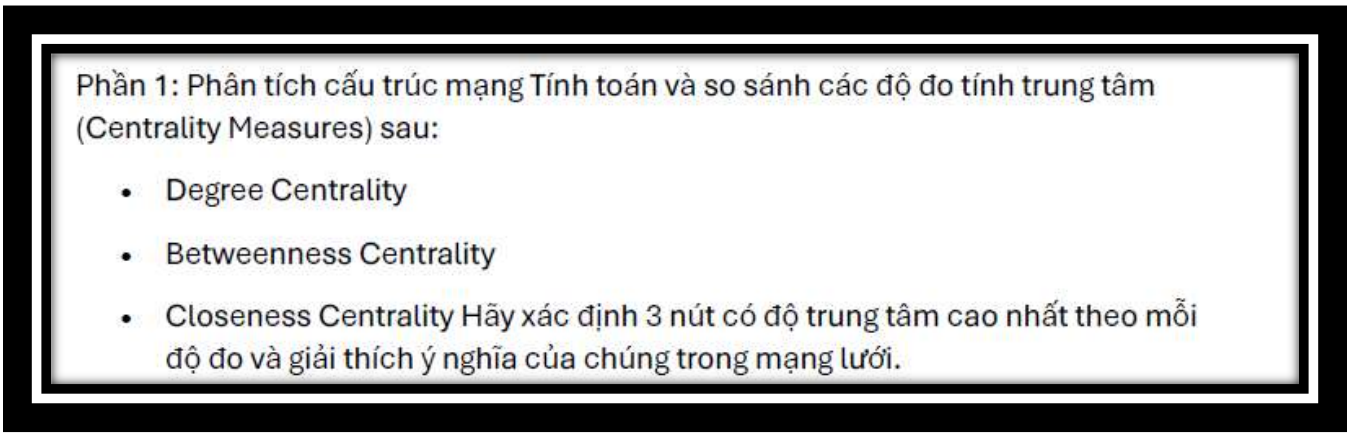
Lab 05-02

Lê Nguyễn Anh Nhật

2186400330

Phần 1



Tập dữ liệu sử dụng là Les Miserables

1. Degree Centrality:

Id	Label	Degree
11	Valjean	36
48	Gavroche	22
55	Marius	19
27	Javert	17
25	Thenardier	16
58	Enjolras	15
23	Fantine	15
62	Courfeyrac	13
64	Bossuet	13
63	Bahorel	12
65	Joly	12
57	Mabeuf	11
59	Combeferre	11
61	Feuilly	11
24	MmeThenard...	11
41	Eponine	11
26	Cosette	11
66	Grantaire	10
68	Gueulemer	10
69	Babet	10
70	Claquesous	10
0	Myriel	10
60	Prouvaire	9
71	Montparnasse	9
16	Tholomyes	9

Hình 1 - Degree Distribution

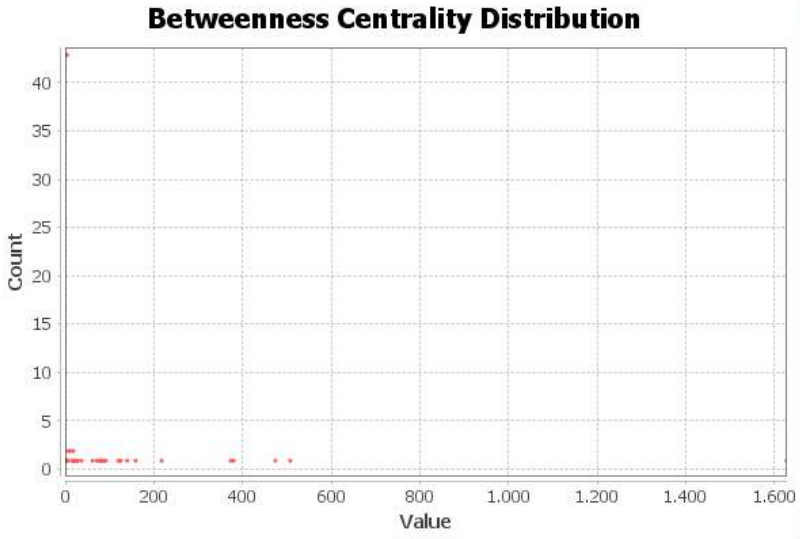
Ở hình này mặc dù không tình được Degree Centrality của từng nút nhưng nó thể hiện rằng một điều. Nút có id là 11 trong bộ dữ liệu Les Miserables có bậc cao nhất và vì thế điều hiển nhiên có có chỉ số DC cao nhất trong dữ liệu về mạng xã hội Les Miserable. Và công thức của Degree Centrality

$$DC(i) = \frac{\text{Degree of node } i}{\text{Total number of nodes in the network} - 1}$$

Dc càng cao thì đồng nghĩa với việc bậc sẽ càng cao khi so sánh với các nút khác

và ở đây nút có ID thứ 11 là cao nhất trong các id vì bậc của nó là 36

2. Betweenness Centrality



Id	Label	Betweenness Central...
11	Valjean	1624.4688
0	Myriel	504.0
48	Gavroche	470.570632
55	Marius	376.292593
23	Fantine	369.486942
25	Thenardier	213.468481
27	Javert	154.844945
51	MlleGillenor...	135.656944
58	Enjolras	121.277067
16	Tholomyes	115.793642
64	Bossuet	87.647903
24	MmeThenard...	82.656893
57	Mabeuf	78.834524
28	Fauchelevent	75.5
47	MmeBurgon	75.0
26	Cosette	67.819322
49	Gillenormand	57.600272
41	Eponine	32.739519
31	Simplice	24.624841
29	Bamatabois	22.916667
39	Pontmercy	19.7375
62	Courfeyrac	15.011035
68	Gueulemer	14.137094
69	Babet	14.137094
70	Claquesous	13.856142

*Hình 2- Bảng và kết quả của **Betweenness Centrality***

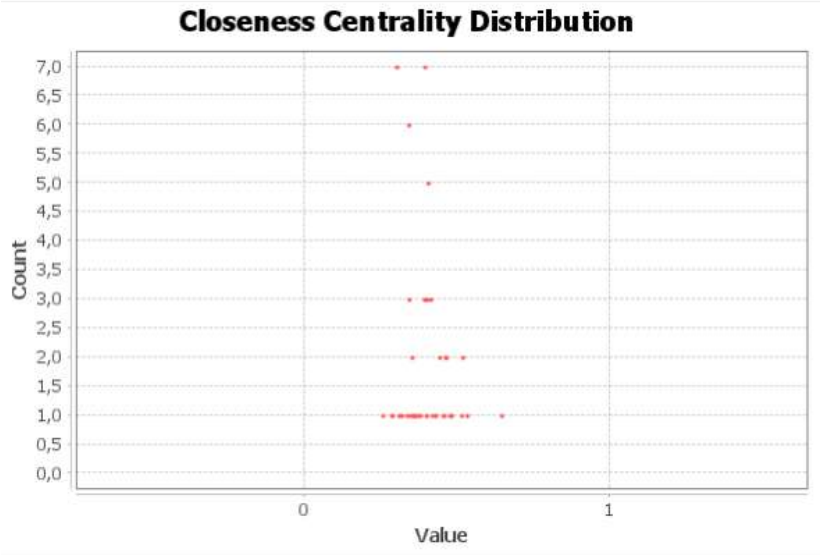
Dựa trên biểu đồ **Betweenness Centrality Distribution** và bảng dữ liệu đi kèm, chúng ta nhận thấy rằng mạng xã hội của **Les Misérables** có một số nút đóng vai trò cực kỳ quan trọng trong việc kết nối các nhóm nhân vật, trong khi phần lớn các nút còn lại có giá trị Betweenness Centrality thấp. Nhân vật **Jean Valjean** có giá trị Betweenness Centrality cao nhất (**1624.4688**), vượt xa các nhân vật khác. Điều này chỉ ra rằng Valjean là trung tâm kết nối của mạng lưới, đóng vai trò cầu nối giữa nhiều nhóm nhân vật và cộng đồng riêng biệt trong câu chuyện. Các nhân vật **Myriel (504.0)**, **Gavroche (470.57)**, và **Marius (376.29)** cũng có giá trị Betweenness Centrality cao. Điều này cho thấy họ là các nhân vật quan trọng trong việc liên kết các tuyến truyện hoặc các cụm nhân vật khác nhau.

Hầu hết các nút trong mạng có giá trị Betweenness Centrality rất thấp (dưới 200), như **MmeBurgon (75.0)** hoặc **Eponine (32.73)**. Điều này cho thấy vai trò trung gian của các nhân vật này rất nhỏ, chủ yếu là thành viên trong các cụm cụ thể mà không kết nối

đáng kể đến các cụm khác. Biểu đồ cũng cho thấy sự phân phối không đồng đều, khi chỉ một số ít nút có giá trị rất cao. Điều này phản ánh một mạng lưới mang tính phân cấp, nơi các nhân vật chính như Valjean giữ vai trò trung gian nổi bật.

Trong bối cảnh câu chuyện, **Jean Valjean** là nhân vật trung tâm cả về mặt cấu trúc mạng lưới lẫn cốt truyện, vì anh có mối quan hệ với hầu hết các nhân vật chính và phụ. Tương tự, **Gavroche** và **Marius** đóng vai trò quan trọng trong việc kết nối các nhóm như cách mạng và các nhân vật khác. Những nhân vật này nếu bị loại bỏ, mạng lưới có thể bị phân tách thành các cụm riêng lẻ, làm mất đi sự liên kết tổng thể. Trong khi đó, các nhân vật với giá trị Betweenness thấp có vai trò hỗ trợ và ít ảnh hưởng đến cấu trúc chung.

3. Closeness Centrality



Id	Label	Closeness Central...
11	Valjean	0.644068
55	Marius	0.531469
25	Thenardier	0.517007

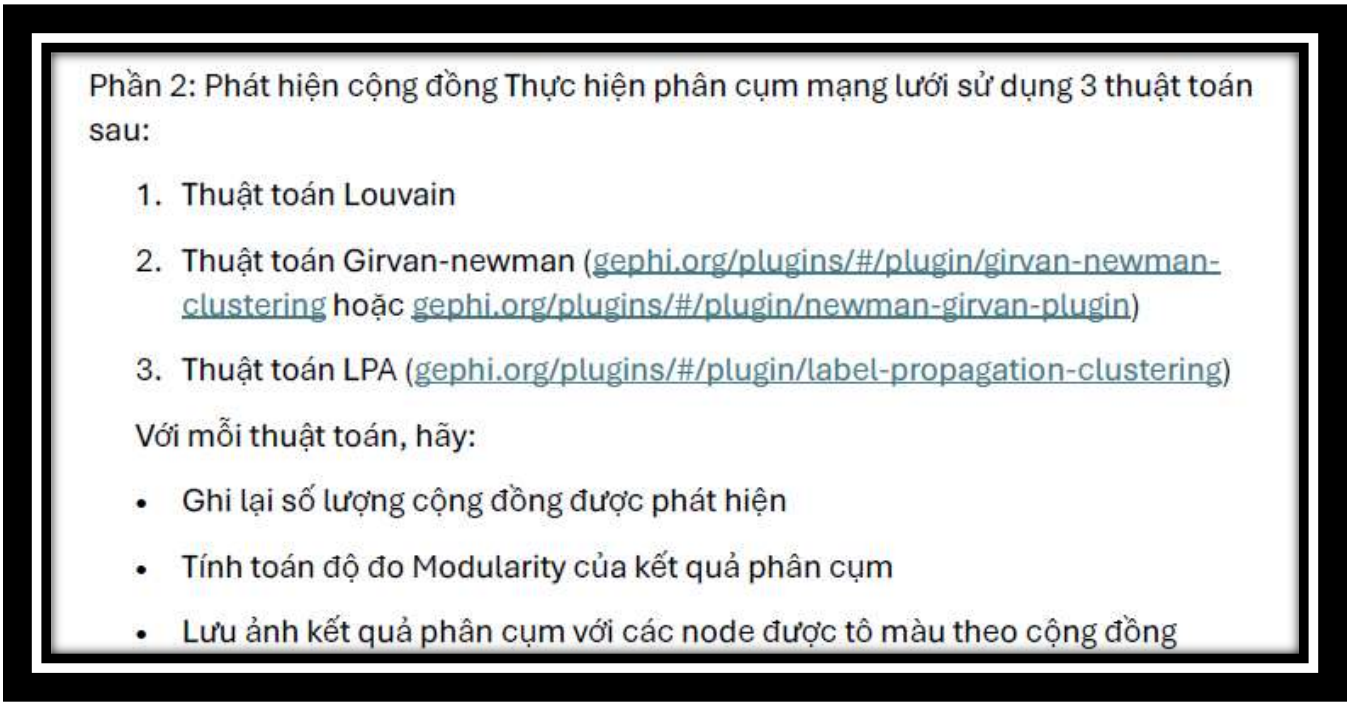
Hình 3 - Bảng và kết quả của Closeness Centrality

Hình ảnh trên trình bày phân bố Closeness Centrality của một mạng xã hội cùng với bảng giá trị Closeness Centrality của các cá nhân trong mạng. Closeness Centrality đo

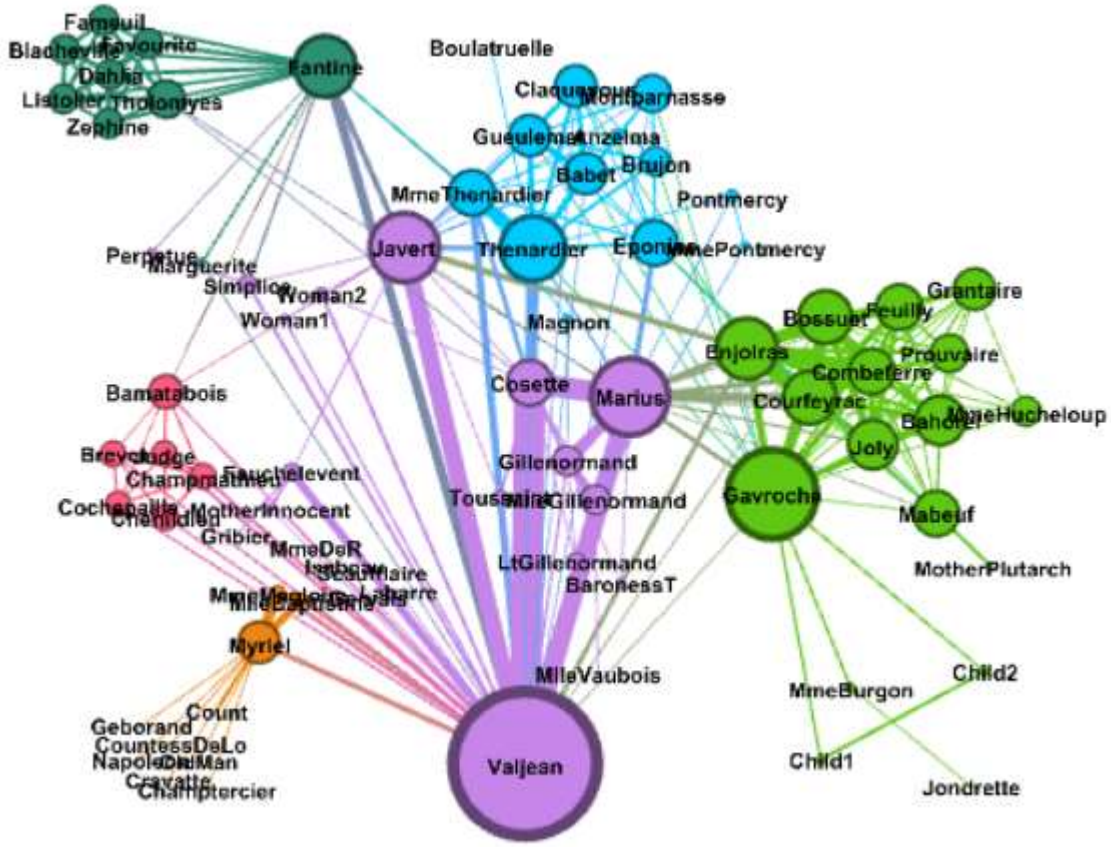
lường mức độ gần gũi của một nút đến tất cả các nút khác trong mạng, thể hiện khả năng tiếp cận thông tin nhanh chóng.

Phân phối cho thấy phần lớn giá trị Closeness Centrality tập trung ở khoảng trung bình, với một số ít giá trị cao (gần 0,6 - 0,7). Điều này gợi ý rằng chỉ một vài cá nhân có vị trí chiến lược với khả năng tiếp cận cao (như "Valjean" với 0,644068), trong khi đa số các cá nhân có khả năng tiếp cận thông tin thấp hơn, người còn lại là Myriel và Thernardier có chủ số cao thứ 2 và 3. Điều này phản ánh một cấu trúc mạng có thể không đồng nhất, nơi một số nút đóng vai trò trung tâm hơn so với phần còn lại.

Phần 2



1. Louvain

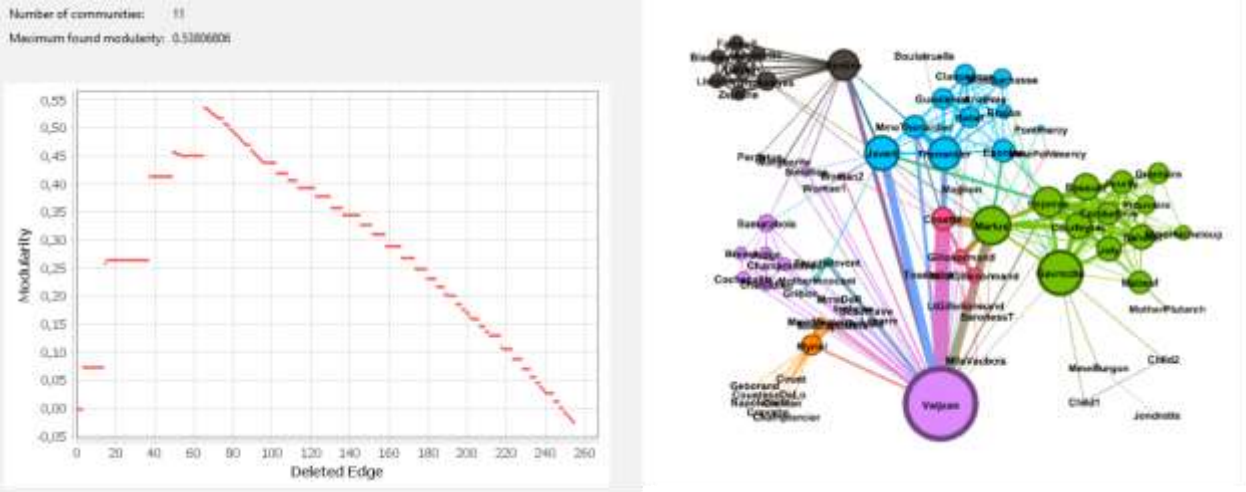


Hình ảnh 4 – Ảnh Louvain phân tách các cộng đồng

Modularity: 0,566
Modularity with resolution: 0,566
Number of Communities: 6

Hình ảnh trên minh họa mạng xã hội được phân cụm bằng thuật toán Louvain, với mỗi cụm được biểu diễn bằng màu sắc khác nhau. Phân cụm này cho thấy các nhóm nhân vật có mối quan hệ chặt chẽ hơn trong mạng. Các nút lớn, như "Valjean," "Fantine," và "Gavroche," đóng vai trò trung tâm trong cụm của mình, thể hiện tầm quan trọng và ảnh hưởng cao. Chỉ số modularity được tối ưu hóa, phản ánh cấu trúc phân cụm rõ rệt, giúp hiểu sâu hơn về cách các cá nhân kết nối trong mạng xã hội.

2. Girvan-Newman

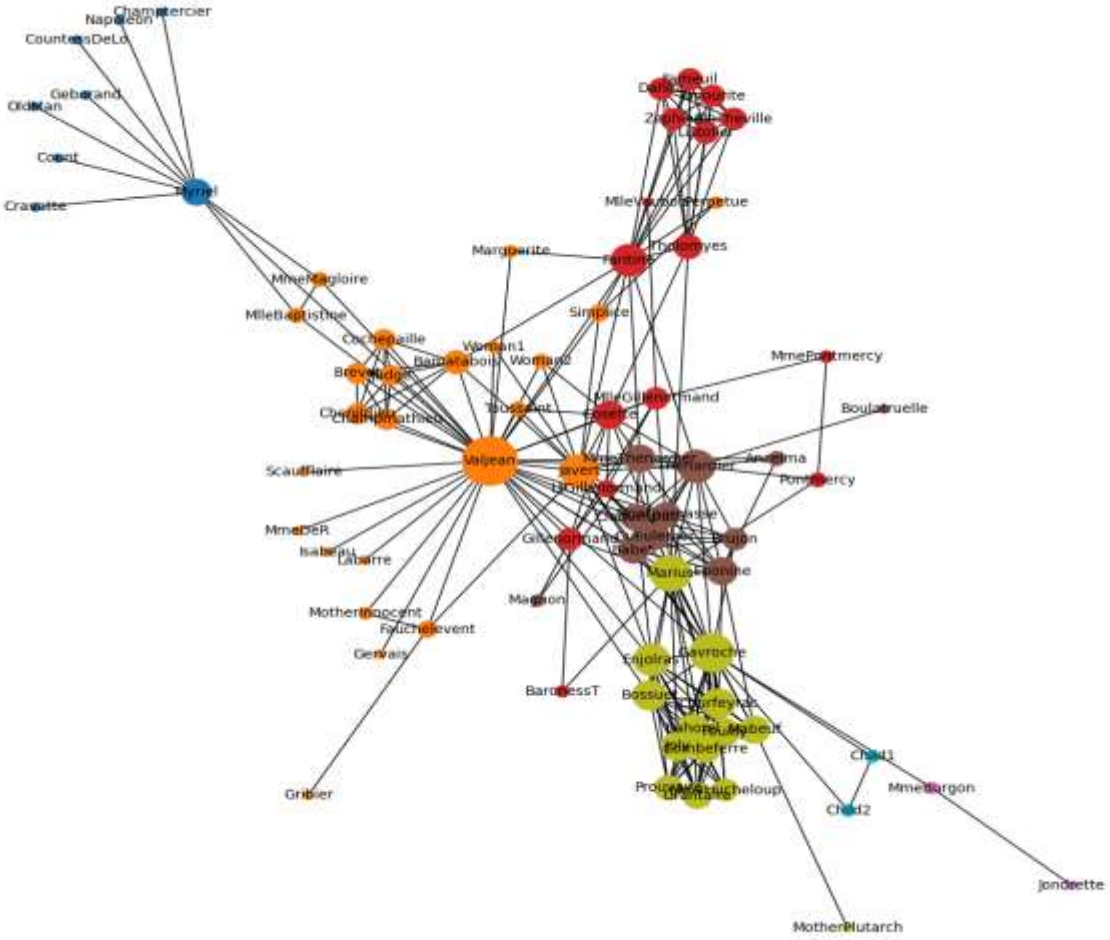


Hình ảnh 5 - Ảnh Girvan-Newman

Hình ảnh trên thể hiện quá trình phân cụm mạng xã hội dựa trên thuật toán Girvan-Newman và giá trị modularity. Biểu đồ bên trái cho thấy sự thay đổi của chỉ số modularity khi các cạnh được xóa dần, đạt giá trị cao nhất là 0,53806806 với 11 cộng đồng được xác định.

Mạng bên phải minh họa kết quả phân cụm, trong đó các nút được chia thành các cộng đồng tương ứng. Các cụm chính, như nhóm "Valjean," "Fantine," và "Gavroche," đại diện cho các nhóm nhân vật có mối liên kết chặt chẽ. Kết quả này cho thấy cấu trúc mạng có tính cộng đồng cao, giúp phân tích sâu hơn về cách các cá nhân tương tác trong mạng.

3. LPA



Hình 6 - Ảnh phân cụm LPA

Có kết quả như sau:

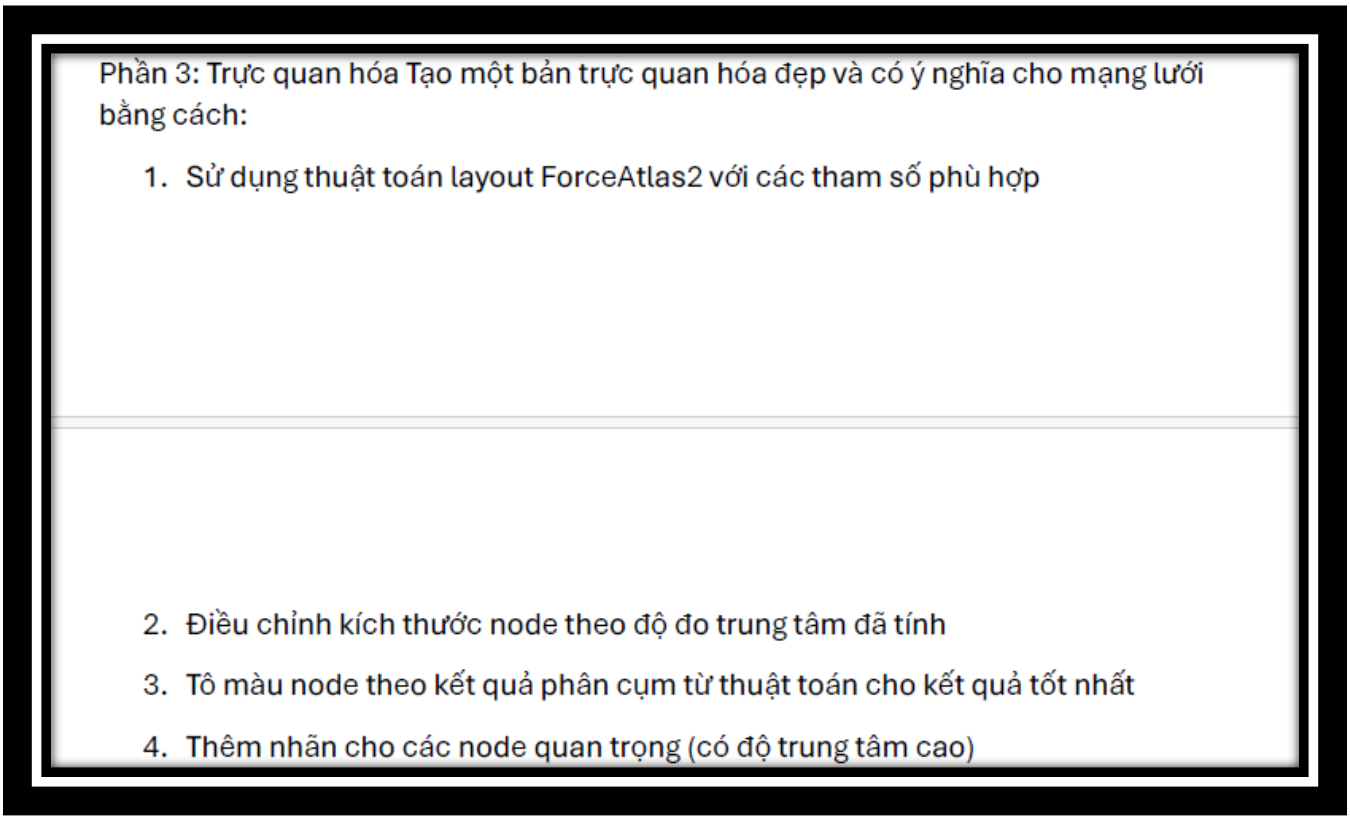
Số cộng đồng: 7

Số modularity: 0.5267

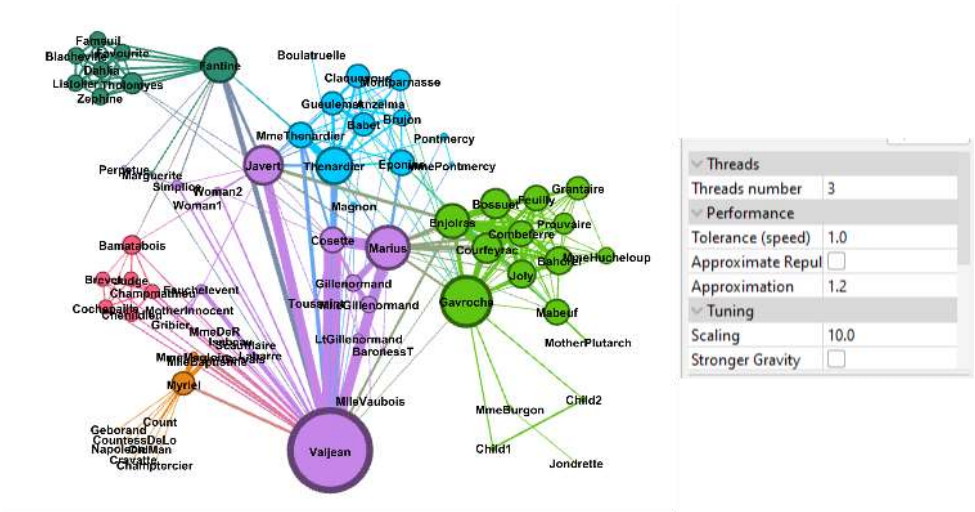
Điều này cho thấy các cộng đồng được xác định có mức độ liên kết nội bộ cao và mối

quan hệ giữa các cộng đồng khác nhau được phân tách rõ ràng, dựa trên giá trị modularity đạt được. Kết quả này phản ánh hiệu quả của thuật toán LPA trong việc phân cụm đồ thị.

Phần 3



1. Layout ForceAtlas2



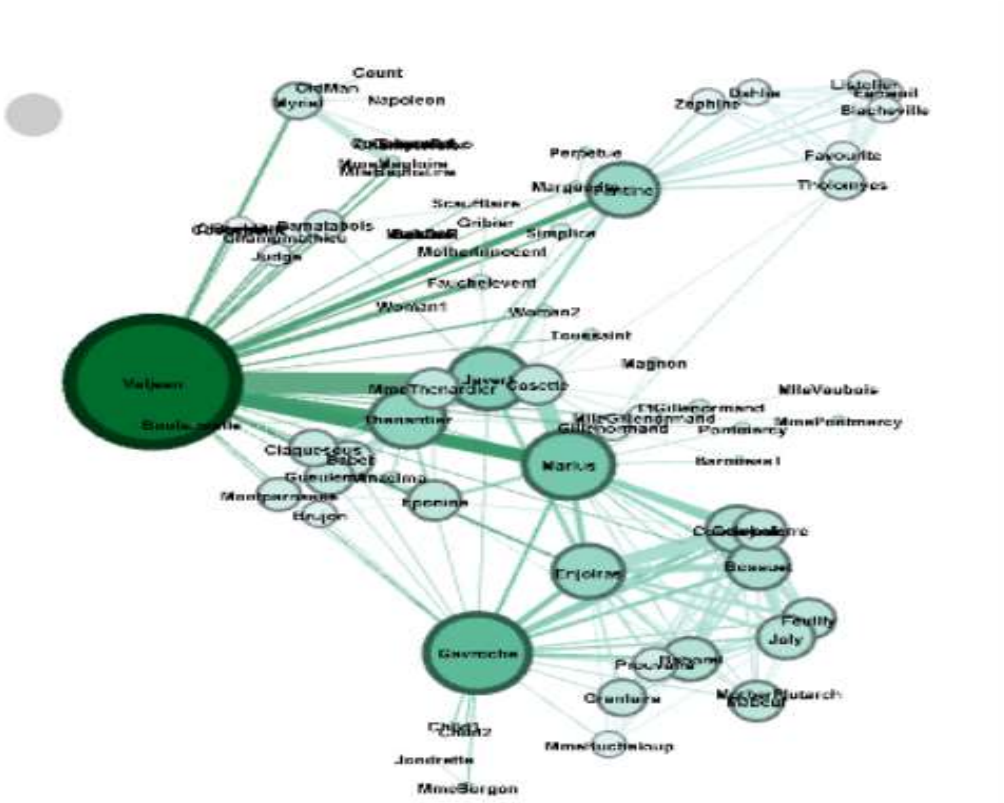
Hình 7 – Ảnh Layout ForceAtlat2

Điều chỉnh tham số ForceAtlas2:

- Scaling: Tăng giá trị này nếu đồ thị quá rối để tăng khoảng cách giữa các node.
- Gravity: Tăng để giữ các node gần trung tâm hơn.
- Prevent Overlap: Chọn mục này để ngăn chồng lấn các node.
- Edge Weight Influence: Điều chỉnh để các cạnh có trọng số lớn ảnh hưởng mạnh hơn.

Dùng các tham số này sẽ tạo ra được hình ảnh như trên với vài kéo thả hình ảnh sẽ đẹp.

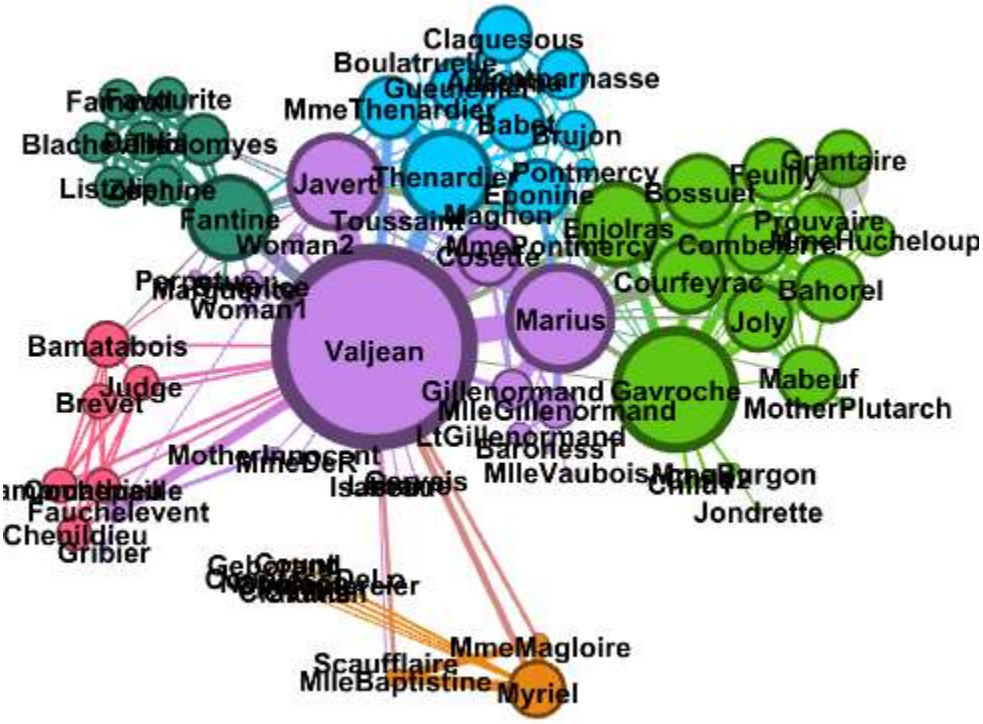
2. Theo độ đo trung tâm



Hình 8 – Điều chỉnh theo trung tâm

Việc điều chỉnh theo trung tâm khiến node Vali sáng và tông màu được chỉnh từ nhạt dần cho đến đậm dần để thấy các node max và min nhìn sẽ rất đẹp.

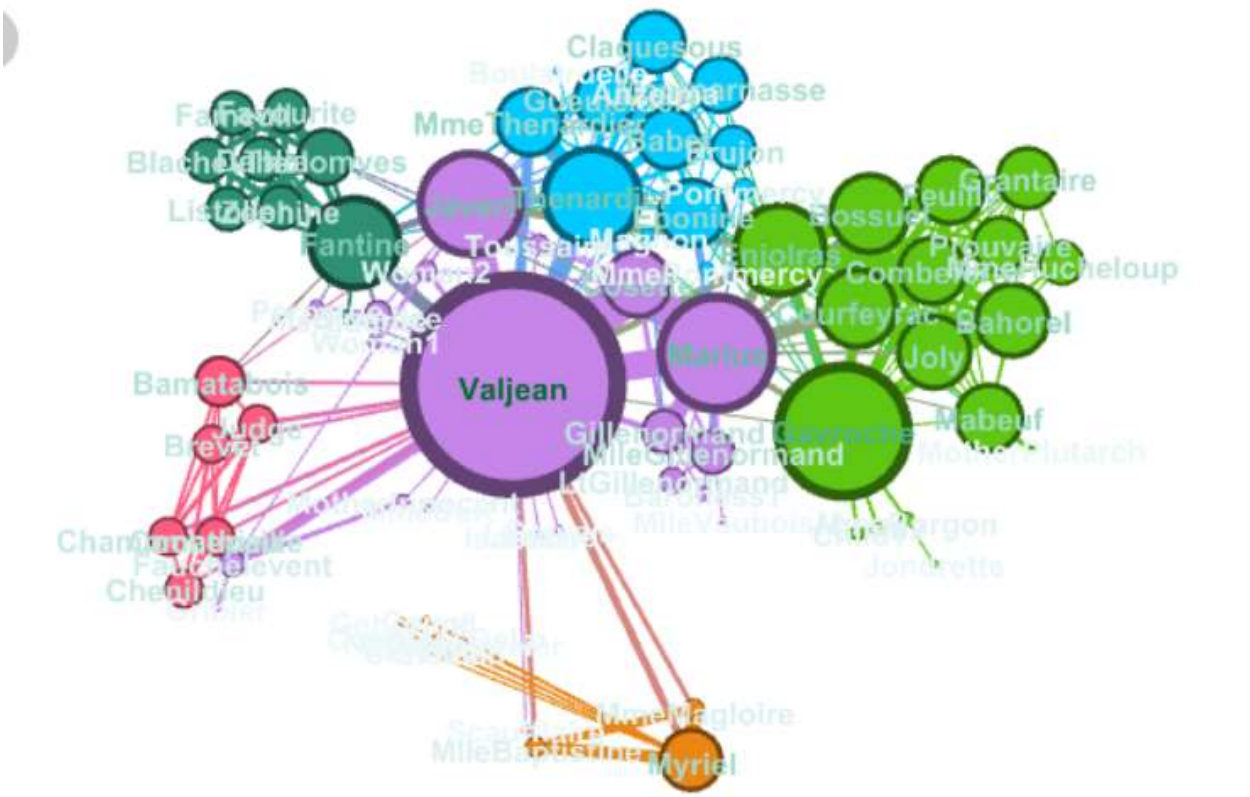
3. Tô màu theo phân cụm



Hình 9 - Màu phân cụm Louvain

Dựa trên kết quả phân cụm, ta nhận thấy rằng đồ thị được chia thành 6 cụm khác nhau, mỗi cụm được phân biệt bằng một màu sắc. Ở trung tâm là cụm có bậc (degree) cao nhất, thể hiện vai trò trung tâm trong mạng lưới. Di chuyển từ trong ra ngoài, các cụm có sắc xanh lá và xanh dương xuất hiện đông đảo hơn, phản ánh sự phân bố của các nút với mức độ liên kết thấp hơn nhưng vẫn duy trì sự kết nối trong hệ thống.

4. Thêm nhãn cho các node quan trọng



Hình 10 – Node quan trọng nhất

Phần 4

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

1. So Sánh

	Lovain	Girvan-Newman	LPA
Số cộng đồng	6	11	7
Modularity	0.566	0.538	0.526

1. **Louvain Algorithm (Louvain)**
   - **Ưu điểm:**
     - **Hiệu quả và nhanh chóng**: Louvain được biết đến là thuật toán phân cụm mạng hiệu quả về mặt tính toán. Nó có thể xử lý mạng lưới lớn mà không yêu cầu quá nhiều tài nguyên tính toán.
     - **Chất lượng cộng đồng cao**: Với modularity đạt 0.566, Louvain cho kết quả phân cụm chất lượng khá tốt, đặc biệt là khi áp dụng cho các mạng xã hội hoặc các dữ liệu có cấu trúc phức tạp.
   - **Nhược điểm:**
     - **Không ổn định với số lượng cộng đồng**: Louvain có thể không cho ra kết quả ổn định khi số lượng cộng đồng thay đổi hoặc với các đồ thị không đồng đều.

- **Phụ thuộc vào khởi tạo**: Kết quả có thể thay đổi với những lần chạy khác nhau, vì thuật toán này có thể bị ảnh hưởng bởi các bước khởi tạo ngẫu nhiên.

2. **Girvan-Newman**
   - **Ưu điểm:**
     - **Đảm bảo tính chính xác**: Girvan-Newman phân tích đồ thị bằng cách loại bỏ các cạnh có độ trung tâm cao, tạo ra cộng đồng rõ ràng. Điều này có thể giúp phát hiện các cộng đồng chính xác, đặc biệt trong các đồ thị có cấu trúc chặt chẽ.
     - **Giải thích được lý thuyết cộng đồng**: Phương pháp này dễ hiểu và có cơ sở lý thuyết rõ ràng.
   - **Nhược điểm:**
     - **Chạy chậm**: Với thời gian tính toán lâu, thuật toán này không thích hợp với các mạng lưới có kích thước lớn, vì nó yêu cầu tính toán nhiều lần độ trung tâm của các cạnh.
     - **Không ổn định trong các đồ thị lớn**: Girvan-Newman dễ bị suy giảm hiệu quả khi áp dụng vào các mạng lưới lớn và phức tạp, dẫn đến thời gian chạy lâu.
3. **LPA (Label Propagation Algorithm)**
   - **Ưu điểm:**
     - **Nhanh chóng và tiết kiệm tài nguyên**: LPA là một thuật toán phân cụm đơn giản và nhanh chóng, không yêu cầu nhiều tài nguyên tính toán. Nó có thể xử lý các đồ thị lớn một cách hiệu quả.
     - **Không cần thông tin ban đầu**: LPA không yêu cầu thông tin về số lượng cộng đồng ban đầu, giúp linh hoạt hơn trong nhiều trường hợp.
   - **Nhược điểm:**
     - **Chất lượng cộng đồng thấp**: Modularity của LPA (0.526) thấp hơn so với Louvain, điều này cho thấy chất lượng phân cụm không cao bằng, đặc biệt trong các đồ thị phức tạp.

- **Không ổn định và dễ bị ảnh hưởng bởi cấu trúc đồ thị**: LPA có thể cho ra kết quả không ổn định, đặc biệt nếu đồ thị có cấu trúc phức tạp hoặc thiếu sự rõ ràng trong các nhóm cộng đồng.

Kết luận:

- **Louvain** là lựa chọn tốt nhất khi cần một phương pháp phân cụm nhanh và chất lượng, đặc biệt trong các trường hợp có mạng xã hội lớn với cấu trúc cộng đồng rõ ràng.
- **Girvan-Newman** thích hợp với những tình huống cần phân tích chính xác cộng đồng, nhưng không phù hợp với đồ thị lớn do tính toán tốn kém.
- **LPA** là một phương pháp nhanh chóng và dễ triển khai, nhưng đôi khi có thể cho ra kết quả chất lượng thấp và không ổn định, đặc biệt trong các đồ thị phức tạp.

2. Giải thích ý nghĩa cộng đồng

Trong ngữ cảnh của mạng xã hội, các cộng đồng được phát hiện thông qua ba thuật toán phân cụm — Louvain, Girvan-Newman và LPA — mang ý nghĩa đặc biệt, phản ánh cách mà các cá nhân hoặc nhóm người trong mạng xã hội tương tác và liên kết với nhau. Mỗi cộng đồng trong mạng xã hội đại diện cho một nhóm người có sự kết nối mạnh mẽ, có thể là các nhóm bạn bè, đồng nghiệp, hoặc những người cùng chia sẻ sở thích và giá trị tương tự. Các thuật toán phân cụm giúp xác định những nhóm này bằng cách phân tích cấu trúc kết nối giữa các người dùng (nodes) và mối quan hệ (edges) trong mạng.

- **Louvain** với khả năng phát hiện các cộng đồng rõ ràng, giúp nhận diện các nhóm có sự gắn kết mạnh mẽ nhất trong mạng xã hội. Kết quả từ thuật toán này cho thấy các cộng đồng có tính đồng nhất cao, điều này có thể ứng dụng trong việc phát hiện các nhóm người dùng tương tác nhiều, như các nhóm bạn bè thân thiết hoặc các cộng đồng hoạt động theo sở thích chung. Việc xác định chính xác các cộng đồng

này có thể giúp hiểu rõ hơn về sự lan truyền thông tin, xu hướng và hành vi trong mạng xã hội.

- **Girvan-Newman**, mặc dù cho ra nhiều cộng đồng hơn (11 cộng đồng trong kết quả của bài toán), lại giúp phân tích chi tiết hơn cấu trúc mạng xã hội. Thuật toán này đặc biệt hiệu quả trong việc phát hiện các cộng đồng có mối quan hệ chặt chẽ, nhưng cũng dễ dàng xác định những mối quan hệ lỏng lẻo hoặc những nhóm người có ít sự tương tác hơn. Điều này có thể giúp xác định các nhóm nhỏ hoặc những phân nhóm ít rõ ràng, những người dùng có thể không tương tác nhiều với nhau nhưng vẫn tồn tại trong mạng xã hội.
- **LPA (Label Propagation Algorithm)**, mặc dù cho kết quả phân cụm đơn giản hơn với chỉ 7 cộng đồng, lại có ưu điểm trong việc xác định các cộng đồng dựa trên sự tương tác tự nhiên mà không cần quá nhiều thông tin ban đầu. Các cộng đồng phát hiện bởi LPA có thể bao gồm các nhóm người dùng có sở thích chung hoặc sự đồng thuận trong hành vi, nhưng vì LPA thường cho kết quả không ổn định và có chất lượng thấp hơn, nên các cộng đồng này đôi khi có thể không rõ ràng như trong Louvain hay Girvan-Newman.

Tóm lại, các cộng đồng được phát hiện trong mạng xã hội thông qua ba thuật toán trên có ý nghĩa rất quan trọng trong việc hiểu rõ hơn về cấu trúc xã hội của người dùng. Các cộng đồng này giúp các nhà nghiên cứu hoặc các tổ chức nhận diện các nhóm người có hành vi tương tự, từ đó đưa ra các chiến lược tiếp thị, nghiên cứu hành vi người dùng, hay phát hiện các xu hướng trong xã hội. Tuy nhiên, việc lựa chọn thuật toán phù hợp phụ thuộc vào mục tiêu phân tích và tính chất của mạng xã hội mà ta đang nghiên cứu.

3. Đề xuất

Dựa trên kết quả phân cụm của ba thuật toán (Louvain, Girvan-Newman và LPA) và các đặc điểm của dữ liệu mạng xã hội, **Louvain** sẽ là phương pháp phân cụm phù hợp nhất cho loại dữ liệu này. Dưới đây là lý do chi tiết:

1. Khả năng phát hiện cộng đồng chất lượng cao

Louvain là một thuật toán phân cụm tối ưu về mặt modularity, với giá trị **0.566** trong kết quả thử nghiệm, cao hơn cả Girvan-Newman (0.538) và LPA (0.526). Modularity cao cho thấy Louvain có khả năng phát hiện ra những cộng đồng có sự gắn kết mạnh mẽ, một yếu tố quan trọng trong mạng xã hội. Các cộng đồng này có thể tương ứng với những nhóm người dùng có sự tương tác và quan tâm chung, như các nhóm bạn bè, nhóm sở thích, hoặc nhóm hoạt động xã hội.

2. Hiệu quả trong mạng xã hội lớn

Trong các mạng xã hội có số lượng người dùng lớn, **Louvain** cho phép phân cụm nhanh chóng và hiệu quả mà không cần phải tính toán quá tốn kém. Điều này đặc biệt quan trọng khi mạng xã hội có quy mô lớn với hàng triệu người dùng, vì Louvain có thể xử lý các đồ thị phức tạp một cách tối ưu mà không gặp phải vấn đề về tài nguyên tính toán.

3. Khả năng tự động phát hiện số lượng cộng đồng

Louvain không yêu cầu số lượng cộng đồng ban đầu mà tự động tìm ra số cộng đồng tối ưu dựa trên cấu trúc kết nối của đồ thị. Điều này giúp giảm thiểu sự phụ thuộc vào các thông số đầu vào và phù hợp với các mạng xã hội có cấu trúc phân bố không đồng đều hoặc chưa xác định rõ số lượng nhóm người dùng.

4. Ổn định và độ chính xác cao

Mặc dù các thuật toán như Girvan-Newman cũng có thể phát hiện cộng đồng chính xác, nhưng thuật toán này tốn kém về mặt tính toán, đặc biệt là với mạng xã hội lớn. LPA, mặc dù nhanh và tiết kiệm tài nguyên, nhưng lại có kết quả phân cụm không ổn định và chất lượng cộng đồng thấp. Do đó, **Louvain** là sự lựa chọn cân bằng giữa hiệu suất tính toán và chất lượng phân cụm, đặc biệt khi áp dụng cho dữ liệu mạng xã hội.