Thực hành Lab05

Bài thực hành số 1 : Tìm hiểu các chỉ số thống kê cơ bản sử dụng Gephi

1.1 Độ đo cơ bản của mạng (Basic Network Metrics)

- **Average Degree**:
  - **Ý nghĩa**: Trung bình số lượng liên kết (degree) của tất cả các nút trong mạng.
  - **Công thức**:
$$\text{Average Degree} = \frac{\sum_{i=1}^{N} d_i}{N}$$
    với d_i là số liên kết của nút i, N là số lượng nút.
  - **Phạm vi**: $[0, N-1]$ đối với đồ thị không hướng, $[0, 2N-1]$ đối với đồ thị có hướng.
  - **Chỉ số tốt**: Giá trị cao cho thấy mạng có mật độ liên kết cao.
  - **Sử dụng**: Để đo lường mức độ kết nối tổng thể của mạng.
- **Network Diameter**:
  - **Ý nghĩa**: Khoảng cách lớn nhất giữa hai nút bất kỳ trong mạng.
  - **Công thức**:
$$\text{Diameter} = \max_{i,j \in V} \text{dist}(i, j)$$
  - **Phạm vi**: $[1, \infty)$.
  - **Chỉ số tốt**: Giá trị nhỏ cho thấy mạng có tính kết nối tốt.
  - **Sử dụng**: Để đo độ rộng của mạng.
- **Graph Density**:
  - **Ý nghĩa**: Đo lường mức độ kết nối giữa các nút trong mạng.
  - **Công thức**:
$$\text{Density} = \frac{2E}{N(N-1)}$$
    với E là số cạnh, N là số nút.
  - **Phạm vi**: $[0, 1]$.

- **Chỉ số tốt**: Giá trị cao thể hiện mạng có nhiều liên kết.
- **Sử dụng**: Để xác định mức độ kết nối tổng quát của mạng.

- **Connected Components**:
  - **Ý nghĩa**: Số lượng các thành phần liên thông trong mạng.
  - **Phạm vi**: $[1, N]$, với N là số lượng nút trong mạng.
  - **Chỉ số tốt**: Số lượng thành phần ít cho thấy mạng kết nối tốt.
  - **Sử dụng**: Để tìm các nhóm nút liên thông.

- **Average Path Length**:
  - **Ý nghĩa**: Độ dài trung bình của các đường đi ngắn nhất giữa mọi cặp nút.
  - **Công thức**:
$$\text{APL} = \frac{\sum_{i \neq j} \text{dist}(i, j)}{N(N-1)}$$
với $\text{dist}(i, j)$ là độ dài đường đi ngắn nhất giữa nút i và nút j.
  - **Phạm vi**: $[1, \infty)$.
  - **Chỉ số tốt**: Giá trị thấp thể hiện tính kết nối hiệu quả.
  - **Sử dụng**: Để đo độ hiệu quả của mạng.

- **Average Clustering Coefficient**:
  - **Ý nghĩa**: Đo mức độ mà các nút trong mạng có xu hướng tạo thành cụm.
  - **Công thức**:
$$C = \frac{1}{N} \sum_{i=1}^{N} \frac{2E_i}{k_i(k_i - 1)}$$
với E_i là số cạnh giữa các nút lân cận của nút i, k_i là bậc (degree) của nút i.
  - **Phạm vi**: $[0, 1]$.
  - **Chỉ số tốt**: Giá trị cao thể hiện tính cụm mạnh.
  - **Sử dụng**: Để đo lường khả năng hình thành cụm của mạng.

1.2 Độ đo tính trung tâm (Centrality Metrics)

- **Degree Centrality**:
  - **Ý nghĩa**: Đo mức độ kết nối của một nút.
  - **Công thức** (đồ thị không hướng):
$$C_D(v) = \frac{\deg(v)}{N-1}$$
  - **Phạm vi**: $[0, 1]$.
  - **Sử dụng**: Để đánh giá mức độ quan trọng của nút trong mạng.

- **Betweenness Centrality**:
  - **Ý nghĩa**: Đo mức độ mà một nút nằm trên các đường đi ngắn nhất giữa các cặp nút khác.
  - **Công thức**:
    $$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$
    với σ_{st} là số đường đi ngắn nhất giữa s và t, $\sigma_{st}(v)$ là số đường đi qua v.
  - **Phạm vi**: $[0, 1]$.
  - **Sử dụng**: Để đo vai trò trung gian của nút trong mạng.
- **Closeness Centrality**:
  - **Ý nghĩa**: Đo mức độ gần gũi của một nút tới các nút khác.
  - **Công thức**:
    $$C_C(v) = \frac{N-1}{\sum_{i \neq v} \text{dist}(v, i)}$$
    với $\text{dist}(v, i)$ là khoảng cách ngắn nhất giữa v và i.
  - **Phạm vi**: $[0, 1]$.
  - **Sử dụng**: Để xác định nút nào gần trung tâm của mạng nhất.
- **Eigenvector Centrality**:
  - **Ý nghĩa**: Đo mức độ quan trọng của một nút dựa trên mức độ kết nối của các nút liên kết với nó.
  - **Phạm vi**: $[0, 1]$.
  - **Sử dụng**: Để xác định các nút có ảnh hưởng lớn trong mạng.
- **PageRank**:
  - **Ý nghĩa**: Đo lường mức độ ảnh hưởng của một nút trong mạng.
  - **Sử dụng**: Để xếp hạng các nút quan trọng, đặc biệt trong công cụ tìm kiếm.
- **HITS (Hub and Authority)**:
  - **Ý nghĩa**:
    - **Hub Score**: Đánh giá mức độ một nút trỏ đến nhiều nút quan trọng.
    - **Authority Score**: Đánh giá mức độ một nút được nhiều nút trỏ tới.
  - **Sử dụng**: Để phân tích các nút trung tâm và nút đáng tin cậy trong mạng.
- **Eccentricity**:
  - **Ý nghĩa**: Khoảng cách lớn nhất từ một nút tới bất kỳ nút nào khác trong mạng.
  - **Sử dụng**: Để đo khoảng cách xa nhất mà nút có thể đạt được trong mạng.

Bài thực hành số 2 : Phân tích cộng đồng trong mạng xã hội sử dụng Gephi

2.1 Phân tích cấu trúc mạng Tính toán và so sánh các độ đo tính trung tâm (Centrality Measures)

- **Degree Centrality**:

Id	Label	Interval	Degree	Closeness Centrality	Betweenness Centrality
55	Kingpin		9	0.23913	0.105542
60	The Punisher		9	0.255814	0.076876
4	Hulk		9	0.179907	0.067637
38	Korg		8	0.195929	0.089776
91	Captain Marvel		7	0.182898	0.069187
14	Falcon		7	0.142857	0.052041
71	Jean Grey		7	0.194937	0.040016
3	Thor		6	0.225146	0.106776
51	Green Goblin		6	0.207547	0.075064
65	Domino		6	0.197436	0.074208
99			6	0.163482	0.069577
84	Silver Surfer		6	0.195431	0.050565
49	Mysterio		6	0.201044	0.045857
63	Deadpool		6	0.206434	0.042953
56	Jessica Jones		6	0.198966	0.040014
30	Peggy Carter		6	0.199482	0.038824
92	Monica Rambeau		6	0.246795	0.032069
61	Ghost Rider		6	0.175399	0.025134
36	Killmonger		6	0.128978	0.024084
31	Red Skull		6	0.240122	0.00615
41	Agatha Harkness		6	1.0	0.004604
73	Beast		5	0.211538	0.090248
5	Black Widow		5	0.191542	0.083645
59	Iron Fist		5	0.195431	0.081386
40	Jane Foster		5	0.222543	0.073555
62	Blade		5	0.228487	0.072148
2	Captain America		5	0.208108	0.054476
83	Pyro		5	0.155556	0.0534

Hình 1: Các nút có Degree Centrality cao nhất.

Phân tích:

- Các nút có Degree Centrality cao nhất là:
  - **Id 4 (Hulk)**: Degree = 9
  - **Id 55 (Kingpin)**: Degree = 9
  - **Id 60 (The Punisher)**: Degree = 9
- Ý nghĩa: Những nút này có số lượng kết nối trực tiếp cao nhất, đóng vai trò như "hub" trong mạng, có khả năng lan truyền thông tin nhanh và ảnh hưởng mạnh đến các nút khác.

- **Betweenness Centrality**:

Id	Label	Interval	Degree	Closeness Centrality	Betweenness Centrality
3	Thor		6	0.225146	0.106776
55	Kingpin		9	0.23913	0.105542
73	Beast		5	0.211538	0.090248
38	Korg		8	0.195929	0.089776
5	Black Widow		5	0.191542	0.083645
59	Iron Fist		5	0.195431	0.081386
86	The Watcher		4	0.194937	0.078833
60	The Punisher		9	0.255814	0.076876
50	Electro		4	0.213889	0.07585
26	Quicksilver		4	0.222543	0.07527
51	Green Goblin		6	0.207547	0.075064
65	Domino		6	0.197436	0.074208
40	Jane Foster		5	0.222543	0.073555
62	Blade		5	0.228487	0.072148
99			6	0.163482	0.069577
91	Captain Marvel		7	0.182898	0.069187
4	Hulk		9	0.179907	0.067637
42	Wong		4	0.177419	0.067232
46	Ego the Living Planet		4	0.190123	0.061949
2	Captain America		5	0.208108	0.054476
83	Pyro		5	0.155556	0.0534
14	Falcon		7	0.142857	0.052041
84	Silver Surfer		6	0.195431	0.050565
97			5	0.25	0.050365
33	Yondu		5	0.239875	0.047776
49	Mysterio		6	0.201044	0.045857
63	Deadpool		6	0.206434	0.042953
90	Moondragon		5	0.179907	0.040808
71	Jean Grey		7	0.194937	0.040016

Hình 2: Các nút có Betweenness Centrality cao nhất.

Phân tích:

- Các nút có Betweenness Centrality cao nhất là:
  - **Id 3 (Thor)**: Betweenness = 0.106776
  - **Id 55 (Kingpin)**: Betweenness = 0.105542
  - **Id 73 (Beast)**: Betweenness = 0.090248
- Ý nghĩa: Những nút này đóng vai trò như cầu nối giữa các cụm trong mạng. Nếu loại bỏ chúng, mạng lưới có thể bị phân mảnh hoặc mất kết nối.

- **Closeness Centrality**:

Id	Label	Interval	Degree	Closeness Centrality	Betweenness Centrality
41	Agatha Harkness		6	1.0	0.004604
48	Vulture		4	1.0	0.002345
37	Hela		2	1.0	0.007833
82	Sabretooth		2	1.0	0.007833
24	Mantis		2	1.0	0.000438
32	Ultron		2	0.75	0.0
9	Black Panther		3	0.666667	0.000593
15	Winter Soldier		2	0.666667	0.000412
12	Wasp		2	0.5	0.000309
20	Rocket Raccoon		2	0.454545	0.0
29	Phil Coulson		4	0.260317	0.007524
60	The Punisher		9	0.255814	0.076876
10	Scarlet Witch		5	0.251634	0.03525
97			5	0.25	0.050365
80	Kitty Pryde		4	0.247588	0.004999
92	Monica Rambeau		6	0.246795	0.032069
21	Groot		4	0.245283	0.0
31	Red Skull		6	0.240122	0.00615
33	Yondu		5	0.239875	0.047776
55	Kingpin		9	0.23913	0.105542
69	Professor X		2	0.237082	0.0
57	Daredevil		4	0.228986	0.00408
62	Blade		5	0.228487	0.072148
3	Thor		6	0.225146	0.106776
47	Quentin Beck		2	0.224784	0.0
40	Jane Foster		5	0.222543	0.073555
26	Quicksilver		4	0.222543	0.07527
44	The Ancient One		3	0.218421	0.0
74	Wolverine		5	0.215686	0.037481

Hình 3: Các nút có Closeness Centrality cao nhất.

Phân tích:

- Các nút có Closeness Centrality cao nhất là:
  - **Id 41 (Agatha Harkness)**: Closeness = 1.0
  - **Id 48 (Vulture)**: Closeness = 1.0
  - **Id 24 (Mantis)**: Closeness = 1.0
- Ý nghĩa: Những nút này nằm ở vị trí trung tâm, có khả năng tiếp cận nhanh nhất đến toàn bộ các nút khác trong mạng lưới.

2.2 Phát hiện cộng đồng Thực hiện phân cụm mạng lưới

2.2.1 Ưu và nhược điểm của mỗi phương pháp

Thuật toán	Ưu điểm	Nhược điểm
Louvain	• Chạy nhanh và hiệu quả với đồ thị lớn. • Modularity cao (0.518) cho thấy cộng đồng được phân cụm rõ ràng và có cấu trúc tốt. • Kết quả ổn định và đáng tin cậy.	• Phụ thuộc vào tham số độ phân giải (resolution). • Có thể bỏ sót các cộng đồng nhỏ.
Girvan-Newman	• Phù hợp với đồ thị nhỏ, cộng đồng được phân tách rõ ràng. • Hữu ích trong việc phân tích chi tiết từng cộng đồng.	• Chạy chậm trên đồ thị lớn. • Modularity thấp (0.453), các cộng đồng kém kết nối chặt chẽ. • Không tối ưu cho mạng lưới dày đặc.

Bảng 1: So sánh ưu và nhược điểm của các thuật toán phân cụm.

2.2.2 Ý nghĩa của các cộng đồng được phát hiện

Các cộng đồng phát hiện từ mạng xã hội nhân vật Marvel mang ý nghĩa quan trọng:

- **Cộng đồng trong Louvain:**
  - Phát hiện 11 cộng đồng với Modularity cao (0.518), cho thấy các cộng đồng được phân chia rõ ràng và có sự kết nối chặt chẽ.
  - Các cộng đồng lớn có thể đại diện cho các nhóm nhân vật chính như Avengers, X-Men hoặc Guardians of the Galaxy.
  - Cộng đồng nhỏ hơn phản ánh các nhân vật có mối quan hệ độc lập hoặc ít tương tác.
- **Cộng đồng trong Girvan-Newman:**
  - Phát hiện 18 cộng đồng với Modularity thấp hơn (0.453), phù hợp để tìm các cộng đồng nhỏ hơn và chi tiết hơn.
  - Kết quả phản ánh những kết nối phức tạp hoặc mối quan hệ phụ giữa các nhân vật, ví dụ như nhóm phụ trong cùng một cộng đồng lớn.

2.2.3 Đề xuất phương pháp phân cụm phù hợp

Đề xuất phương pháp: Thuật toán Louvain.
Lý do:

- Modularity cao (0.518) chứng tỏ cộng đồng được phân chia rõ ràng và có cấu trúc tốt.
- Hiệu quả và ổn định, phù hợp cho mạng lưới lớn như mạng xã hội nhân vật Marvel.
- Thích hợp cho phân tích tổng quan, giúp phát hiện các nhóm lớn và ảnh hưởng chính trong mạng.

Tuy nhiên:

- Nếu mục tiêu là khám phá chi tiết mối quan hệ giữa các nhân vật nhỏ hơn, thuật toán Girvan-Newman có thể hữu ích như một công cụ bổ sung, dù Modularity thấp hơn.