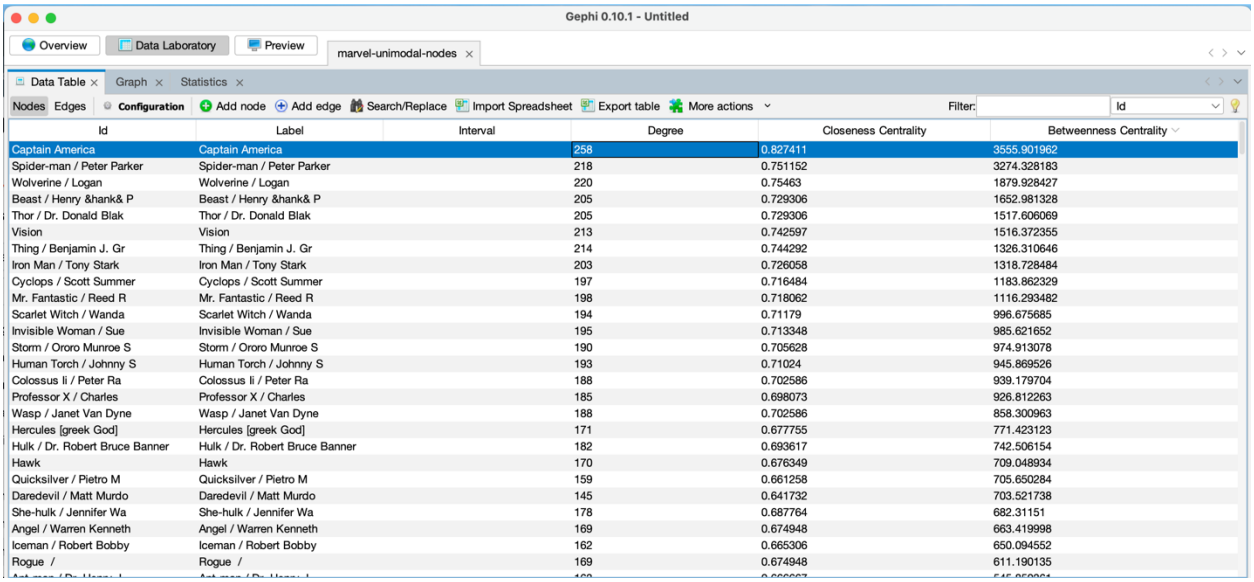
Phần 1: Phân tích cấu trúc mạng Tính toán và so sánh các độ đo tính trung tâm (Centrality Measures) sau:

• Degree Centrality
• Betweenness Centrality
• Closeness Centrality Hãy xác định 3 nút có độ trung tâm cao nhất theo mỗi độ đo và giải thích ý nghĩa của chúng trong mạng lưới.

Bài làm



Id	Label	Interval	Degree	Closeness Centrality	Betweenness Centrality
Captain America	Captain America		258	0.827411	3555.901962
Spider-man / Peter Parker	Spider-man / Peter Parker		218	0.751152	3274.328183
Wolverine / Logan	Wolverine / Logan		220	0.75463	1879.928427
Beast / Henry &hank& P	Beast / Henry &hank& P		205	0.729306	1652.981328
Thor / Dr. Donald Blak	Thor / Dr. Donald Blak		205	0.729306	1517.606069
Vision	Vision		213	0.742597	1516.372355
Thing / Benjamin J. Gr	Thing / Benjamin J. Gr		214	0.744292	1326.310646
Iron Man / Tony Stark	Iron Man / Tony Stark		203	0.726058	1318.728484
Cyclops / Scott Summer	Cyclops / Scott Summer		197	0.716484	1183.862329
Mr. Fantastic / Reed R	Mr. Fantastic / Reed R		198	0.718062	1116.293482
Scarlet Witch / Wanda	Scarlet Witch / Wanda		194	0.71179	996.675685
Invisible Woman / Sue	Invisible Woman / Sue		195	0.713348	985.621652
Storm / Ororo Munroe S	Storm / Ororo Munroe S		190	0.705628	974.913078
Human Torch / Johnny S	Human Torch / Johnny S		193	0.71024	945.869526
Colossus Ii / Peter Ra	Colossus Ii / Peter Ra		188	0.702586	939.179704
Professor X / Charles	Professor X / Charles		185	0.698073	926.812263
Wasp / Janet Van Dyne	Wasp / Janet Van Dyne		188	0.702586	858.300963
Hercules [greek God]	Hercules [greek God]		171	0.677755	771.423123
Hulk / Dr. Robert Bruce Banner	Hulk / Dr. Robert Bruce Banner		182	0.693617	742.506154
Hawk	Hawk		170	0.676349	709.048934
Quicksilver / Pietro M	Quicksilver / Pietro M		159	0.661258	705.650284
Daredevil / Matt Murdo	Daredevil / Matt Murdo		145	0.641732	703.521738
She-hulk / Jennifer Wa	She-hulk / Jennifer Wa		178	0.687764	682.31151
Angel / Warren Kenneth	Angel / Warren Kenneth		169	0.674948	663.419998
Iceman / Robert Bobby	Iceman / Robert Bobby		162	0.665306	650.094552
Rogue /	Rogue /		169	0.674948	611.190135

Degree: 258 / 327(tổng số node) = 0,79 ; 218/327 =0,667 ; 220/327 = 0,672
Betweenness Centrality: 3555,9 ; 3274 ; 1879,9
Closeness Centrality: 0,827; 0,75 ; 0,754

Ý nghĩa các chỉ số:

1. Degree Centrality

- **0,79**: Nút này có số lượng kết nối trực tiếp rất lớn, đóng vai trò trung tâm trong việc liên kết các nút khác.
- **0,667** và **0,672**: Hai nút này cũng có nhiều kết nối, nhưng tầm quan trọng thấp hơn nút đầu.

2. Betweenness Centrality

- **3555,9**: Nút này đóng vai trò là cầu nối quan trọng nhất, điều phối luồng thông tin giữa các phần của mạng.
- **3274**: Cũng là một nút quan trọng trong việc kết nối, nhưng kém hơn một chút so với nút đầu.
- **1879,9**: Vai trò trung gian ít quan trọng hơn hai nút trên.

3. Closeness Centrality

- **0,827**: Nút này có vị trí chiến lược, gần với tất cả các nút khác trong mạng, giúp truy cập thông tin nhanh chóng.
- **0,75** và **0,754**: Hai nút này cũng nằm ở vị trí thuận lợi trong mạng, nhưng không bằng nút đầu.

Tổng kết:

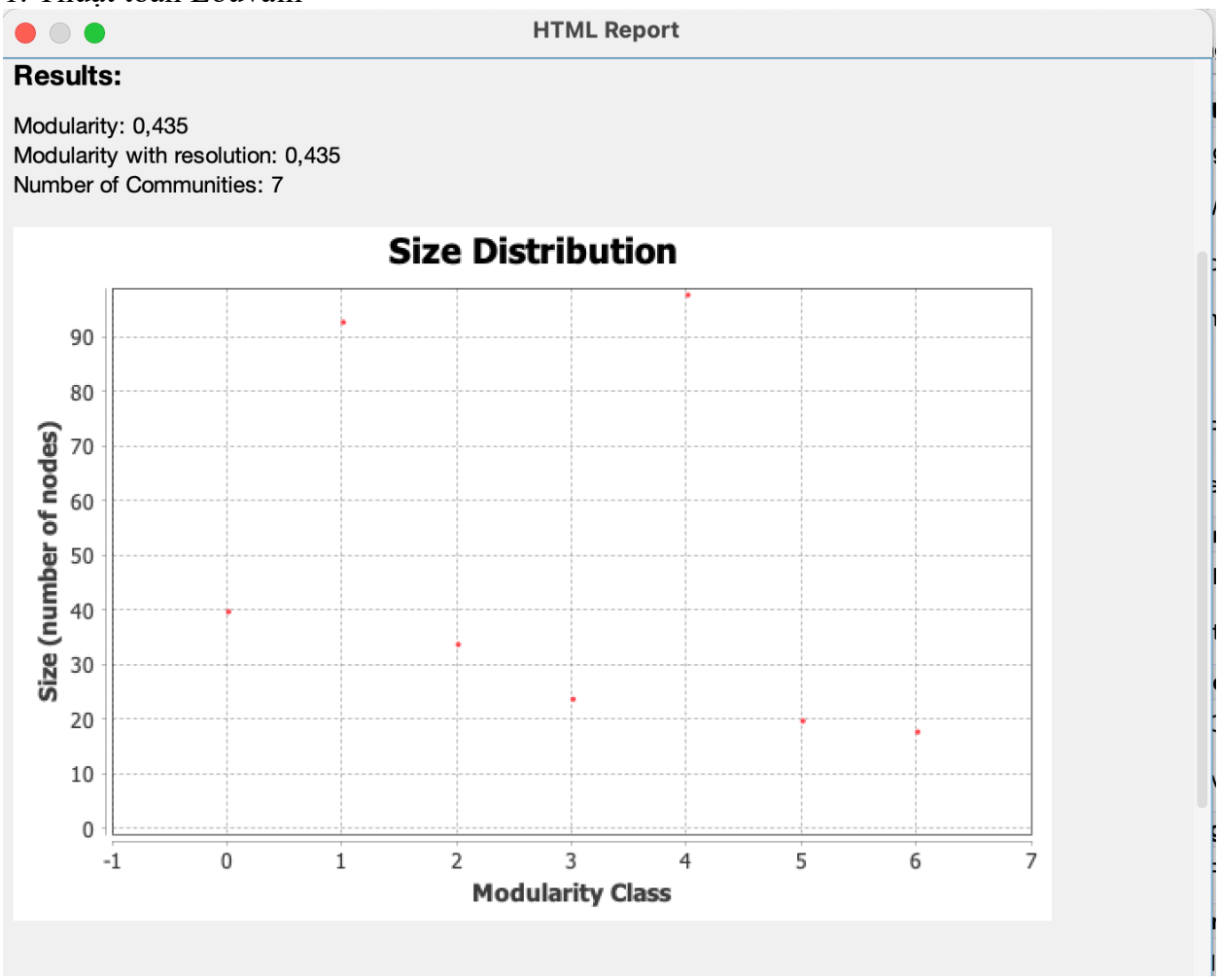
Nút có **Degree Centrality** 0,79, **Betweenness Centrality** 3555,9, và **Closeness Centrality** 0,827 là nút quan trọng nhất trong mạng, vừa có nhiều kết nối trực tiếp, vừa là cầu nối quan trọng, vừa dễ dàng tiếp cận các nút khác.

Phần 2: Phát hiện cộng đồng Thực hiện phân cụm mạng lưới sử dụng 3 thuật toán sau:
1. Thuật toán Louvain
2. Thuật toán Girvan-newman (gephi.org/plugins/#/plugin/girvan-newman-clustering hoặc gephi.org/plugins/#/plugin/newman-girvan-plugin)
3. Thuật toán LPA (gephi.org/plugins/#/plugin/label-propagation-clustering)
Với mỗi thuật toán, hãy:
• Ghi lại số lượng cộng đồng được phát hiện
• Tính toán độ đo Modularity của kết quả phân cụm

• Lưu ảnh kết quả phân cụm với các node được tô màu theo cộng đồng

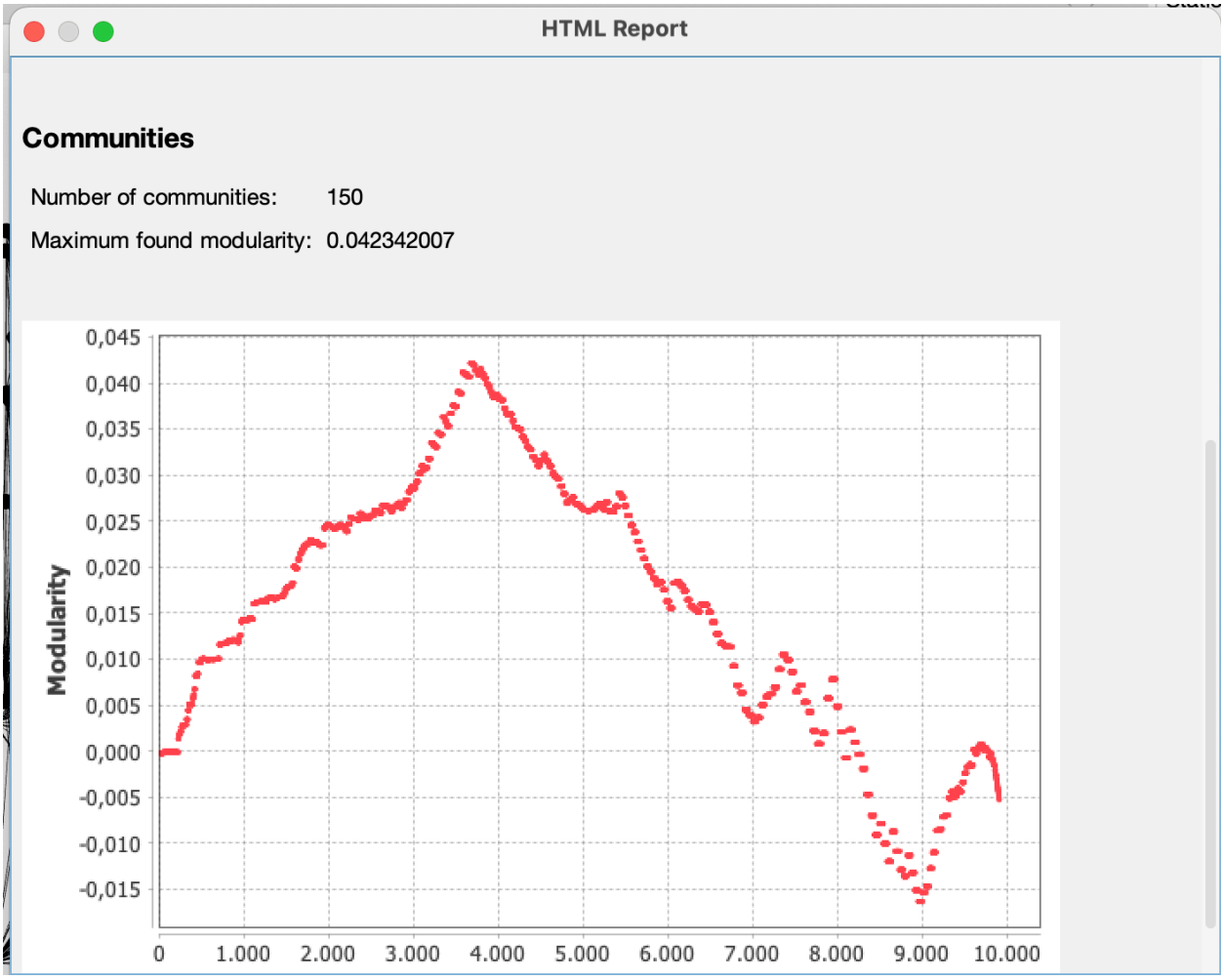
Bài làm

1. Thuật toán Louvain

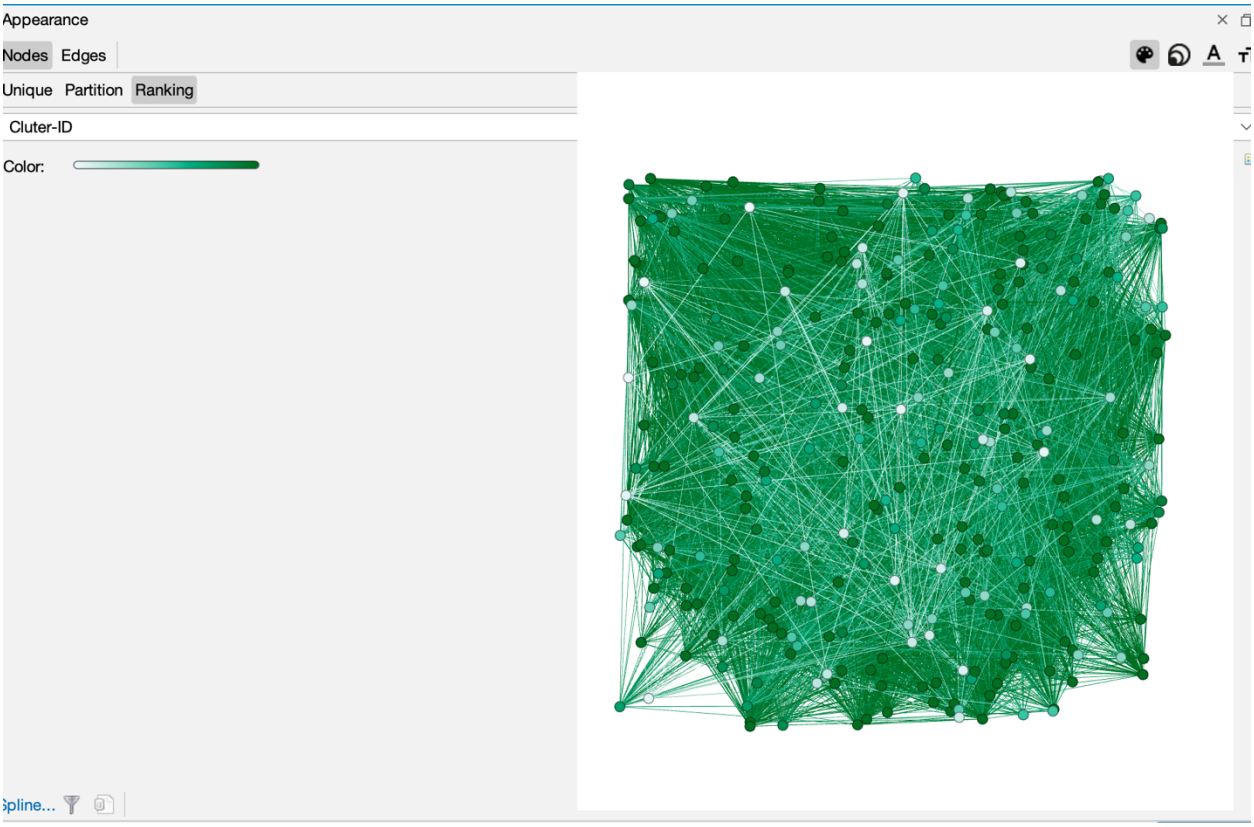


Tổng số cụm: 7 cụm
Độ đo: 0,435

2. Thuật toán Girvan-newman (gephi.org/plugins/#/plugin/girvan-newman-clustering hoặc gephi.org/plugins/#/plugin/newman-girvan-plugin)

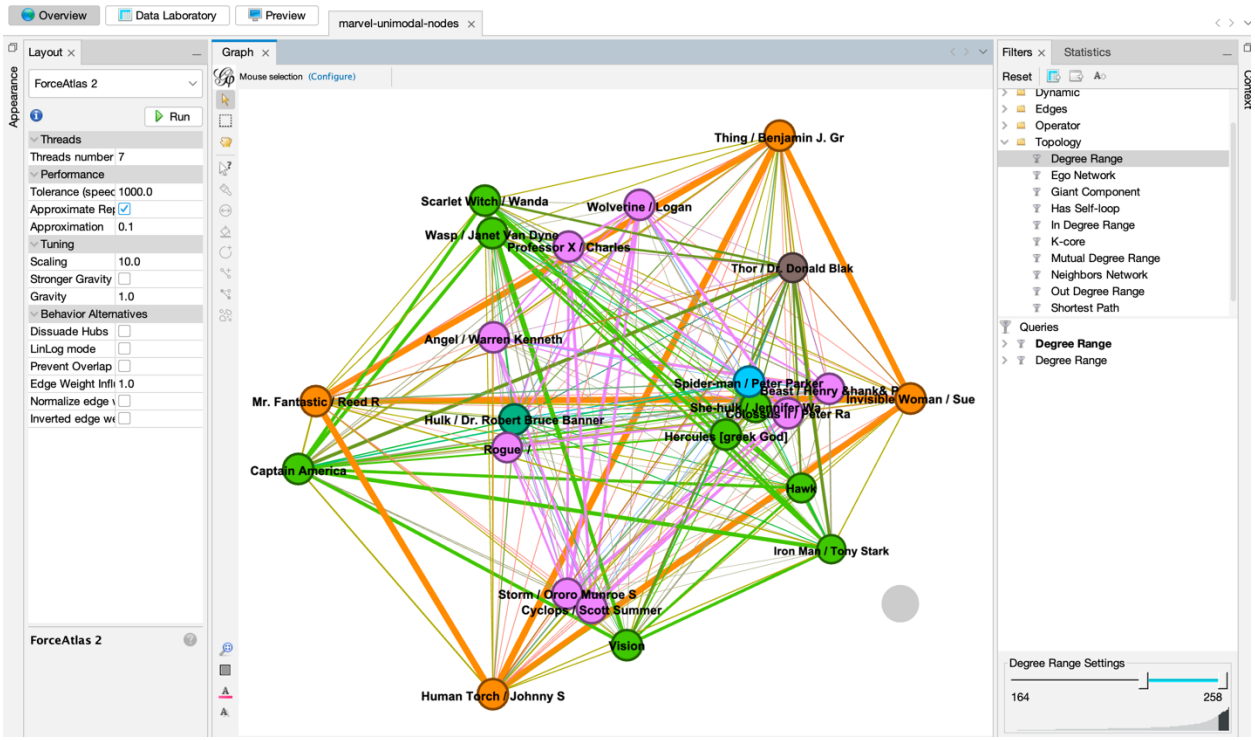


Tổng: 150 cộng đồng
Độ đo = 0,04



Phần 3: Trực quan hóa Tạo một bản trực quan hóa đẹp và có ý nghĩa cho mạng lưới bằng cách:
1. Sử dụng thuật toán layout ForceAtlas2 với các tham số phù hợp2. Điều chỉnh kích thước node theo độ đo trung tâm đã tính
3. Tô màu node theo kết quả phân cụm từ thuật toán cho kết quả tốt nhất
4. Thêm nhãn cho các node quan trọng (có độ trung tâm cao)

Bài làm



Phần 4: Báo cáo và đánh giá Viết báo cáo ngắn bao gồm:
1. So sánh kết quả của 3 thuật toán phân cụm, nêu ưu và nhược điểm của mỗi phương pháp
2. Giải thích ý nghĩa của các cộng đồng được phát hiện trong ngữ cảnh của mạng xã hội
3. Đề xuất phương pháp phân cụm phù hợp nhất cho loại dữ liệu này và lý do

Bài làm

1. So sánh kết quả của 3 thuật toán phân cụm, ưu và nhược điểm
a. K-Means
Kết quả: Chia các nút thành các nhóm dựa trên sự tương đồng về đặc điểm, thường phù hợp với dữ liệu số.
Ưu điểm:
Dễ triển khai, tính toán nhanh.
Hiệu quả với dữ liệu có dạng hình cầu.

Dễ hiểu và trực quan.
Nhược điểm:
Không tốt với dữ liệu phi tuyến hoặc không có hình dạng cụ thể.
Nhạy cảm với giá trị ngoại lai và cần biết trước số cụm kk.
b. Hierarchical Clustering (Phân cụm thứ bậc)
Kết quả: Tạo cấu trúc cây phân cấp (dendrogram) để thể hiện sự tương đồng giữa các nút.
Ưu điểm:
Không cần xác định trước số cụm.
Cung cấp thông tin trực quan về mối quan hệ giữa các nút.
Nhược điểm:
Tốn thời gian và tài nguyên khi dữ liệu lớn.
Kết quả phụ thuộc vào phương pháp liên kết (linkage).
c. Community Detection (Louvain hoặc Girvan-Newman)
Kết quả: Xác định các cộng đồng tự nhiên trong mạng xã hội dựa trên mối quan hệ giữa các nút.
Ưu điểm:
Tốt với dữ liệu mạng xã hội, tập trung vào mối quan hệ hơn là đặc điểm nút.
Tự động xác định số cộng đồng.
Nhược điểm:
Tốn tài nguyên nếu mạng lớn.
Không xử lý tốt dữ liệu ngoài mạng xã hội.

2. Giải thích ý nghĩa của các cộng đồng được phát hiện
Cộng đồng trong mạng xã hội:
Các cộng đồng thường đại diện cho các nhóm người có chung đặc điểm, mối quan tâm, hoặc tương tác mạnh với nhau.
Ví dụ:
Một cộng đồng gồm các thành viên trong một tổ chức, câu lạc bộ.
Cộng đồng người dùng cùng theo dõi một chủ đề hoặc hashtag.
Ý nghĩa trong mạng xã hội:
Hiểu hành vi, sở thích người dùng để cá nhân hóa nội dung.
Xác định các cá nhân có sức ảnh hưởng lớn trong mạng lưới.
Hỗ trợ chiến lược marketing hoặc các chiến dịch xã hội.

3. Đề xuất phương pháp phân cụm phù hợp nhất và lý do
Phương pháp đề xuất: **Community Detection (Louvain)**
Lý do:

- Dữ liệu mạng xã hội thường dựa trên quan hệ giữa các nút, không chỉ dựa vào đặc điểm của nút.

- Louvain tối ưu hóa độ đo modularity, giúp phát hiện các cộng đồng có ý nghĩa trong ngữ cảnh xã hội.
- Không cần biết trước số lượng cụm, phù hợp với tính phức tạp và đa dạng của mạng xã hội.