

**Khoa: Công nghệ Thông tin****Social Networks Course****LAB 5.2 (Gephi)****Họ và tên sinh viên:**

Nguyễn Hoàng Khang ..... MSSV: 2186400244.....Lớp: 21DKHA1

Ngành : Khoa học dữ liệu

**PHẦN 1. Top3 theo từng Metric****- Cao nhất theo Degree Centrality (~Degree):**

Id	Label	Interval	Eccentricity	Closeness Centrality	Harmonic Closeness Centrality	Betweenness Centrality	Degree ▾	Weighted Degree	Cluter-ID
14	Grin		6.0	0.376543	0.494262	0.061972	12	12.0	4
37	SN4		6.0	0.398693	0.498361	0.13857	11	11.0	4
45	Topless		7.0	0.346591	0.468462	0.04067	11	11.0	3

- **Node 14 (Grin): Degree = 12**
- **Node 37 (SN4): Degree = 11**
- **Node 45 (Topless): Degree = 11**

**Nhận xét chi tiết:****Node 14 (Grin):**

- Với **Degree = 12**, Grin là node có số lượng kết nối trực tiếp nhiều nhất trong mạng lưới cá heo.
- Điều này cho thấy Grin đóng vai trò là trung tâm chính về kết nối trực tiếp, có thể đảm nhiệm việc giao tiếp hoặc liên lạc với nhiều thành viên khác trong mạng lưới.

**Node 37 (SN4):**

- SN4 vẫn giữ vai trò quan trọng với **Degree = 11**.
- Dựa trên vị trí của nó trong các chỉ số trung tâm khác như Betweenness Centrality và Closeness Centrality, SN4 vừa có số kết nối trực tiếp cao vừa là một trung gian kết nối quan trọng giữa các node khác.

**Node 45 (Topless):**

- Topless cũng có **Degree = 11**, cho thấy nó là một cá heo có nhiều kết nối trực tiếp, tương đương với SN4.

- Tuy nhiên, giá trị **Betweenness Centrality** thấp hơn (0.04067), điều này cho thấy rằng vai trò trung gian của Topless không mạnh như SN4.

### Ý nghĩa trong mạng lưới:

- Grin** là trung tâm với số lượng kết nối cao nhất, có thể đại diện cho một cá heo có vai trò "leader" hoặc "connector" mạnh mẽ nhất trong mạng lưới.
- SN4** không chỉ có nhiều kết nối mà còn đóng vai trò quan trọng trong việc truyền tải thông tin giữa các cụm khác nhau.
- Topless**, mặc dù có số lượng kết nối cao, nhưng không đóng vai trò trung gian quan trọng, có thể là thành viên chính trong một cụm cụ thể nhưng không ảnh hưởng mạnh đến các cụm khác.

### Số sánh với các chỉ số khác:

- Grin** với Degree cao nhất cũng có **Betweenness Centrality** = **0.061972**, thấp hơn SN4, cho thấy vai trò trung gian của Grin bị hạn chế hơn.
- Topless**, với **Betweenness Centrality** = **0.04067**, có ảnh hưởng trung gian yếu nhất trong nhóm.

- Cao nhất theo Closeness Centrality:

Id	Label	Interval	Eccentricity	Closeness Centrality	Harmonic Closeness Centrality	Betweenness Centrality	Degree	Weighted Degree	Cluter-ID
36	SN100		5.0	0.417808	0.480055	0.248237	7	7.0	4
40	SN9		5.0	0.403974	0.480055	0.14315	8	8.0	4
37	SN4		6.0	0.398693	0.498361	0.13857	11	11.0	4

- Top 3 Actor Nodes:**

- Node 36 (SN100): Closeness Centrality = 0.418
- Node 40 (SN9): Closeness Centrality = 0.404
- Node 37 (SN4): Closeness Centrality = 0.399

- Giải thích:**

- Node 36 (SN100) có Closeness Centrality cao nhất, cho thấy nó có khả năng tiếp cận các node khác nhanh hơn so với các node khác trong mạng.
- Node 40 (SN9) và Node 37 (SN4) cũng có Closeness Centrality cao, cho thấy chúng nằm ở vị trí gần trung tâm của mạng lưới.

- Ý nghĩa trong mạng lưới:**

- Các node này có thể tiếp cận nhanh chóng với các thành viên khác trong mạng lưới, đóng vai trò như điểm trung tâm để lan truyền thông tin hoặc tài nguyên.

- Cao nhất theo Betweenness Centrality:

Id	Label	Interval	Eccentricity	Closeness Centrality	Harmonic Closeness Centrality	Betweenness Centra...	Degree	Weighted Degree	Cluter-ID
36	SN100		5.0	0.417808	0.480055	0.248237	7	7.0	4
1	Beescratch		5.0	0.371951	0.448634	0.213324	8	8.0	0
40	SN9		5.0	0.403974	0.480055	0.14315	8	8.0	4
37	SN4		6.0	0.398693	0.498361	0.13857	11	11.0	4

### • Top 3 Actor Nodes:

- Node 36 (SN100): Betweenness Centrality = 0.248
- Node 1 (Beescratch): Betweenness Centrality = 0.213
- Node 40 (SN9): Betweenness Centrality = 0.143

### • Giải thích:

- Node 36 (SN100) có Betweenness Centrality cao nhất, chứng tỏ nó là cầu nối quan trọng giữa các nhóm/cộng đồng trong mạng. Các đường đi ngắn giữa các node khác thường đi qua node này.
- Node 1 (Beescratch) và Node 40 (SN9) có Betweenness Centrality tương đối cao, cho thấy chúng cũng đóng vai trò trung gian trong việc kết nối các phần khác nhau của mạng.

### • Ý nghĩa trong mạng lưới:

- Các node này có vai trò như "broker" hoặc "mediator", đảm bảo sự gắn kết và truyền tải thông tin giữa các cụm (clusters) hoặc cộng đồng trong mạng lưới.

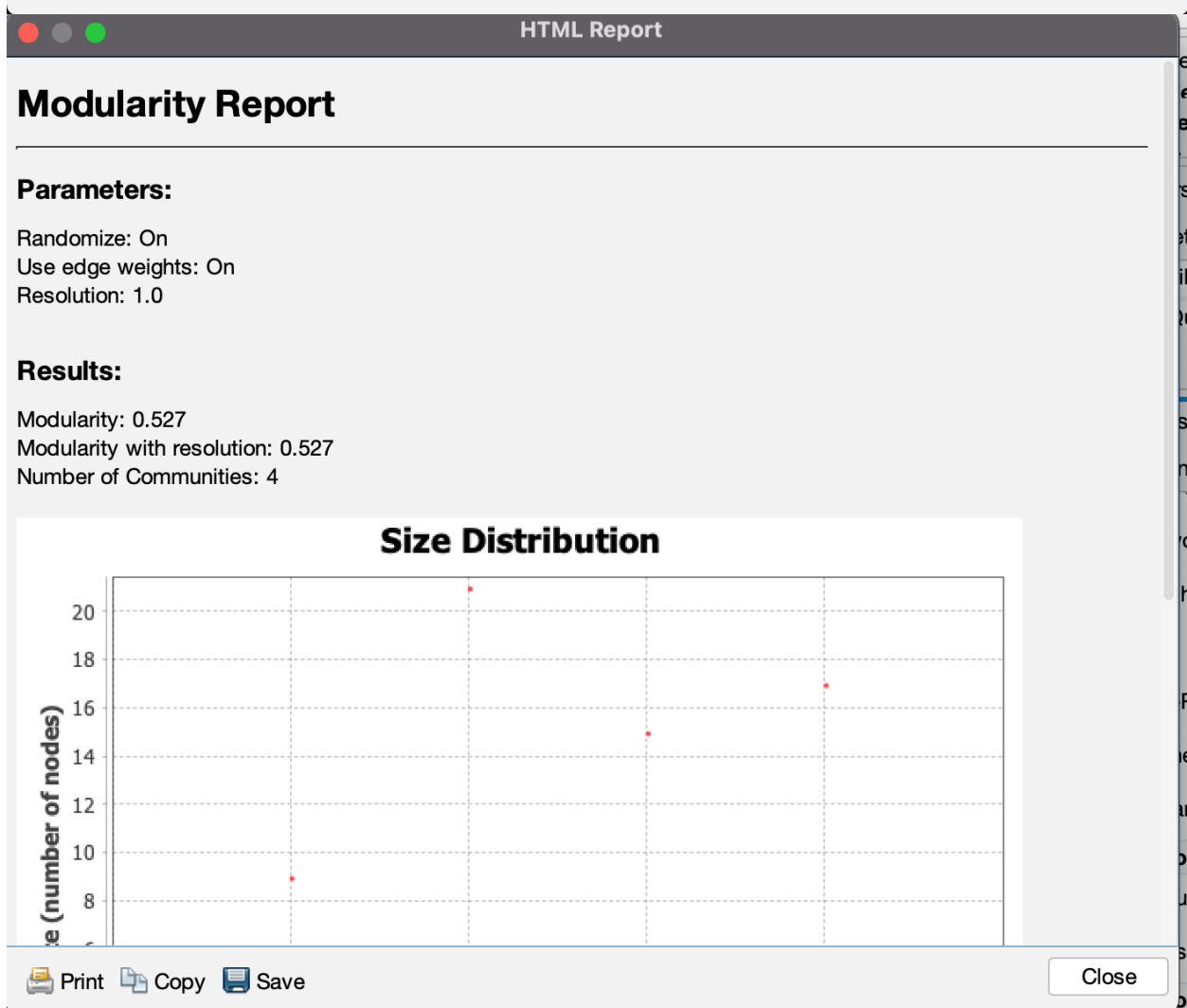
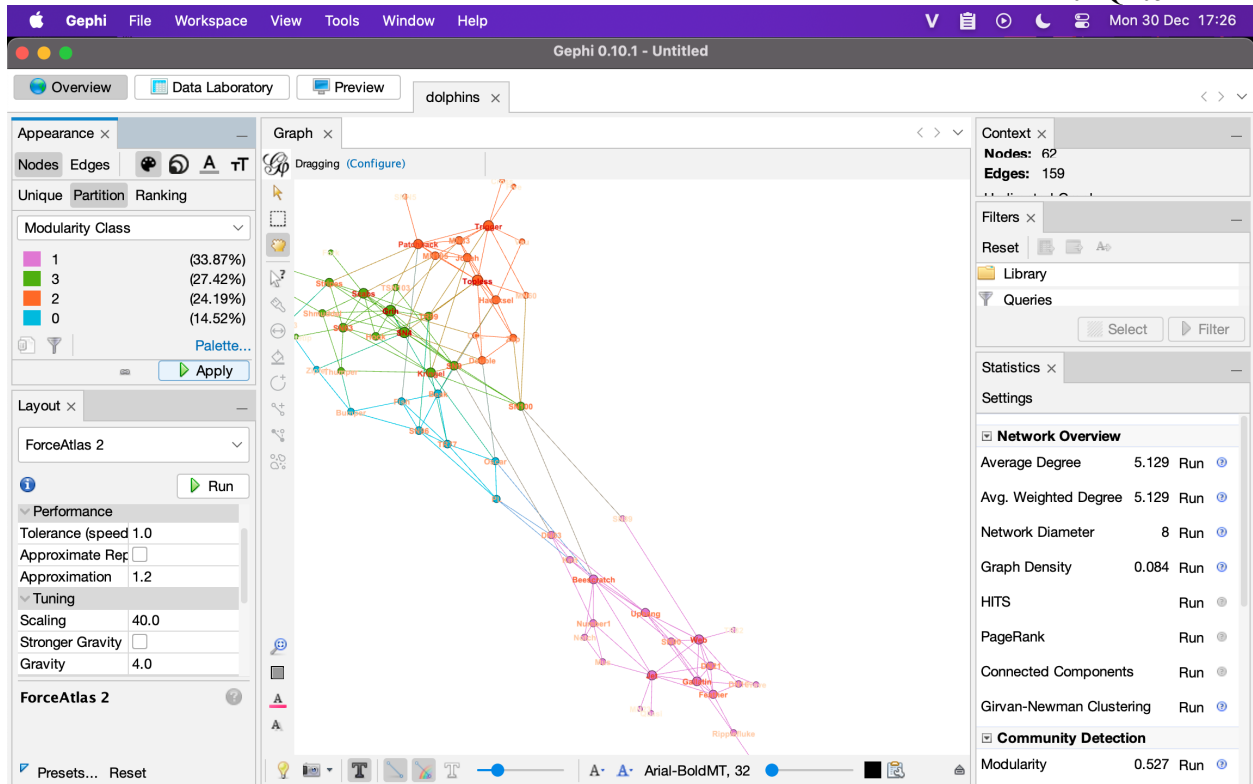
### Kết luận:

- Node **Grin** (14) là trung tâm chính về kết nối trực tiếp trong mạng lưới.
- Node **SN4** (37) là một trong những node quan trọng nhất, vừa có nhiều kết nối vừa đảm bảo liên kết giữa các cụm.
- Node **Topless** (45), dù có nhiều kết nối, nhưng vai trò trung gian yếu, cho thấy ảnh hưởng của nó chỉ tập trung trong một cộng đồng hoặc khu vực cụ thể.

Hệ quả là, Grin có thể là "nucleus" của mạng lưới, trong khi SN4 vừa là "connector" vừa là cầu nối trung gian giữa các cộng đồng.

## PHẦN 2:

### 2.1 Louvian:



## 2.2 Givan-Newman:

HTML Report

## Girvan-Newman Report

---

### Parameters:

Respect edge type for shortest path betweenness: no

Respect parallel edges for shortest path betweenness: no

Respect edge type for modularity computation: no

Respect parallel edges for modularity computation: no

### Processed Graph Data

Nodes: 62




Edges 159

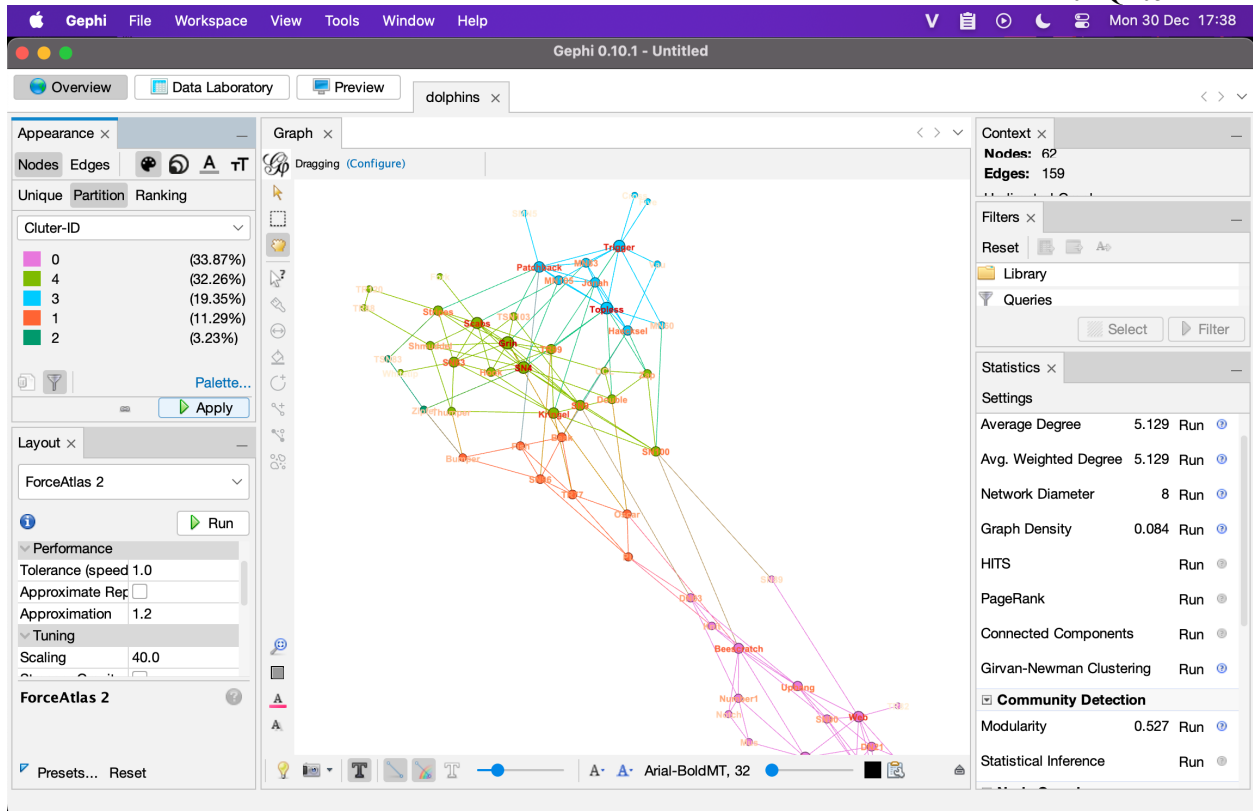
Processing time: 0.063 sec.

### Communities

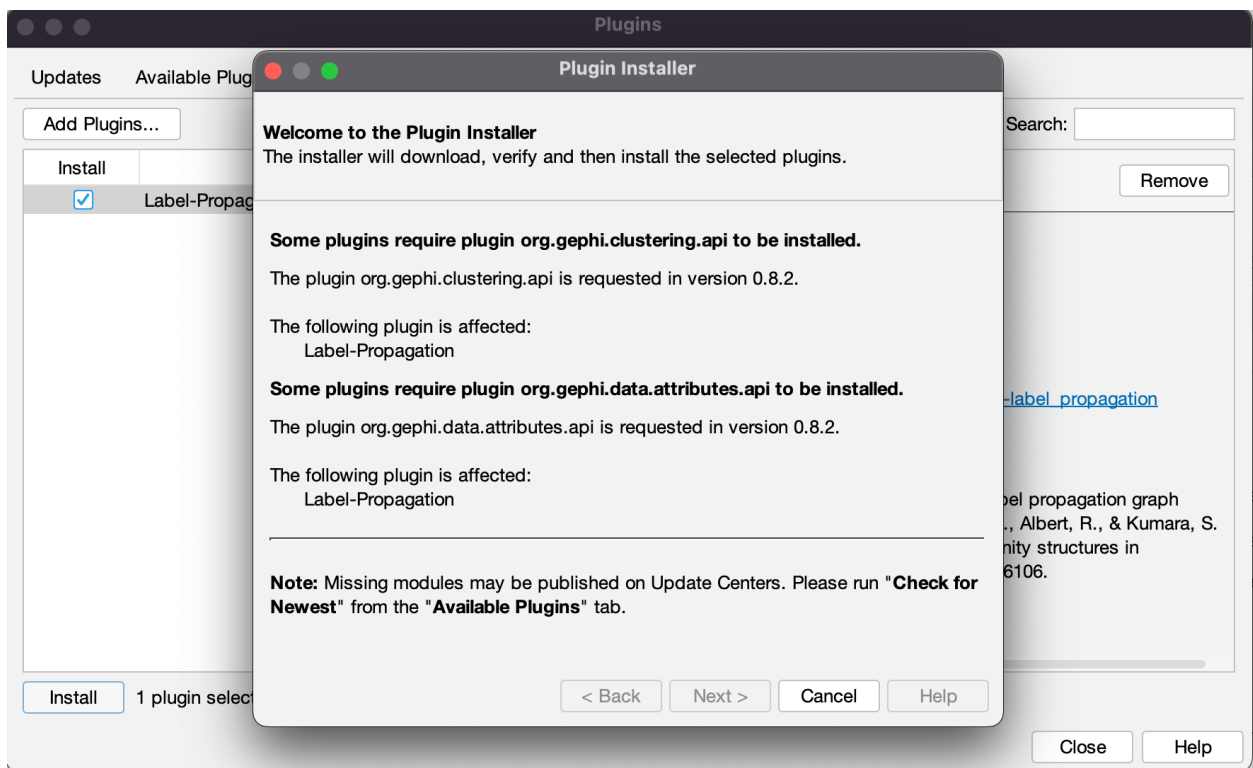
Number of communities: 5

Maximum found modularity: 0.5193822

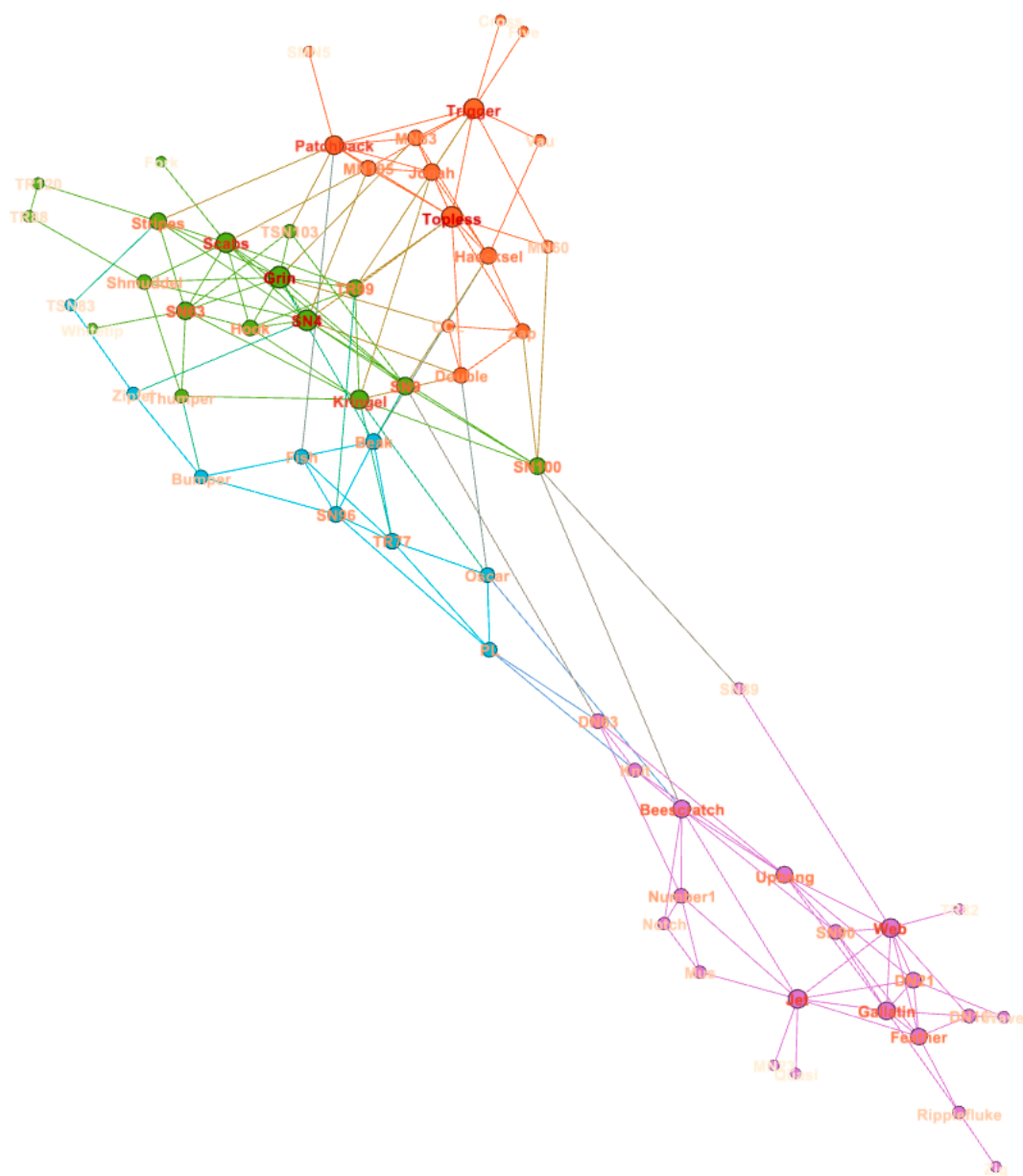
 Print  Copy  Save Close

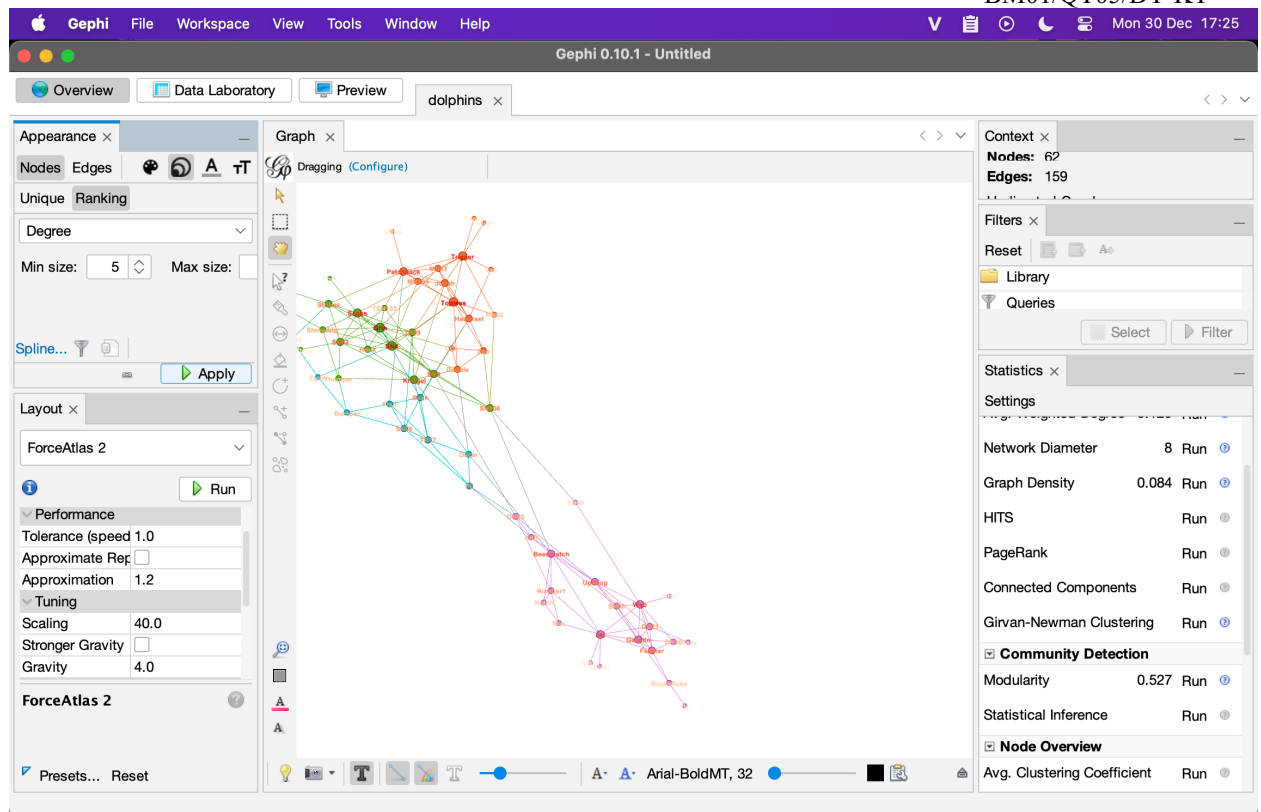


## 2.3 LPA



PHẦN 3: tốt nhất là Louvian





## PHẦN 4: So sánh

### 1. So sánh kết quả của thuật toán Louvain và Girvan-Newman

#### Thuật toán Louvain

- Lý thuyết toán học:**

Thuật toán Louvain tối ưu hóa Modularity (QQQ) theo từng cấp độ (hierarchical).  
Modularity được định nghĩa là:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j)$$

- $A_{ij}$ : Trọng số giữa hai đỉnh  $i$  và  $j$ .
- $k_i, k_j$ : Tổng trọng số các cạnh của  $i$  và  $j$ .
- $m$ : Tổng số cạnh trong mạng.
- $\delta(c_i, c_j)$ : Hàm Kronecker, bằng 1 nếu  $c_i = c_j$ , 0 nếu khác.

Louvain thực hiện tối ưu hóa Modularity qua hai giai đoạn:

- Tái phân phối các đỉnh vào các cụm để tăng QQQ.
  - Tập hợp các cụm thành siêu đỉnh và lặp lại.
- Ưu điểm:**



- Tối ưu hóa Modularity toàn cục, dẫn đến kết quả ổn định và rõ ràng.
- Rất nhanh và hiệu quả, đặc biệt cho mạng lớn.
- Modularity cao nhất trong các thuật toán, ví dụ:  $Q=0.527Q = 0.527Q=0.527$ .
- **Nhược điểm:**
  - Tạo ra các cụm lớn hơn, ít nhạy với các cộng đồng nhỏ.
  - Có thể bỏ qua các chi tiết cục bộ trong mạng.
- **Kết quả từ mạng cá heo:**
  - Số lượng cộng đồng: 444.
  - Modularity:  $Q=0.527Q = 0.527Q=0.527$ .

## Thuật toán Girvan-Newman

- **Lý thuyết toán học:**  
 Thuật toán Girvan-Newman dựa trên việc loại bỏ các cạnh có giá trị Betweenness Centrality cao nhất. Betweenness Centrality ( $CB(e)$ ) của một cạnh  $e$  được tính như:

$$CB(e) = \sum_{s \neq t} \sigma_{st}(e) \sigma_{st} C_B(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}} CB(e) = \sum_{s \neq t} \sigma_{st}(e)$$

- $\sigma_{st}$ : Số đường đi ngắn nhất giữa  $s$  và  $t$ .
- $\sigma_{st}(e)$ : Số đường đi ngắn nhất giữa  $s$  và  $t$  đi qua cạnh  $e$ .

Các cạnh có  $CB(e)$  cao được coi là "cầu nối" giữa các cộng đồng. Khi các cạnh này bị loại bỏ, mạng lưới phân tách dần thành các cụm rời rạc.

- **Ưu điểm:**
  - Phù hợp để phát hiện các cộng đồng nhỏ và rời rạc.
  - Mang tính chất giải thích rõ ràng về cách các cạnh đóng vai trò trung gian trong mạng.
  - Dựa trên lý thuyết đường đi ngắn nhất, giúp xác định cấu trúc mạng chính xác hơn.
- **Nhược điểm:**
  - Hiệu suất kém trên mạng lớn do phải tính toán Betweenness Centrality cho từng cạnh.
  - Modularity thấp hơn Louvain, ví dụ:  $Q=0.519382Q = 0.519382Q=0.519382$ .
- **Kết quả từ mạng cá heo:**
  - Số lượng cộng đồng: 555.
  - Modularity:  $Q=0.519382Q = 0.519382Q=0.519382$ .

## Lý thuyết của thuật toán LPA (Label Propagation Algorithm)

- **Lý thuyết toán học:**  
 LPA hoạt động dựa trên sự lan truyền nhãn giữa các đỉnh. Ban đầu, mỗi đỉnh được gán một nhãn duy nhất. Trong mỗi lần lặp:

$$L(v) = \arg \max_{\{l\}} \sum_{u \in \Gamma(v)} I(L(u)=l) L(v) = \arg \max_{\{l\}} \sum_{u \in \Gamma(v)} I(L(u)=l) L(v)$$

- $L(v)$ : Nhãn của đỉnh  $v$ .
- $\Gamma(v)$ : Tập các đỉnh lân cận của  $v$ .
- $I(L(u)=l)$ : Hàm chỉ thị, bằng 1 nếu  $L(u)=l$ , 0 nếu không.

Thuật toán hội tụ khi tất cả các đỉnh có nhãn ổn định.

- **Ưu điểm:**
  - Thời gian xử lý nhanh, lý tưởng cho mạng rất lớn.
  - Không cần tham số.
- **Nhược điểm:**
  - Kết quả không ổn định, phụ thuộc vào trạng thái khởi tạo.
  - Không tối ưu hóa Modularity trực tiếp.

## 2. Giải thích ý nghĩa của các cộng đồng được phát hiện trong ngữ cảnh của mạng xã hội

- **Cộng đồng trong mạng cá heo:**
  - Các cụm đại diện cho nhóm cá heo tương tác thường xuyên, như nhóm gia đình hoặc bạn bè thân thiết.
  - Cộng đồng lớn hơn (từ Louvain) có thể chỉ ra sự phân chia theo vai trò hoặc địa lý trong nhóm cá heo.
  - Các cộng đồng nhỏ hơn (từ Girvan-Newman) cho thấy các nhóm rời rạc với mối liên kết mạnh nội tại nhưng yếu ngoại vi.
- **Ứng dụng trong ngữ cảnh xã hội:**
  - Phát hiện các nhóm tương tự trong mạng xã hội người, giúp nhận diện nhóm sở thích hoặc mối quan tâm.
  - Tối ưu hóa thông điệp quảng cáo hoặc chiến dịch tiếp cận theo cộng đồng.

## 3. Đề xuất phương pháp phân cụm phù hợp nhất

**Phương pháp đề xuất: Louvain**

- **Lý do:**
  - Modularity cao hơn ( $Q=0.527$ ), đảm bảo cộng đồng được xác định rõ ràng.
  - Hiệu suất xử lý nhanh, phù hợp với quy mô mạng cỡ trung bình.
  - Độ ổn định cao trong các lần chạy, kết quả nhất quán.

**Khi nên dùng Girvan-Newman:**

- Nếu cần hiểu rõ cấu trúc cạnh và vai trò trung gian trong mạng.
- Thích hợp cho mạng nhỏ hoặc cần phân tích sâu các cạnh quan trọng.

Với mạng xã hội cá heo, Louvain là lựa chọn tối ưu để cân bằng hiệu suất và chất lượng phân cụm.