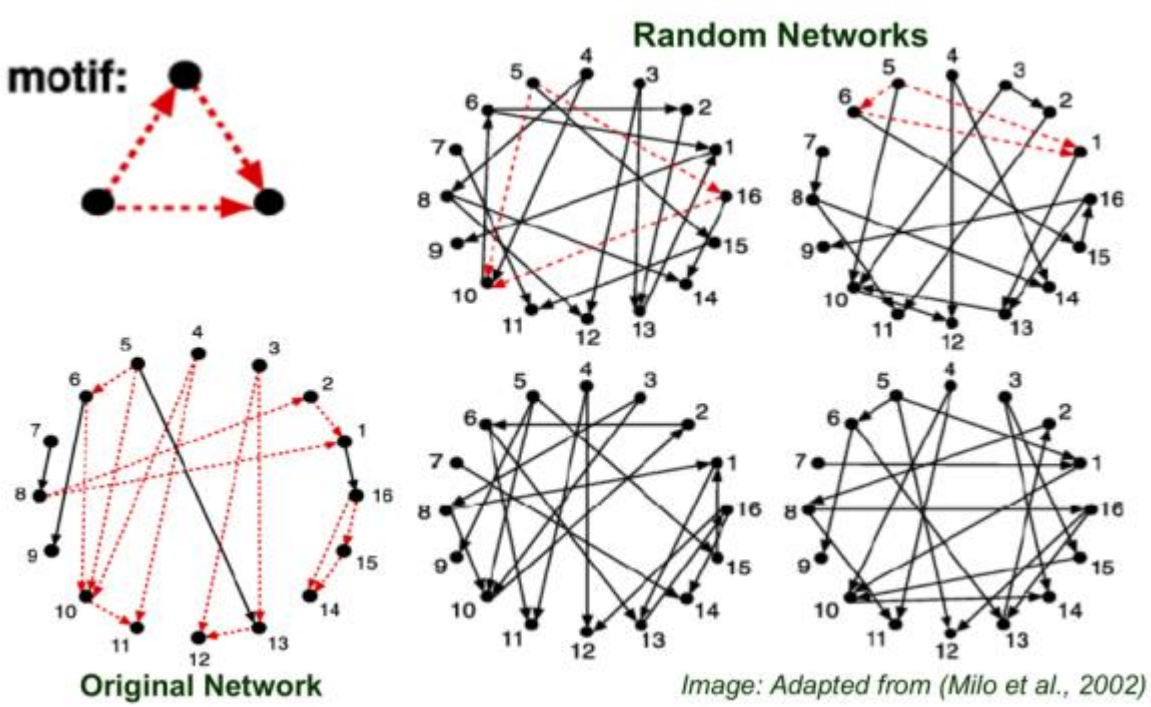
- Giúp chúng ta hiểu cách mạng hoạt động.
- Giúp chúng ta dự đoán hoạt động và phản ứng của mạng trong một tình huống nhất định.
- Ví dụ:
  - **Các vòng lặp chuyển tiếp**: được tìm thấy trong mạng lưới tế bào thần kinh.
  - **Các vòng lặp song song**: được tìm thấy trong mạng lưới thức ăn.



4. MÔ TÍP MẠNG

Dò tìm mô típ mạng

- Đếm số đồ thị con i trong mạng G^{real}.
- Đếm số đồ thị con i trong mạng ngẫu nhiên G^{rand}.
  - **Mô hình đồ thị ngẫu nhiên**: Mỗi G^{rand} có cùng số đỉnh, số cạnh và phân phối độ đo với G^{real}.



4. MÔ TÍP MẠNG

Dò tìm mô típ mạng

- Gán độ đo Z-score (*có ý nghĩa thống kê của mô típ* i) cho i:
  - $Z_i = (N_i^{real} - \bar{N}_i^{rand}) / std(N_i^{rand})$
  - Trong đó:
    - N_i^{real}: Tổng số đồ thị con i trong mạng G^{real}.
    - N_i^{rand}: Tổng số đồ thị con i trong mạng ngẫu nhiên G^{rand}.
- **Độ đo Z-score cao: Đồ thị con i là một mô típ mạng của G.**

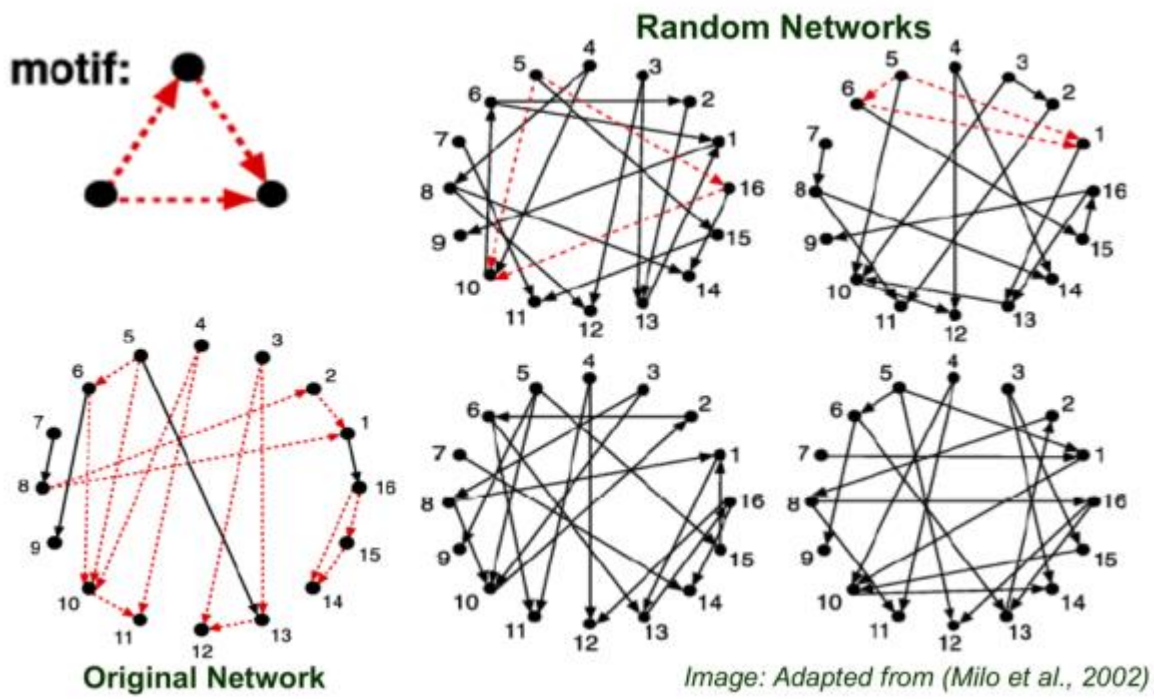


Image: Adapted from (Milo et al., 2002)

NỘI DUNG

1. Thuộc tính của mạng (*Network Properties*)
2. Mô hình đồ thị ngẫu nhiên (*Random Graph Model*)
3. Xây dựng mạng (*Network Construction*)
4. Mô típ mạng (*Network Motifs*)
5. **Lan truyền thông tin (*Diffusion Process – SIR*)**

5. LAN TRUYỀN THÔNG TIN

Lan truyền qua các mạng

- **Hành vi xếp tầng (Cascading behavior)**
  - Khi mọi người được kết nối trong mạng với nhau thì họ có thể ảnh hưởng đến hành vi và quyết định của nhau.
- **Sự lan tỏa của những đổi mới**
  - Sự lan tỏa của lý thuyết đổi mới là một giả thuyết vạch ra cách thức những tiến bộ công nghệ mới và các tiến bộ khác lan truyền khắp các xã hội và nền văn hóa, từ khi du nhập đến khi được áp dụng rộng rãi.
- **Hiệu ứng mạng**
  - Hiệu ứng mạng là một hiện tượng theo đó số lượng người hoặc người tham gia tăng lên sẽ nâng cao giá trị của một hàng hóa hoặc dịch vụ.
- **Dịch tễ**

5. LAN TRUYỀN THÔNG TIN

Tiến trình lan truyền thông tin (Diffusion Process)

- Các hành vi phân tầng từ nút này sang nút khác giống như một bệnh dịch.

Số sinh sản cơ bản

R_0 = dự kiến số ca mắc bệnh mới do một cá nhân gây ra.

- Nếu $R_0 < 1$ thì bệnh kết thúc.
- Nếu $R_0 > 1$ thì bệnh có cơ hội lây lan rộng.

Điều này có ý nghĩa gì đối với các chính sách tiêm chủng?

5. LAN TRUYỀN THÔNG TIN

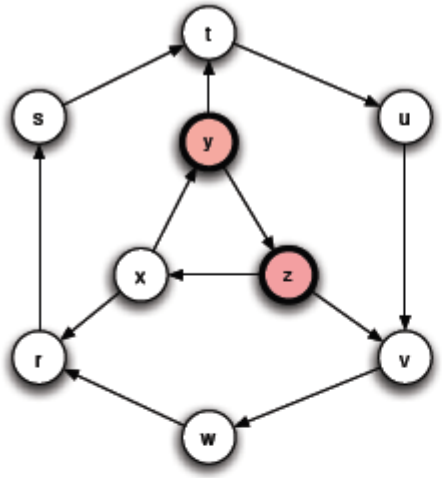
Mô hình SIR

- **Susceptible**
  - Dễ mắc bệnh: có thể mắc bệnh từ hàng xóm trong mạng.
- **Infectious**
  - Truyền nhiễm: đã mắc bệnh và có thể truyền bệnh.
- **Removed**
  - Đã loại bỏ: đã hồi phục hoặc đã chết, nhưng không lây nhiễm trong mọi trường hợp.

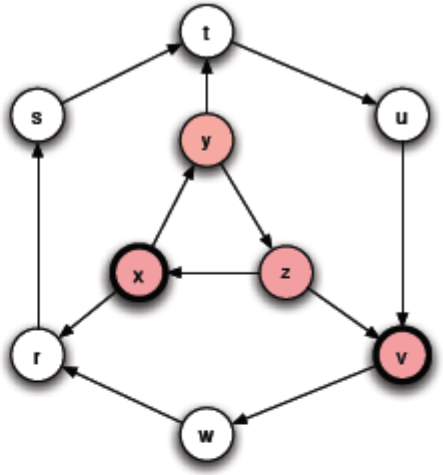
5. LAN TRUYỀN THÔNG TIN

Mô hình SIR

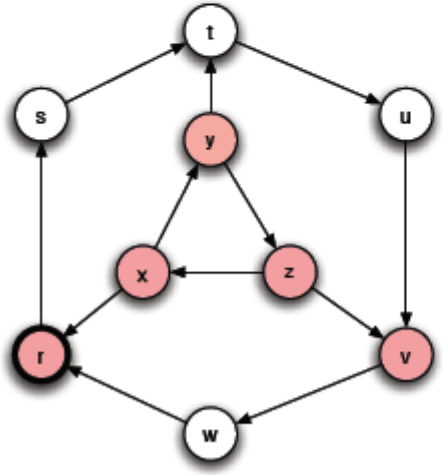
- **Ví dụ**:
  - Mỗi nút vẫn lây nhiễm (*đường viền đậm*) trong một khoảng thời gian và lây nhiễm cho hàng xóm với xác suất p.



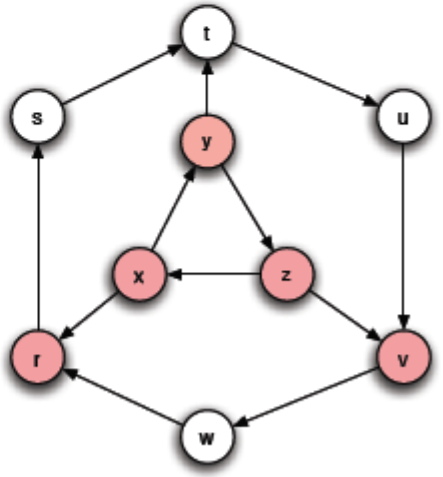
(a)



(b)



(c)



(d)


Q & A