

Mạng Xã Hội - Lab05.01: Tìm hiểu các chỉ số thống kê cơ bản sử dụng Gephi

Phan Phú Hào - 2186400335

Ngày 30 tháng 12 năm 2024

Chương 1. Độ đo cơ bản của mạng

1.1. Average Degree (Bậc trung bình)

- Định nghĩa:** Là giá trị trung bình của số lượng cạnh kết nối đến mỗi đỉnh trong mạng
- Công thức:** $\bar{k} = \frac{2|E|}{|V|}$ với $|E|$ là số cạnh và $|V|$ là số đỉnh
- Phạm vi:** $[0, |V| - 1]$
- Đánh giá:** Chỉ số càng cao thể hiện mạng càng kết nối chặt chẽ. Tuy nhiên, cần xem xét kết hợp với các chỉ số khác để đánh giá toàn diện.

1.2. Network Diameter (Đường kính mạng)

- Định nghĩa:** Là độ dài đường đi ngắn nhất lớn nhất giữa hai đỉnh bất kỳ trong mạng
- Công thức:** $D = \max_{i,j} d(i,j)$ với $d(i,j)$ là đường đi ngắn nhất giữa đỉnh i và j
- Phạm vi:** $[1, |V| - 1]$
- Đánh giá:** Đường kính nhỏ thường tốt hơn, thể hiện thông tin có thể truyền nhanh trong mạng.

1.3. Graph Density (Mật độ đồ thị)

- **Định nghĩa:** Tỷ lệ giữa số cạnh thực tế và số cạnh tối đa có thể có trong mạng
- **Công thức:** $\rho = \frac{2|E|}{|V|(|V|-1)}$ với đồ thị vô hướng
- **Phạm vi:** $[0, 1]$
- **Đánh giá:** Mật độ cao (gần 1) thể hiện mạng kết nối dày đặc, mật độ thấp (gần 0) thể hiện mạng thưa thớt.

1.4. Connected Components (Thành phần liên thông)

- **Định nghĩa:** Là số lượng các nhóm đỉnh mà trong đó có đường đi giữa mọi cặp đỉnh
- Không có công thức cụ thể, được tính bằng thuật toán DFS hoặc BFS
- **Phạm vi:** $[1, |V|]$
- **Đánh giá:** Số thành phần liên thông = 1 là tốt nhất, thể hiện mạng hoàn toàn kết nối.

1.5. Average Path Length (Độ dài đường đi trung bình)

- **Định nghĩa:** Trung bình của tất cả các đường đi ngắn nhất giữa mọi cặp đỉnh
- **Công thức:** $L = \frac{1}{n(n-1)} \sum_{i \neq j} d(i, j)$
- **Phạm vi:** $[1, D]$ với D là đường kính mạng
- **Đánh giá:** Giá trị nhỏ thể hiện mạng "nhỏ gọn", thông tin truyền đi hiệu quả.

1.6. Average Clustering Coefficient (Hệ số phân cụm trung bình)

- **Định nghĩa:** Mức độ các đỉnh láng giềng của một đỉnh kết nối với nhau
- **Công thức:** $C = \frac{1}{n} \sum_i C_i$ với $C_i = \frac{2L_i}{k_i(k_i-1)}$
 - L_i : số cạnh giữa các láng giềng của đỉnh i

- k_i : số láng giềng của đỉnh i
- **Phạm vi:** $[0, 1]$
- **Đánh giá:** Giá trị cao thể hiện xu hướng tạo cụm mạnh trong mạng.

Chương 2. Độ đo tính trung tâm

2.1. Degree Centrality

- **Định nghĩa:** Số lượng kết nối trực tiếp của một đỉnh
- **Công thức:** $C_D(v) = \frac{\deg(v)}{|V|-1}$
- Với đồ thị có hướng:
 - In-degree: $C_D^{in}(v) = \frac{\deg_{in}(v)}{|V|-1}$
 - Out-degree: $C_D^{out}(v) = \frac{\deg_{out}(v)}{|V|-1}$
- **Phạm vi:** $[0, 1]$
- **Đánh giá:** Giá trị cao thể hiện đỉnh có nhiều kết nối trực tiếp, quan trọng về mặt kết nối cục bộ.

2.2. Betweenness Centrality

- **Định nghĩa:** Mức độ một đỉnh nằm trên các đường đi ngắn nhất giữa các cặp đỉnh khác
- **Công thức:** $C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$
 - σ_{st} : số đường đi ngắn nhất từ s đến t
 - $\sigma_{st}(v)$: số đường đi ngắn nhất từ s đến t đi qua v
- **Phạm vi:** $[0, 1]$ (sau chuẩn hóa)
- **Đánh giá:** Giá trị cao thể hiện đỉnh quan trọng trong việc kết nối các phần khác nhau của mạng.

2.3. Closeness Centrality

- **Định nghĩa:** Mức độ gần gũi của một đỉnh với tất cả các đỉnh khác trong mạng
- **Công thức:** $C_C(v) = \frac{|V|-1}{\sum_{u \neq v} d(v,u)}$
- **Phạm vi:** $[0, 1]$
- **Đánh giá:** Giá trị cao thể hiện đỉnh có thể tiếp cận nhanh đến các đỉnh khác.

2.4. Eigenvector Centrality

- **Định nghĩa:** Đo lường ảnh hưởng của một đỉnh dựa trên tầm quan trọng của các đỉnh kết nối với nó
- **Công thức:** $x_v = \frac{1}{\lambda} \sum_{t \in N(v)} x_t$
 - λ là giá trị riêng lớn nhất của ma trận kề
 - $N(v)$ là tập các đỉnh kề với v
- **Phạm vi:** $[0, 1]$ (sau chuẩn hóa)
- **Đánh giá:** Giá trị cao thể hiện đỉnh kết nối với nhiều đỉnh quan trọng khác.

2.5. PageRank

- **Định nghĩa:** Biến thể của Eigenvector Centrality, tính đến cả hướng và trọng số của cạnh
- **Công thức:** $PR(v) = \frac{1-d}{N} + d \sum_{u \in M(v)} \frac{PR(u)}{L(u)}$
 - d : hệ số giảm (damping factor), thường = 0.85
 - $M(v)$: tập các đỉnh trỏ đến v
 - $L(u)$: số cạnh ra của đỉnh u
- **Phạm vi:** $[0, 1]$
- **Đánh giá:** Giá trị cao thể hiện đỉnh được nhiều đỉnh quan trọng trỏ đến.

2.6. HITS (Hub and Authority)

- **Định nghĩa:** Phân biệt hai loại đỉnh quan trọng:
 - Hub: đỉnh trở đến nhiều authority tốt
 - Authority: đỉnh được nhiều hub tốt trở đến
- **Công thức:**
 - Hub: $h_i = \sum_{j \in O(i)} a_j$
 - Authority: $a_i = \sum_{j \in I(i)} h_j$
 - $O(i)$: tập các đỉnh mà i trở đến
 - $I(i)$: tập các đỉnh trở đến i
- **Phạm vi:** $[0, 1]$ (sau chuẩn hóa)
- **Đánh giá:** Giá trị hub/authority cao thể hiện vai trò quan trọng trong cấu trúc liên kết.

2.7. Eccentricity

- **Định nghĩa:** Khoảng cách lớn nhất từ một đỉnh đến bất kỳ đỉnh nào khác trong mạng
- **Công thức:** $e(v) = \max_{u \in V} d(v, u)$
- **Phạm vi:** $[1, D]$ với D là đường kính mạng
- **Đánh giá:** Giá trị nhỏ thể hiện đỉnh nằm ở vị trí "trung tâm" của mạng.