Bài thực hành số 2

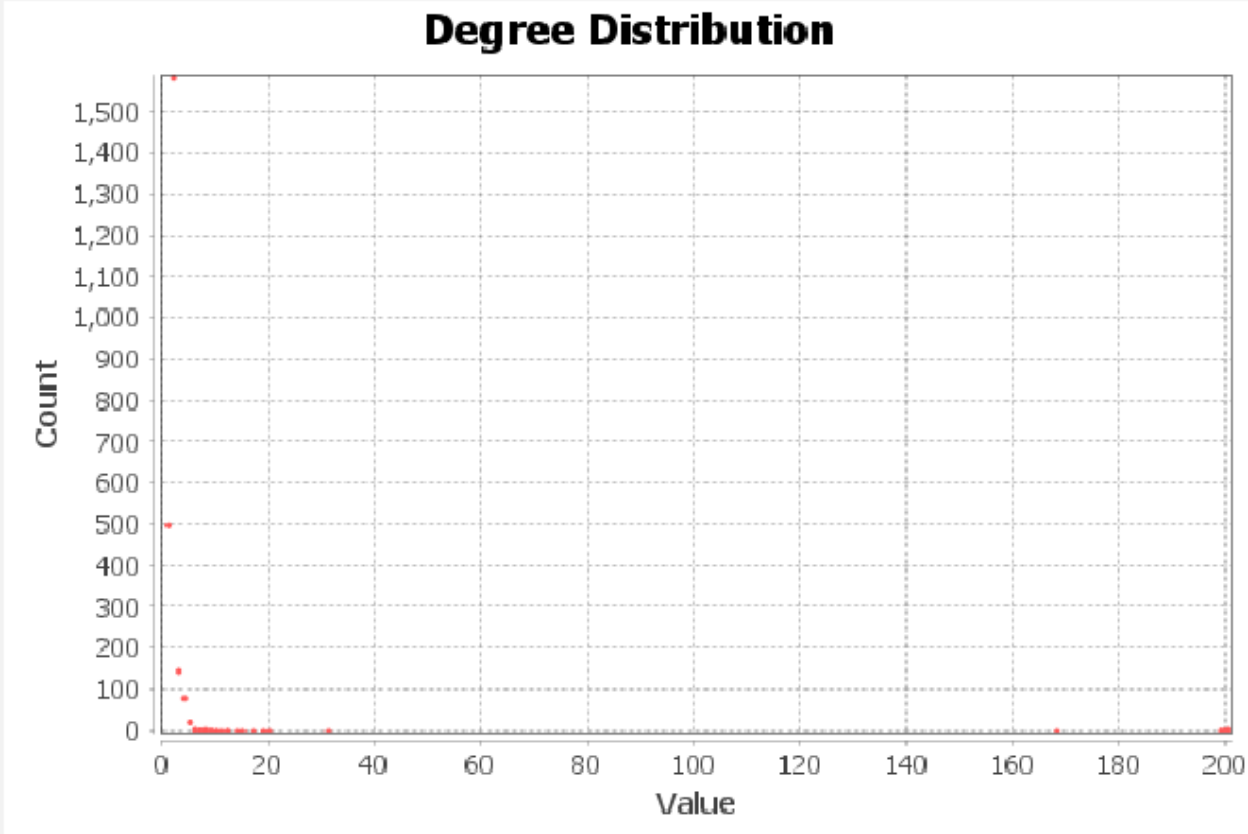
Phần 1: Bộ dữ liệu Spotify

Degree

Id	Label	Interval	type	In-Degree	Out-Degree	Degree
Country	Country		Genre	200	0	200
Rock	Rock		Genre	200	0	200
Classical	Classical		Genre	200	0	200
Rap	Rap		Genre	200	0	200
Hip-Hop	Hip-Hop		Genre	200	0	200
R&B	R&B		Genre	200	0	200
Pop	Pop		Genre	199	0	199
Jazz	Jazz		Genre	199	0	199
Blues	Blues		Genre	199	0	199
Wolfgang Amad...	Wolfgang Amad...		Artist	0	168	168
Juice WRLD	Juice WRLD		Artist	0	31	31
The Black Keys	The Black Keys		Artist	0	20	20
XXXTENTACION	XXXTENTACION		Artist	0	19	19
Ariana Grande	Ariana Grande		Artist	0	17	17
2 Chainz	2 Chainz		Artist	0	15	15
Post Malone	Post Malone		Artist	0	14	14
Cage The Elepha...	Cage The Elepha...		Artist	0	12	12
Leon Bridges	Leon Bridges		Artist	0	12	12
T-Pain	T-Pain		Artist	0	11	11
Khalid	Khalid		Artist	0	10	10
Drake	Drake		Artist	0	10	10
Billie Eilish	Billie Eilish		Artist	0	9	9
Johann Sebastia...	Johann Sebastia...		Artist	0	9	9
Eminem	Eminem		Artist	0	9	9
Kelsea Ballerini	Kelsea Ballerini		Artist	0	8	8

- Các thể loại như **Country**, **Rock**, **Classical**, và **Rap** đều có giá trị "Degree" cao nhất (200), cho thấy đây là các thể loại được kết nối mạnh mẽ và phổ biến trong mạng lưới.
- Điều này có thể phản ánh rằng các thể loại này có sự tương tác rộng rãi với nhiều nghệ sĩ, album, hoặc bài hát khác nhau.

Bảng dữ liệu phản ánh rõ ràng sự phổ biến của các thể loại âm nhạc so với nghệ sĩ, đồng thời cho thấy mức độ liên kết và tương tác của từng nghệ sĩ trong mạng lưới. Những giá trị này có thể được sử dụng để xác định các nút trung tâm và phân tích sự lan tỏa trong mạng lưới âm nhạc



Betweenness Centrality

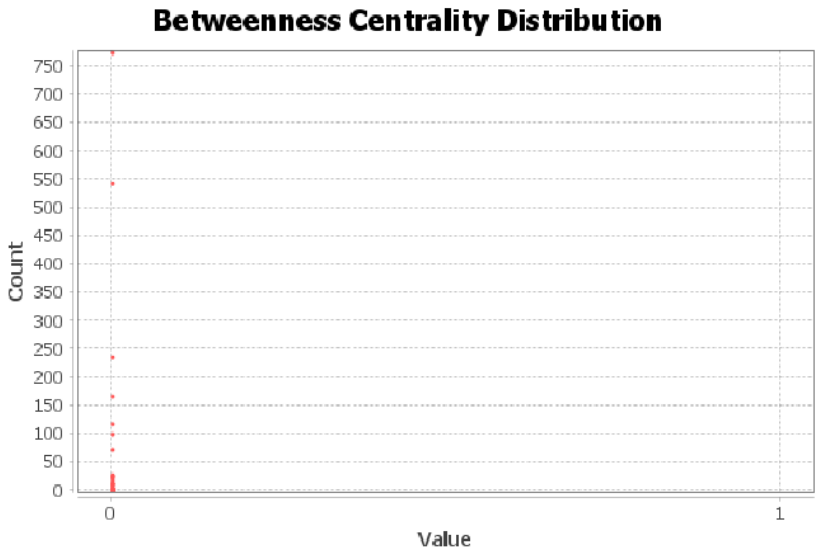
Results:

Diameter: 3

Radius: 0

Average Path length: 1.213529682466636

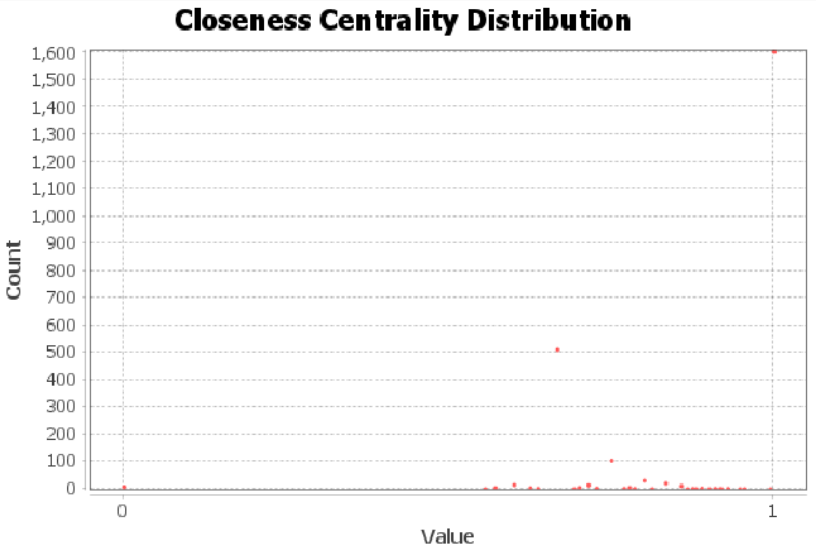
Betweenness Centrality
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.000001
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0
0.0



Các giá trị "Betweenness Centrality" trong bảng dữ liệu đều rất thấp, dao động từ **0.000001** đến **0.0**. Điều này cho thấy rằng các nút (thể loại hoặc nghệ sĩ) trong mạng lưới không đóng vai trò quan trọng trong việc kết nối giữa các cụm hoặc đường đi ngắn nhất.

3. Closeness Centrality

Closeness Centrality
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0
1.0

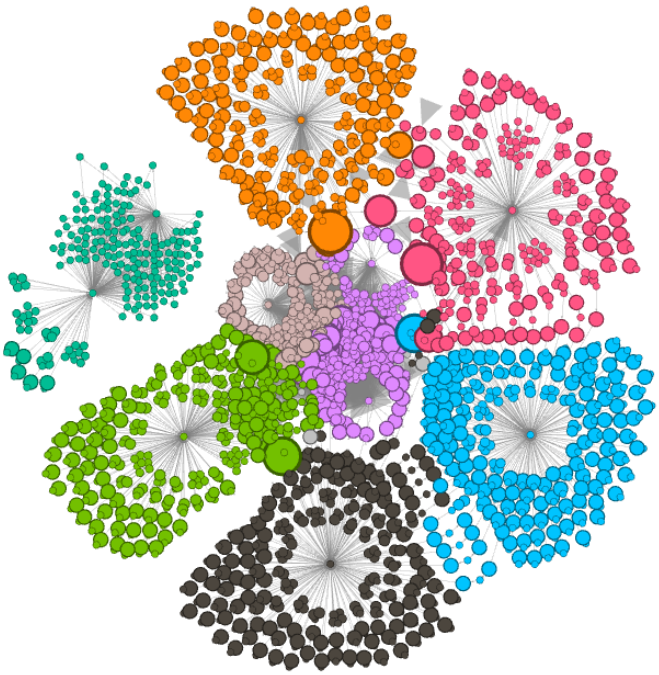


3 giá trị cao nhất trong cột đều bằng 1.0 chỉ ra rằng các nút trong mạng lưới có khoảng cách trực tiếp hoặc ngắn nhất với tất cả các nút khác. Điều này gợi ý rằng mạng lưới có cấu trúc rất tập trung, nơi mọi nút đều liên kết chặt chẽ và dễ dàng tiếp cận các nút khác.

Giá trị này cho thấy các nút nằm trong một mạng mà các đường đi ngắn nhất giữa chúng là tối ưu hoặc gần như không có nút nào ở vị trí bất lợi so với các nút khác.

Phần 2:

Louvain



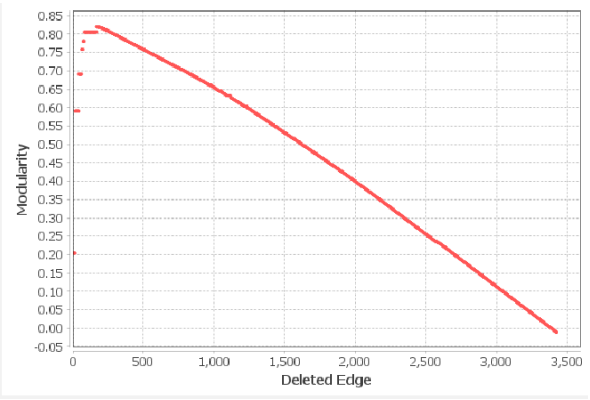
Hình ảnh được phân cụm bằng Louvain

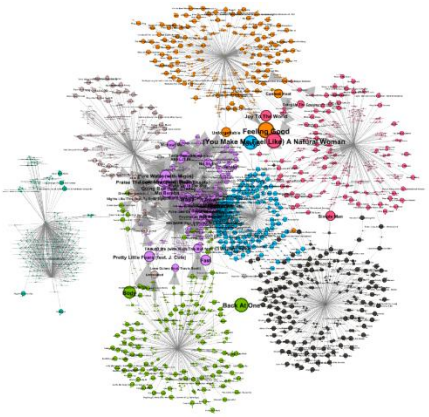
Modularity: 0.819
Modularity with resolution: 0.819
Number of Communities: 8

Số lượng cộng đồng 8

- Các cụm có kích thước khác nhau, cho thấy mức độ tập trung của các kết nối trong từng nhóm.
- Một số cụm lớn (ví dụ, cụm màu cam, xanh lam, và hồng) có sự tập trung cao, biểu thị rằng các nút trong các cụm này có liên kết mạnh mẽ.
- Các cụm nhỏ hơn (ví dụ, cụm xanh lục hoặc cụm nhỏ nằm rải rác) có thể đại diện cho các nút ít kết nối hoặc nhóm ngoại biên.

Givan-Newman



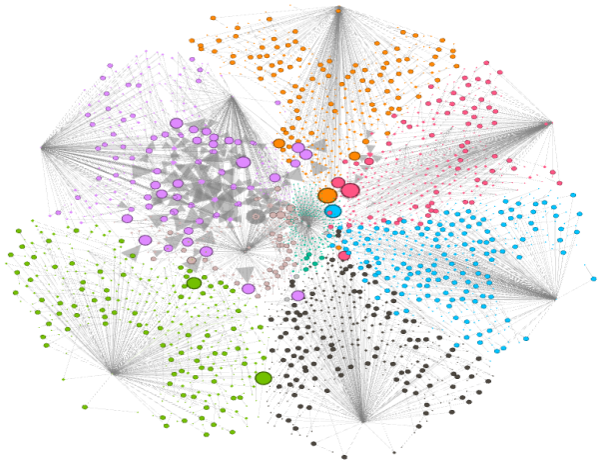


Số lượng cộng đồng được phát hiện là 8

Modularity khi phân cụm 0.819

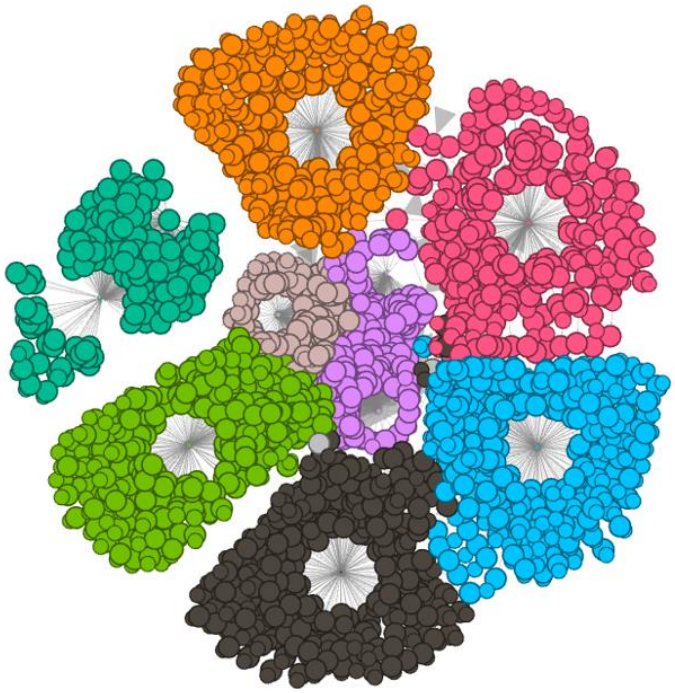
Phần 3:

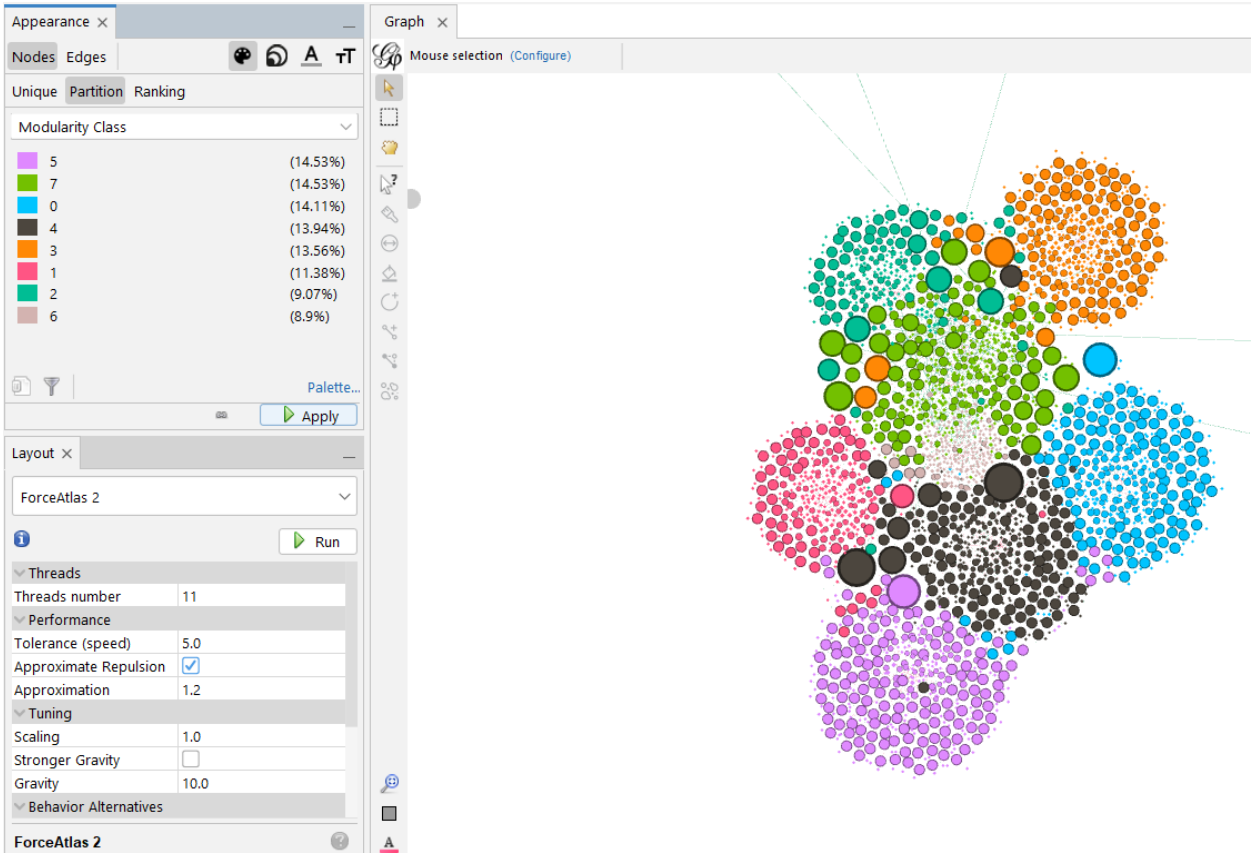
layout ForceAtlas2



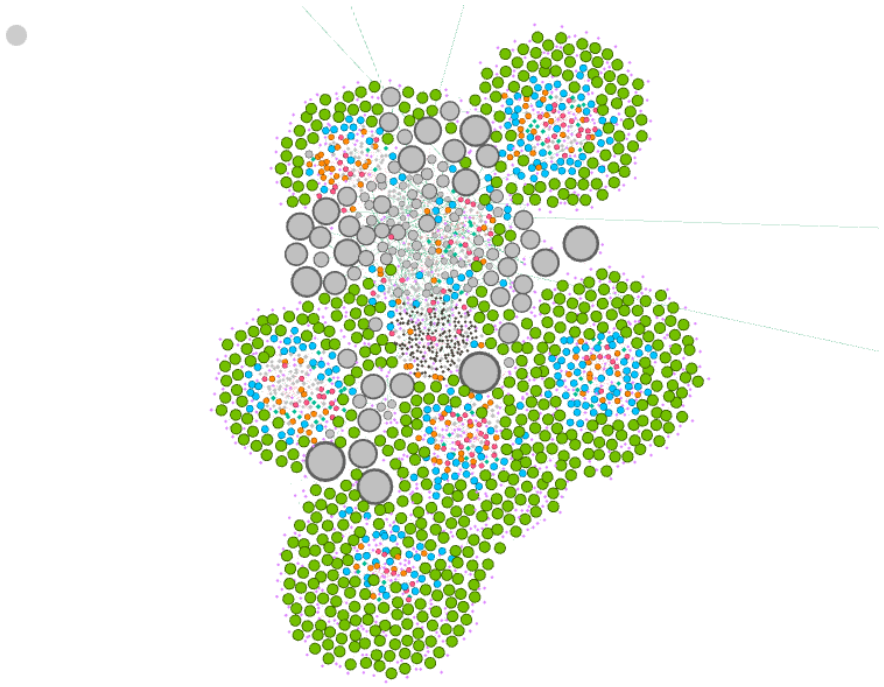
Threads	
Threads number	11
Performance	
Tolerance (speed)	1.0
Approximate Repulsion	☐
Approximation	1.2
Tuning	
Scaling	1.0
Stronger Gravity	☐
Gravity	5.0
Behavior Alternatives	
Dissuade Hubs	☐
LinLog mode	☐
Prevent Overlap	☐
Edge Weight Influence	1.0
Normalize edge weight	☑
Inverted edge weights	☐

Điều chỉnh kích thước cho mỗi cụm





Tô màu cho kết quả tốt nhất với Modularity class



Tô màu cho kết quả tốt nhất với Betweenness Centrality

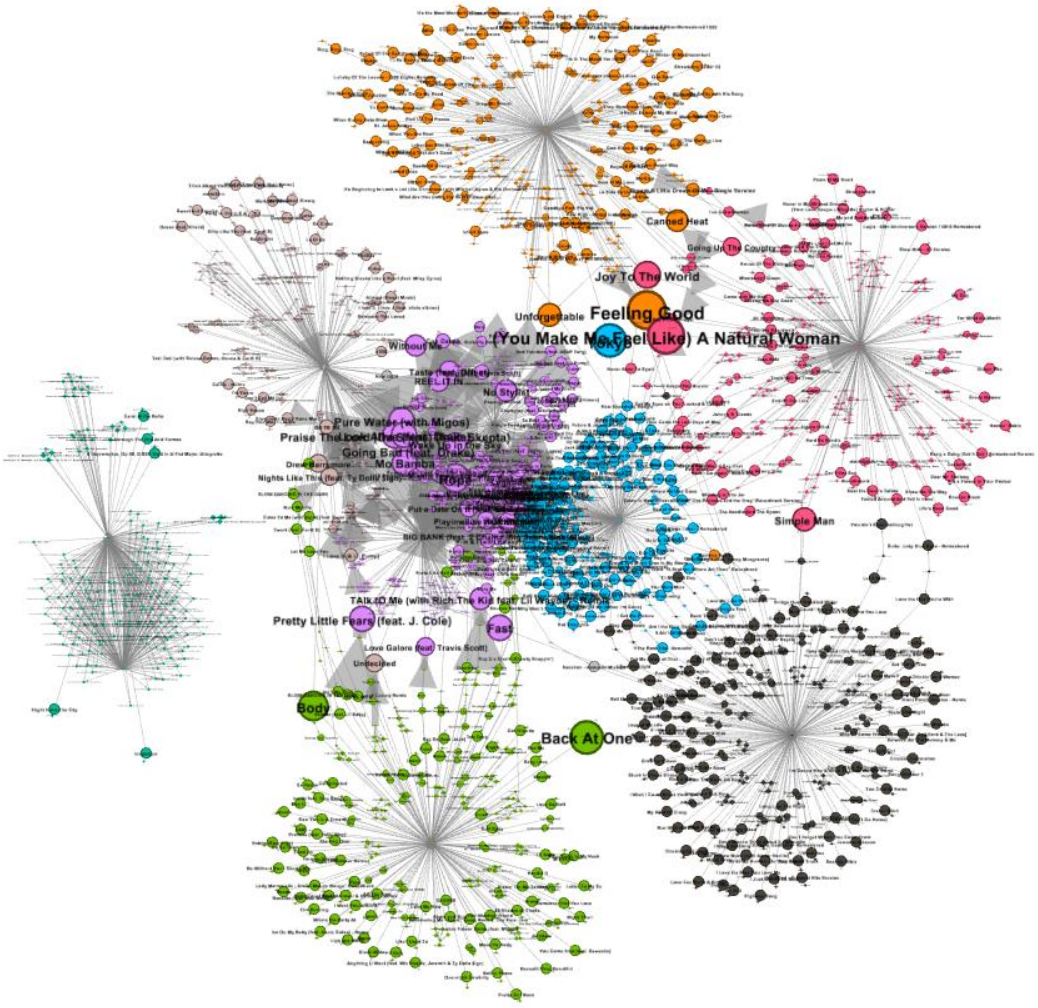
Cụm **màu xanh lá** là lớn nhất, chiếm phần lớn node trong mạng, cho thấy đây là cụm có sự kết nối mạnh mẽ hoặc chi phối mạng lưới.
Các cụm khác (xanh dương, cam, tím) nhỏ hơn, nằm rải rác xung quanh cụm xanh lá, cho thấy tính phân tầng hoặc sự đa dạng trong mạng.

Tổng thể: • **Điểm mạnh:**

- Đồ thị thể hiện tốt sự phân cụm và nổi bật các node quan trọng.
- Màu sắc và kích thước node được mã hóa trực quan, giúp dễ dàng xác định vai trò của các node trong mạng.

  - **Điểm cần cải thiện:**

- Trung tâm đồ thị bị chồng chéo nhiều, nên tăng khoảng cách giữa các node trung tâm để dễ đọc hơn.
- Một số cụm nhỏ hơn (màu tím, cam) chưa thể hiện rõ ràng vai trò của chúng trong mạng.



Thêm nhãn cho mỗi node

Đồ thị cho thấy một **mạng lưới phân cụm rõ ràng**, với các cụm màu sắc khác nhau biểu thị cho các cộng đồng trong mạng lưới.
Các cụm lớn (như cụm hồng, xanh lá, xanh lam) tập trung nhiều node và có mật độ kết nối cao, phản ánh các nhóm liên kết chặt chẽ trong mạng lưới.
Các cụm nhỏ hơn ở rìa (ví dụ: cụm xanh nhạt, nâu) có thể đại diện cho các nhóm độc lập hoặc ít liên kết hơn.

- Nhãn đã được thêm cho các node quan trọng (các node trung tâm có giá trị trung tâm cao), ví dụ như:
  - **"Feeling Good," "Simple Man," "Back At One."**
- Điều này giúp dễ dàng nhận diện các node có ảnh hưởng lớn hoặc đóng vai trò quan trọng trong mạng.

Các node quan trọng được hiển thị với kích thước lớn hơn, giúp chúng nổi bật so với các node khác. Đây là cách hiệu quả để trực quan hóa tầm ảnh hưởng của từng node trong mạng.

Node có kích thước lớn thường nằm ở trung tâm cụm hoặc tại vị trí cầu nối giữa các cụm.

Bố trí sử dụng thuật toán **Force Atlas 2** hoặc tương tự, tạo sự phân bố đồng đều và dễ nhìn giữa các cụm.

Có sự phân tách rõ ràng giữa các cụm, giúp dễ dàng phân biệt các nhóm trong mạng.

Tuy nhiên, nhãn của một số node bị chồng chéo (ví dụ: trong cụm tím và đen), có thể gây khó khăn trong việc đọc thông tin.

Phần 4: Báo cáo và đánh giá

1. Louvain Algorithm:

- **Số lượng cộng đồng phát hiện:** 8
- **Modularity:** 0.819 (ổn định với và không có điều chỉnh độ phân giải)

2. Louvain with resolution adjustment:

- **Số lượng cộng đồng phát hiện:** 8 (tương tự Louvain thông thường)
- **Modularity tối đa:** 0.819 (không cải thiện so với Louvain tiêu chuẩn)

Ưu và nhược điểm của các thuật toán:

1. Louvain Algorithm:

- **Ưu điểm:**
  - Hiệu quả tính toán cao, phù hợp cho các mạng xã hội lớn.
  - Tự động xác định số lượng cộng đồng mà không cần tham số đầu vào.
  - Modularity tối ưu ở mức cao (0.819), cho thấy mạng lưới có cấu trúc cộng đồng rõ ràng.
- **Nhược điểm:**
  - Có thể bỏ qua các cụm nhỏ hoặc bị lệ thuộc vào mức độ phân giải (resolution).
  - Không luôn đảm bảo tìm được modularity cao nhất.

2. Girvan-Newman Algorithm:

- **Ưu điểm:**
  - Tốt trong việc phân tích mạng nhỏ và trung bình.
  - Modularity cao hơn (0.8224271), có thể phát hiện được các cộng đồng nhỏ mà Louvain bỏ qua.
  - Dễ hiểu về mặt thuật toán (dựa trên việc loại bỏ cầu nối).
- **Nhược điểm:**
  - Tốn kém tính toán, không phù hợp cho mạng xã hội lớn.
  - Số lượng cộng đồng (10) có thể quá chi tiết, gây khó khăn trong phân tích tổng thể.

Ý nghĩa của các cộng đồng trong ngữ cảnh mạng xã hội:

1. **Cộng đồng được phát hiện:**
    - Mỗi cộng đồng đại diện cho một nhóm nghệ sĩ hoặc các đối tượng có sự liên kết chặt chẽ với nhau trong mạng.
    - Ví dụ, trong cột **Nodes**, các nghệ sĩ như **Cam**, **Kevin Fowler**, và **Chris Cagle** có thể nằm trong cùng một cụm do có mối liên hệ về thể loại âm nhạc hoặc sở thích của người dùng.
2. **Ngữ cảnh mạng xã hội:**
    - Các cụm này có thể phản ánh:
        - **Mối quan tâm chung:** Người dùng theo dõi hoặc thích các nghệ sĩ trong cùng một cụm.
        - **Sự ảnh hưởng:** Nghệ sĩ có giá trị trung tâm cao trong một cụm có thể là nhân tố ảnh hưởng lớn đến người dùng trong mạng xã hội.
        - **Sự phân hóa:** Các cộng đồng khác nhau cho thấy sự phân hóa về sở thích hoặc mối quan tâm giữa các nhóm người dùng.
3. **Ý nghĩa modularity cao:**
    - Giá trị modularity cao (>0.8) cho thấy các cộng đồng được phân biệt rõ ràng trong mạng, đồng nghĩa với việc các nghệ sĩ hoặc đối tượng trong một cộng đồng có mối liên hệ mật thiết, ít tương tác với các cộng đồng khác.

Đề xuất phương pháp phân cụm phù hợp nhất:

Phương pháp được đề xuất: Louvain Algorithm

- **Lý do:**
    - Hiệu quả tính toán vượt trội, phù hợp với dữ liệu mạng xã hội lớn.
    - Phát hiện các cộng đồng rõ ràng mà không cần điều chỉnh thủ công.
    - Modularity đủ cao (0.819) để phân tích cấu trúc mạng và không quá chi tiết như Girvan-Newman.
    - Dễ dàng áp dụng và tùy chỉnh nếu cần so sánh hoặc phân tích sâu hơn