

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

MẠNG XÃ HỘI

*GVHD: Ths.Nguyễn Thị Kim Phụng*

1. Nguyễn Thị Ngọc Hà - 17520421
2. Nguyễn Thị Thu Phương - 17520928
3. Nguyễn Minh Thư - 17521105
4. Trần Hoài Thanh - 17521059

MỤC LỤC

[I. GIỚI THIỆU: 4](#_Toc57880664)

[1. Đề tài 4](#_Toc57880665)

[2. Bộ dữ liệu 4](#_Toc57880666)

[3. R Studio 6](#_Toc57880667)

[4. Gephi 7](#_Toc57880668)

[II. GEOMETRIC MEASURES (Độ đo hình học) 12](#_Toc57880669)

[1. Degree Centrality (Độ trung tâm dựa trên bậc của nút): 12](#_Toc57880670)

[1.1. Sử dụng R 12](#_Toc57880671)

[1.2. Dùng Gephi: 15](#_Toc57880672)

[2. Closeness centrality (Độ trung tâm dựa trên sự gần gũi) 18](#_Toc57880673)

[2.1. Sử dụng R 18](#_Toc57880674)

[2.2. Sử dụng Gephi 20](#_Toc57880675)

[III. SPECTRAL MEASURES (Độ đo phổ) 24](#_Toc57880676)

[1. Eigenvector-centrality 24](#_Toc57880677)

[1.1. Sử dụng R 24](#_Toc57880678)

[1.2. Sử dụng Gephi 25](#_Toc57880679)

[2. Page rank 28](#_Toc57880680)

[2.1. Sử dụng R 28](#_Toc57880681)

[2.2. Sử dụng Gephi 30](#_Toc57880682)

[IV. PATH-BASED MEASURES (Độ đo dựa trên đường đi) 33](#_Toc57880683)

[1. Betweenness centrality (Độ trung tâm ở giữa) 33](#_Toc57880684)

[1.1. Sử dụng R 33](#_Toc57880685)

[1.2. Sử dụng Gephi 34](#_Toc57880686)

[V. THUẬT TOÁN GOM CỤM 38](#_Toc57880687)

[1. Thuật toán Louvain: 38](#_Toc57880688)

[1.1. Sử dụng R 39](#_Toc57880689)

[1.2. Sử dụng Gephi 40](#_Toc57880690)

[2. 28Thuật toán Girvan-Newman 42](#_Toc57880691)

[2.1. Sử dụng R 42](#_Toc57880692)

[2.2. Sử dụng Gephi 43](#_Toc57880693)

[VI. BẢNG PHÂN CÔNG CÔNG VIỆC 47](#_Toc57880694)

[VII. TÀI LIỆU THAM KHẢO 47](#_Toc57880695)

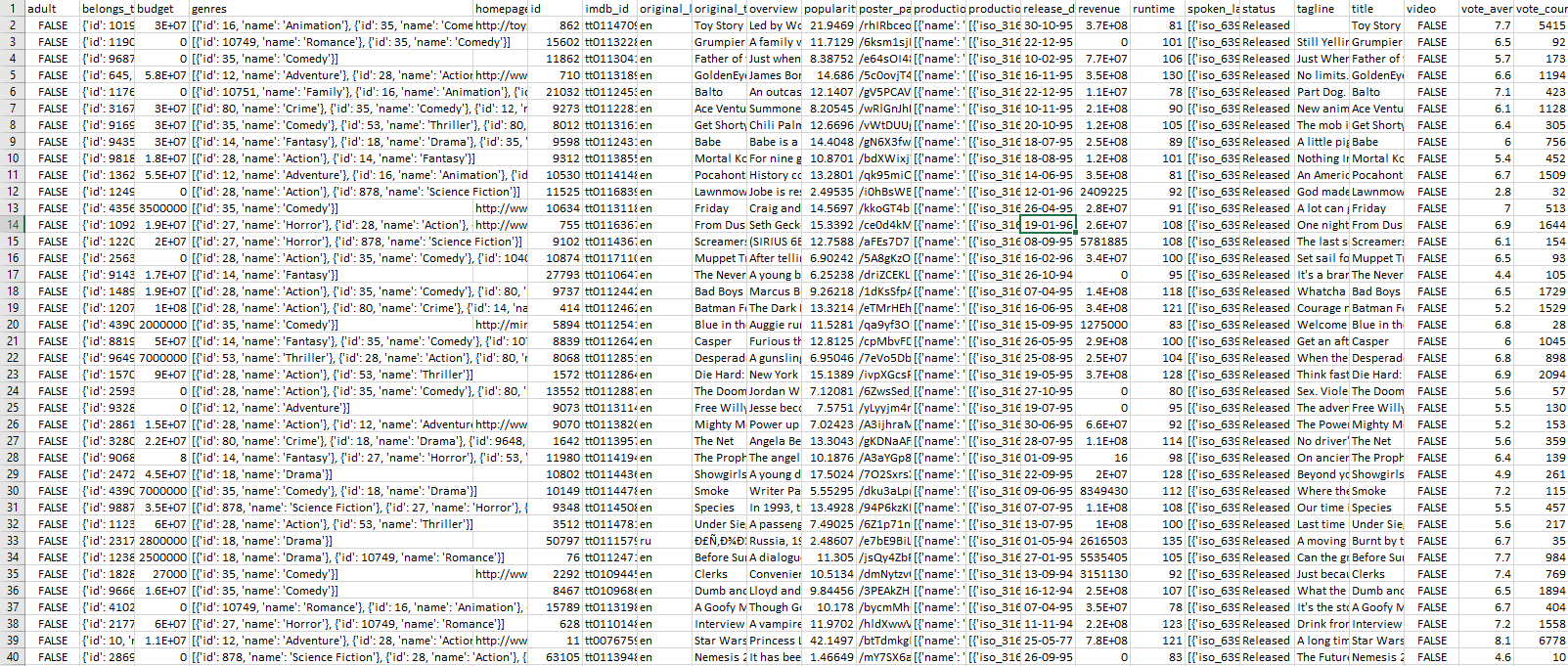
# GIỚI THIỆU:

## Đề tài

Đề tài chúng em hướng đến là phân tích mối liên hệ giữa các bộ phim thông qua thể loại (genres). Sử dụng các Độ đo (Độ đo trung tâm, Closeness, Betweeness), Thuật toán Phân tích phổ (Page Rank, Eigenvector), Path-Based (Node Betweeness), để biểu diễn các mối quan hệ, từ đó, tìm ra bộ phim có số mối liên hệ về thể loại nhiều nhất.

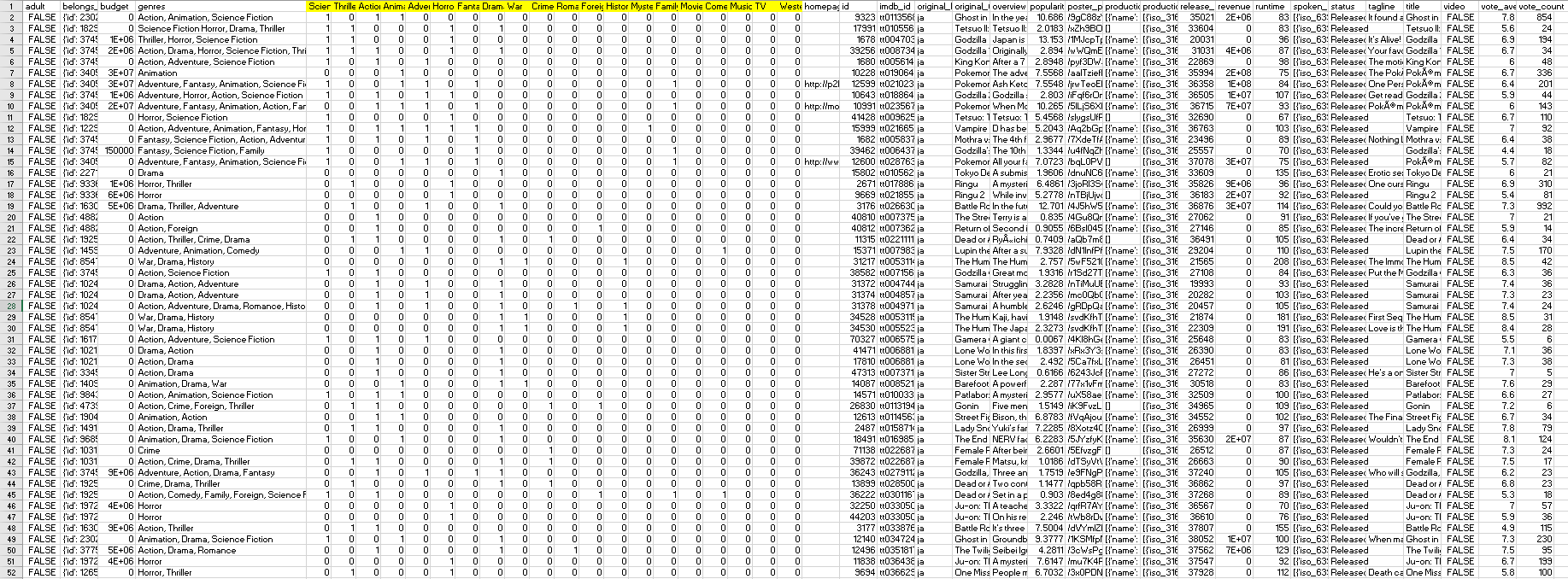
## Bộ dữ liệu

Bộ dữ liệu được lấy từ Kaggle: <https://www.kaggle.com/rounakbanik/the-movies-dataset> trong mục Data: “*movies\_metadata.csv*” gồm 4422 dòng.



Tiến hành lọc các thuộc tính lỗi font và NULL trong bộ dữ liệu.

Data chúng em lấy gồm dữ liệu trong số các bộ phim có nước sản xuất là: cn (Trung Quốc), ja (Nhật Bản), pt (Bồ Đào Nha), es (Tây Ban Nha) và fi (Phần Lan).



Lấy ngẫu nhiên 251 nodes từ bộ dữ liệu sau khi lọc, xác định trọng số, nhóm chúng em có tập kết quả sau:



Với các nút (Nodes) là các bộ phim có được từ bộ dữ liệu, cạnh (edges) là những liên kết giữa 2 bộ phim khi có cùng thể loại (genres). Chúng em sẽ xác định mối liên hệ về thể loại (genres) để biểu diễn đồ thị và kết quả trên ứng dụng Gephi và R Studio.

## R Studio

* Đọc dữ liệu:

edges=read.csv("E:/data.csv", header=TRUE)

* Review dữ liệu:

head(edges)

* Đọc thư viện:

library(igraph)

* Tạo graph:

net<-graph\_from\_data\_frame(data=edges, directed=F)

* Hiển thị graph:

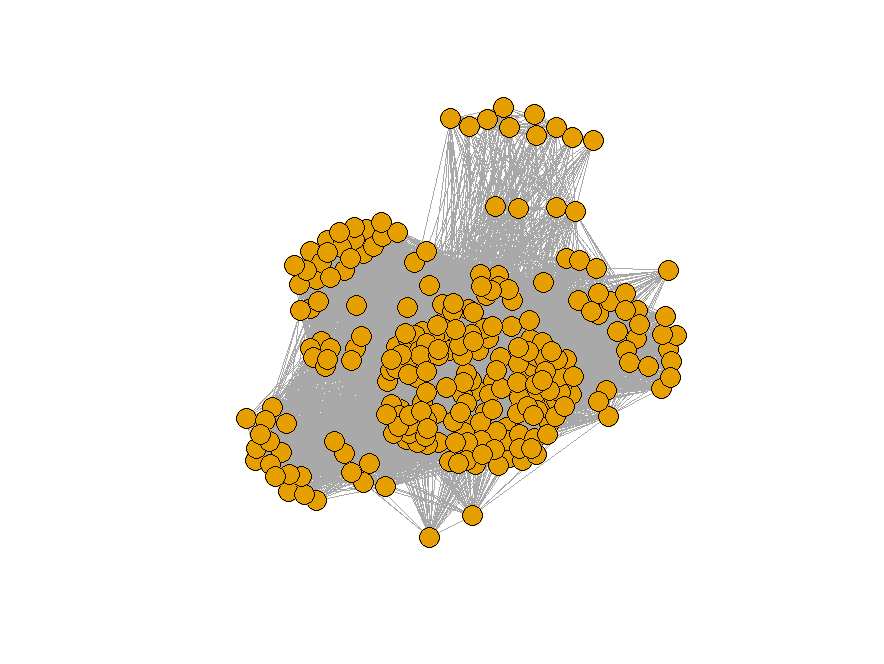
plot(net, vertex.size=4 ,vertex.label=NA)

Với vertex.size: size của node

vertex.label=NA: không hiển thị tên node trên graph

directed= F: đồ thị vô hướng

Hình ảnh Graph được hiển thị:



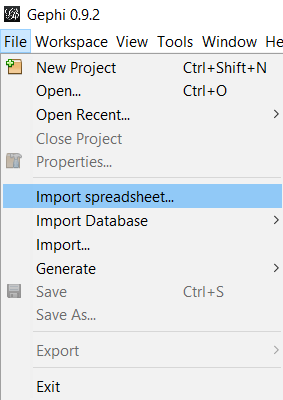
## Gephi

* Đọc tập dữ liệu:

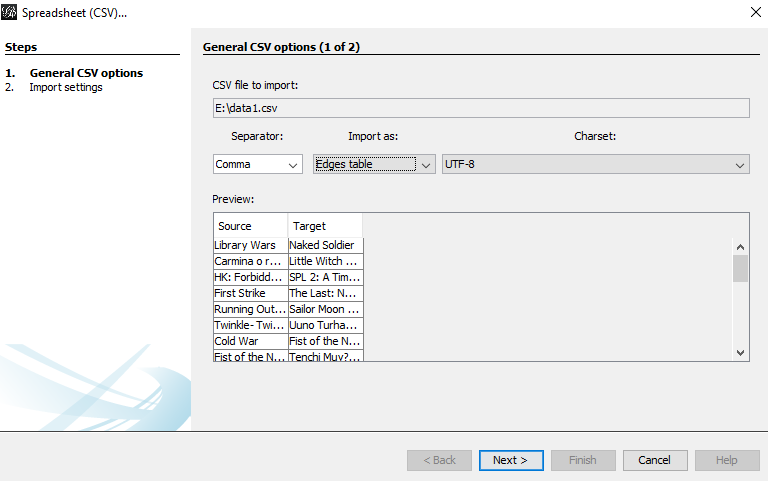
Tập dữ liệu gồm:

Tập data*.csv*: Bao gồm dữ liệu về số cạnh (edges), mối liên hệ giữa các node giữa chúng.

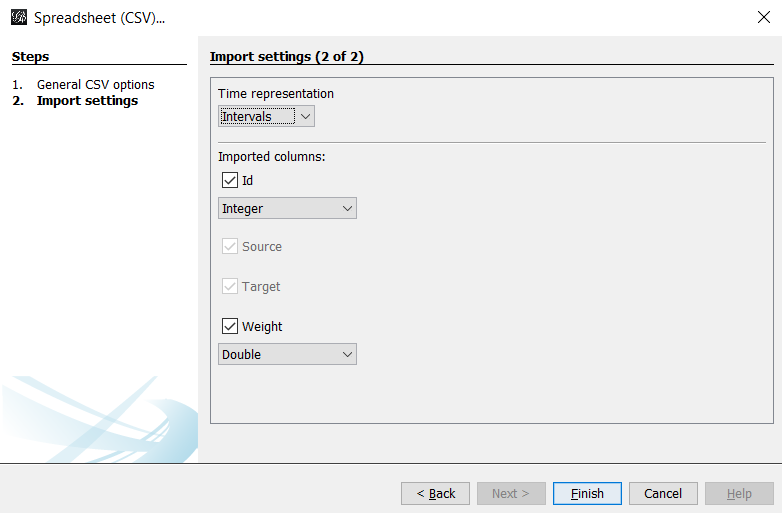
B1: Chọn *Import Spreadsheet,* import file dữ liệu



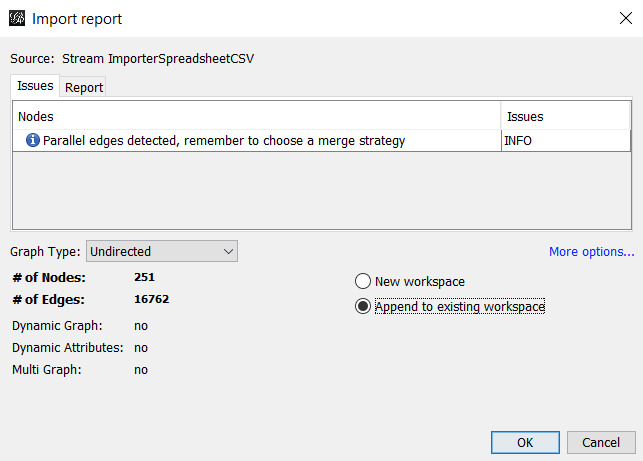
B2: Ở mục Import as, chọn *Edges table-> Next.*



B3: Chọn Interval -> Tích chọn ID và Weight



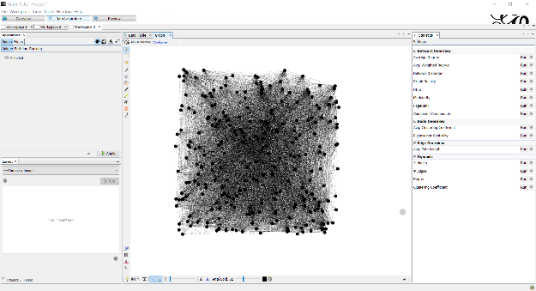
B4: Ở mục Graph type -> Chọn Undirected

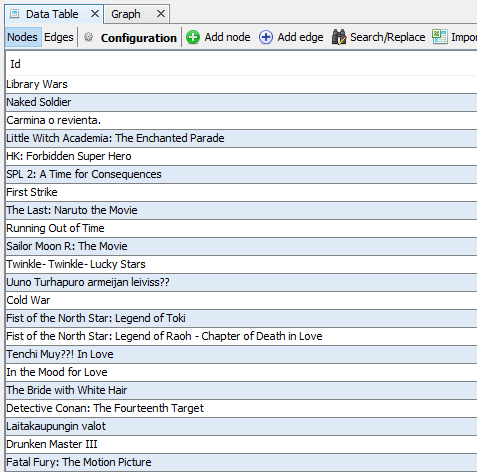


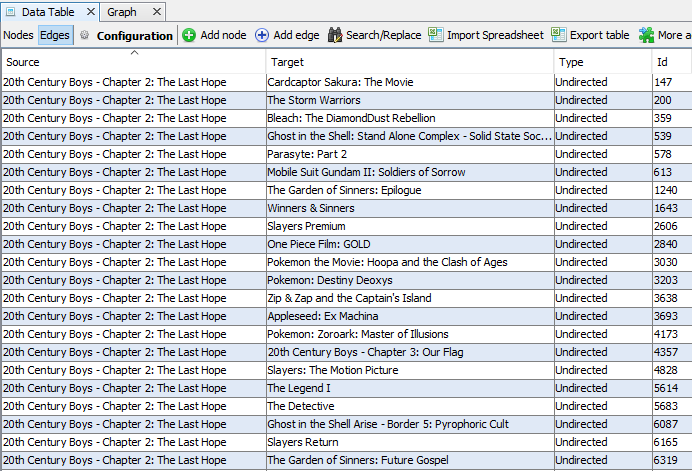
Đồ thị bao gồm 251 nodes, 16.762 cạnh

B5: Tích chọn Append to existing workspace -> OK

Hình ảnh Visualize của mạng đồ thị ta thu được như sau



Dữ liệu các Node được đọc vào bảng:

Dữ liệu các Edge được đọc vào bảng:

# GEOMETRIC MEASURES (Độ đo hình học)

## Degree Centrality (Độ trung tâm dựa trên bậc của nút):

Độ đo trung tâm: Là độ đo được định nghĩa bởi số lượng liên kết(bậc) của nút đó.

Gồm 2 thuộc tính: In Degree và Out Degree. Trong đó:

• In Degree: Số lượng các liên kết hướng đến nút đó.

• Out Degree: Số lượng liên kết mà nút đó hướng đến nút khác.

Ở đây, đồ thị của nhóm chúng em là đồ thị Undirected – Tức đồ thị không có hướng. Do đó giá trị của 2 thuộc tính này là như nhau.

Với thuộc tính Độ đo trung tâm, đồ thị được vẽ ra sẽ biểu thị mối liên kết giữa các nút (ở đây là phim) với các nút (phim) khác. Mối liên kết đó được tạo nên bởi các nút phim có cùng thể loại. Nút có giá trị Degree Centrality cao hơn sẽ có liên kết nhiều hơn trong mạng.

### Sử dụng R

Sử dụng câu lệnh trong R:

degree(net, v = V(net), mode = c( "total"))

Thuộc tính:

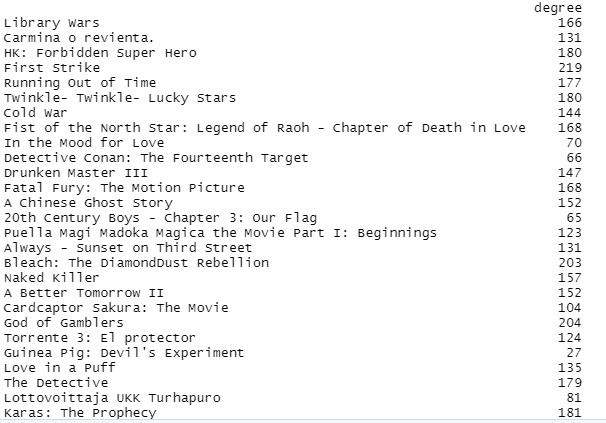
* Thuộc tính *net* được tạo từ câu lệnh sau:

*net <- graph\_from\_data\_frame(data=edges, directed=F)*

Trong đó: *graph\_from\_data\_frame* là hàm tạo đồ thị từ data frame là file csv đầu vào. *Data = edges* là chọn thuộc tính cạnh để đếm. *Directed = F* tức đây là đồ thị không có hướng.

* Thuộc tính *v = v(net)* tức chọn ra các đỉnh (nút) từ Đồ thị tạo bởi hàm net.
* Thuộc tính *mode = c(“total”)* tức chọn thước đo cho độ đo tương ứng. Thuộc tính này bao gồm các thước đo: In Degree, Out Degree và Freeman (Total). Chọn Total là thước đo mặc định do đây là đồ thị thuộc dạng không hướng.

Kết quả sau khi chạy dòng lệnh như sau:



Sau đó, để làm rõ hơn các nút theo giá trị độ đo trung tâm, chúng em chọn màu sắc nút đi từ đỏ đến xanh, ứng với dộ đo từ nhỏ đến lớn.

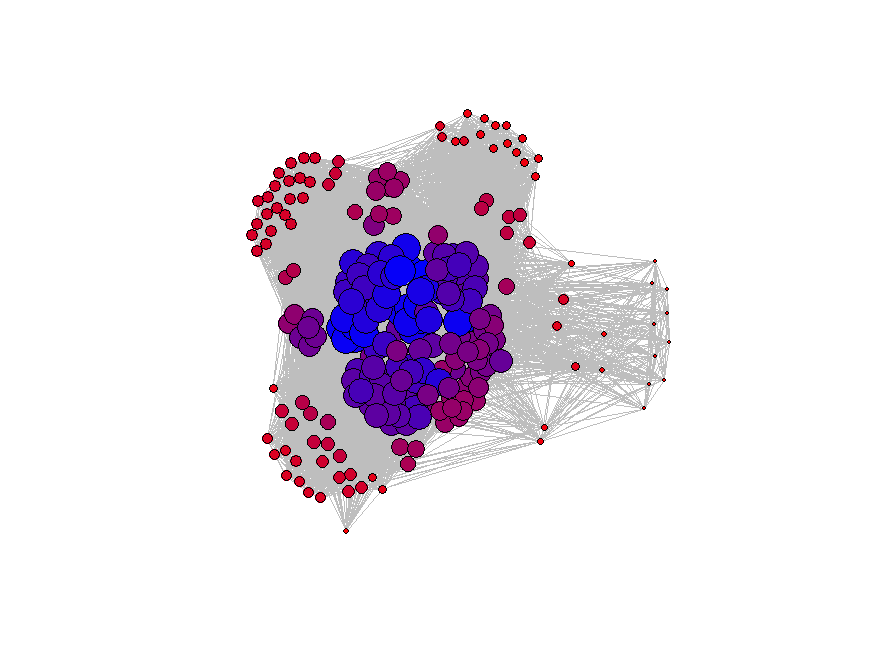
degree(net, v = V(net), mode = c( "total"))

cols=setNames(colorRampPalette(c("red","blue"))(length(unique(dg))), sort(unique(dg)))

plot(net,vertex.label=NA,vertex.color=cols[as.character(dg)],

vertex.size=dg\*0.07,edge.color="grey")

Hình ảnh sau khi chạy hàm có được:



Từ kết quả Độ đo trung tâm có được, nhóm chúng em chạy hàm liệt kê top 10 độ đo trung tâm lớn nhất như sau:

library(dplyr)

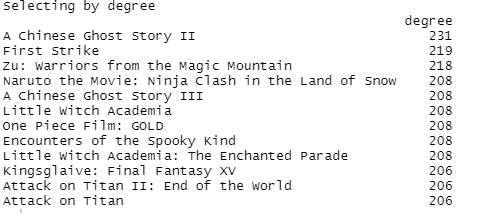
top\_n(top\_dg,10)%>% arrange(desc(degree))

Đầu tiên, install Plugin thư viện dplyr: Đây là một thư viện nằm trong hộp công cụ Tidyverse. Plugin này cung cấp những function hoán chuyển và thao tác trên dữ liệu sau khi nó đã được tải vào R.

Sử dụng hàm *top\_n* trong đó:

* top\_dg là hàm lấy giá trị chạy hàm degree bên trên, 10 là số lượng giá trị muốn lấy.
* Hàm arrange là hàm sắp xếp, desc là thứ tự từ lớn tới nhỏ.

Kết quả hiển thị như sau:



### Dùng Gephi:

Sau khi bước Import dữ liệu và Visualize, ta chọn Windown và mở hộp công cụ Statistic. Trong đó, có chức năng Avg.Degree. Ta nhấn Run chức năng này.

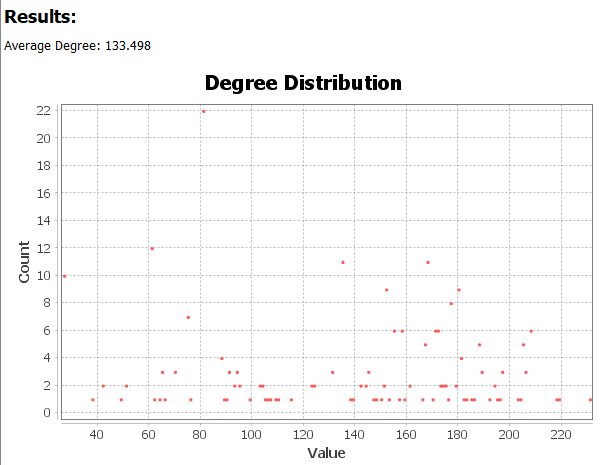
Kết quả Degree Centrality sau khi chạy:

Graphical user interface, application

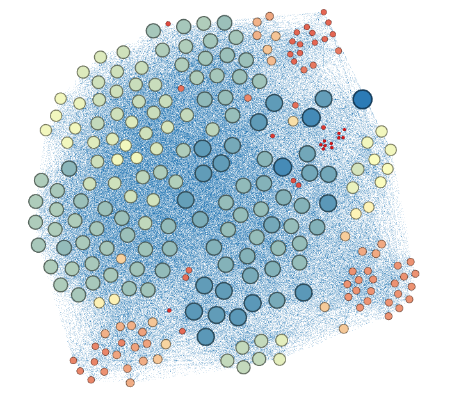
Description automatically generated

Ta có giá trị trung bình của độ đo Degree= 133.498

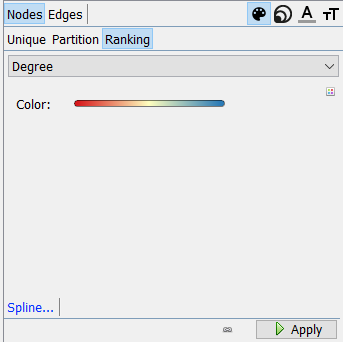
Biểu đồ phân phối mật độ của độ đo Degree Centrality được hiển thị như sau:



Với thuộc tính Value là giá trị Độ đo trung tâm tính được và thuộc tính Count là số lượng nút có giá trị Độ đo tương ứng.

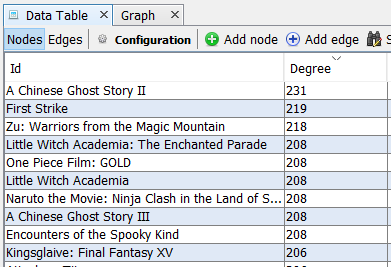
Biểu đồ mối liên hệ giữa các Node theo giá trị Degree Centrality:

Trong đó:



Màu sắc xếp loại các Node đi từ đỏ đến xanh – tương đương với giá trị từ nhỏ đến lớn.

* Top 10 phim xếp theo giá trị Degree Centrality thu được là:



## Closeness centrality (Độ trung tâm dựa trên sự gần gũi)

Là độ đo dựa trên sự gần gũi, chỉ ra một nút trong mạng có thể truy cập nhanh tới nhiều nút khác trong mạng.

Ở đây, Closeness centrality cho biết phim có thể loại gần với các phim khác nhất, là phim có mối liên kết cho phép nó truy cập đến tất cả các node khác trong mạng nhanh hơn bất kỳ nút (phim) nào khác.

### Sử dụng R

Sử dụng ngôn ngữ R

closeness(net, vids = V(net), normalized = TRUE)

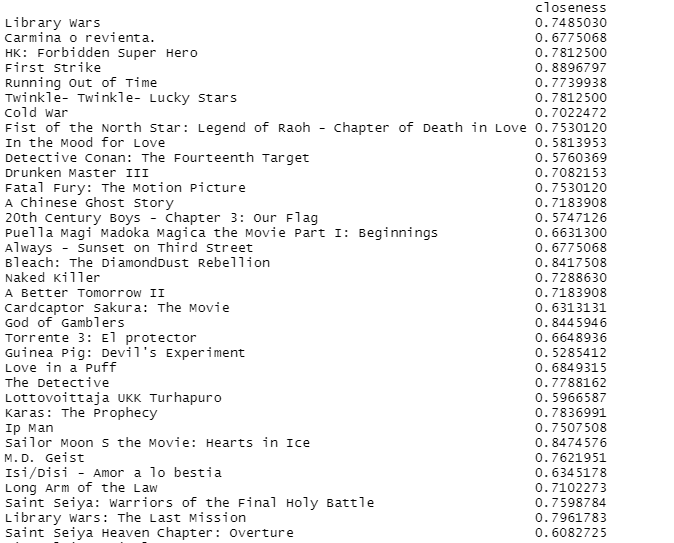
Trong đó:

• Vids = V(net) tức chọn ra các đỉnh (nút) từ Đồ thị tạo bởi hàm net.

• Normalized = TRUE tức thực hiện chuẩn hóa hay không.

Gía trị chạy hàm có được:

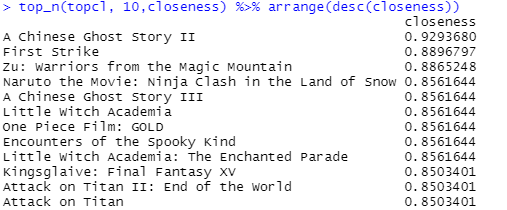
Ta có kết quả như sau:



Sau đó, ta chạy hàm top\_n để liệt kê top 10 giá trị Closeness có được:

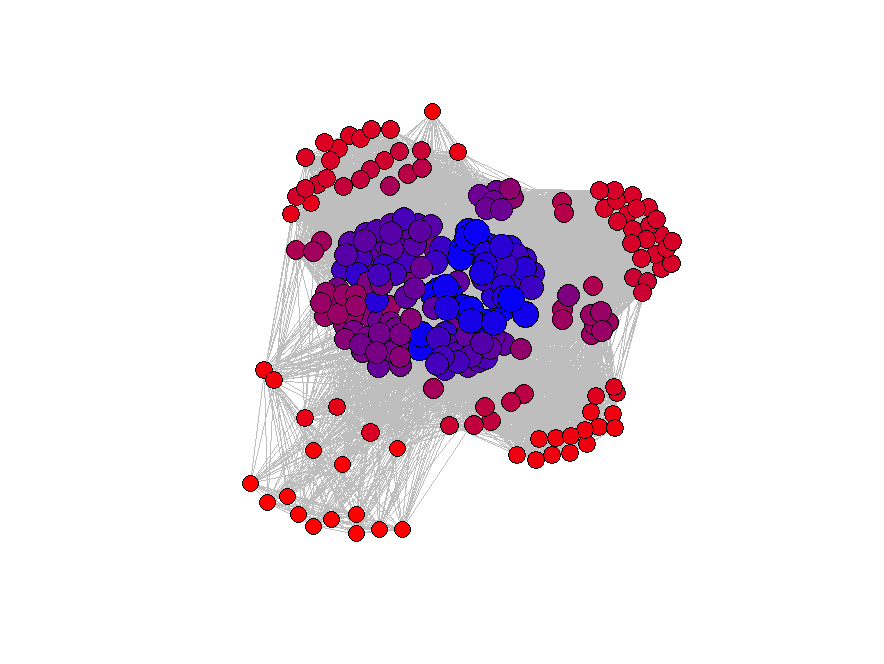
* Top 10 bộ phim xếp theo độ đo Closeness:

top\_n(topcl, 10,closeness) %>% arrange(desc(closeness))



Sắp xếp này cho thấy tựa(nút) phim: A Chinese Ghost Story II có giá trị độ đo dựa trên sự gần gũi và có mức ảnh hưởng tới các nút khác trong mạng cao nhất.

Biểu đồ sau khi chạy hàm closeness được vẽ như sau, với màu sắc từ đỏ tới xanh ứng với giá trị độ đo Closeness từ nhỏ đến lớn.



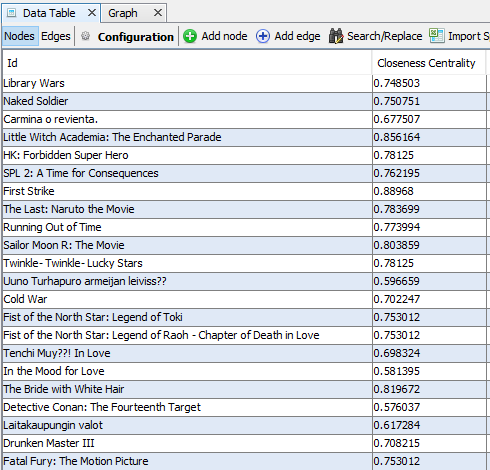
### Sử dụng Gephi

Ta chọn Statistic -> Network Diameter -> Run.

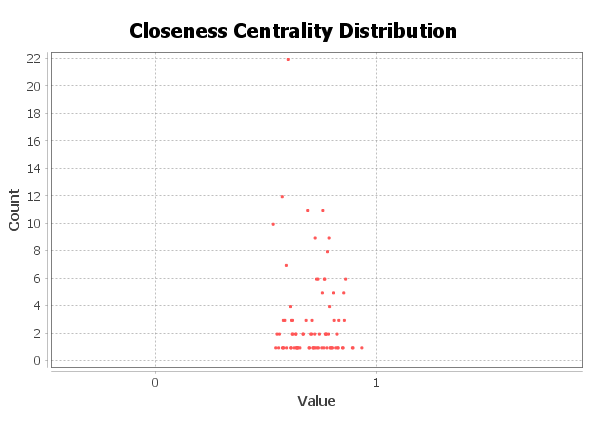
Graphical user interface, text, application, email

Description automatically generated

Kết quả Closeness Centrality sau khi chạy

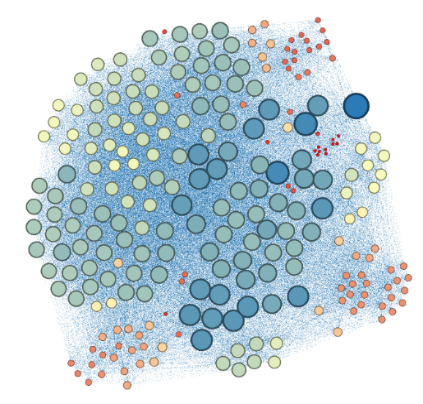


Biểu đồ phân phối mật độ của Closeness Centrality:

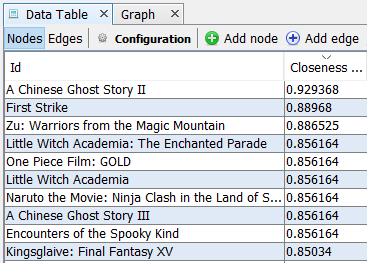


Trong đó, Thuộc tính Value là giá trị Độ đo Closeness của các nút thu được và Thuộc tính Count là số lượng nút có giá trị Closeness tương ứng.

Biểu đồ mối liên hệ giữa các nút theo giá trị Closeness Centrality



* Top 10 bộ phim xếp theo giá trị Closeness Centrality thu được là:



# SPECTRAL MEASURES (Độ đo phổ)

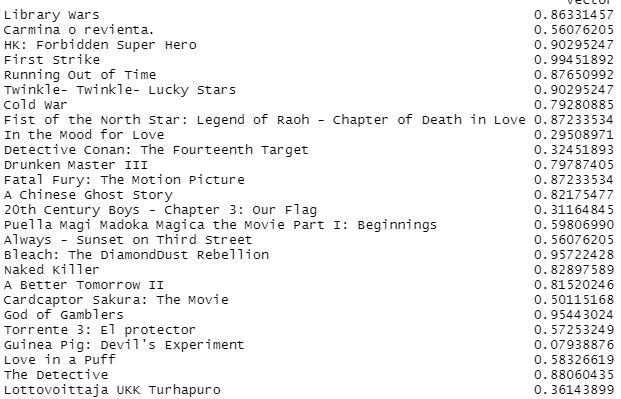
## Eigenvector-centrality

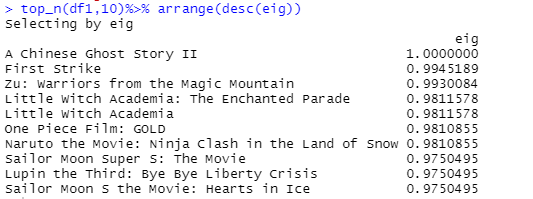
Độ đo Eigenvector (còn gọi là độ trung tâm hoặc điểm uy tín) là thước đo mức độ ảnh hưởng của một nút trong mạng. Nó chỉ định điểm số tương đối cho tất cả các nút trong mạng dựa trên khái niệm rằng các kết nối đến các nút có điểm cao đóng góp nhiều hơn vào điểm của nút được đề cập hơn là các kết nối bằng nhau đến các nút có điểm thấp.

Với 2 node có cùng độ đo trung tâm (Degree Centrality) nếu một nút được kết nối với nhiều nút quan trọng, nó cũng sẽ là một nút quan trọng, và nếu nó chỉ được kết nối với một vài nút không quan trọng thì nó sẽ không quan trọng. Nếu A và B có cùng mức độ trung tâm, nhưng A gắn với tất cả những người có trình độ cao và B gắn với tất cả những người có trình độ thấp, thì bằng trực giác chúng ta muốn thấy A có điểm cao hơn B.

### Sử dụng R

eigen\_centrality(net)$vector

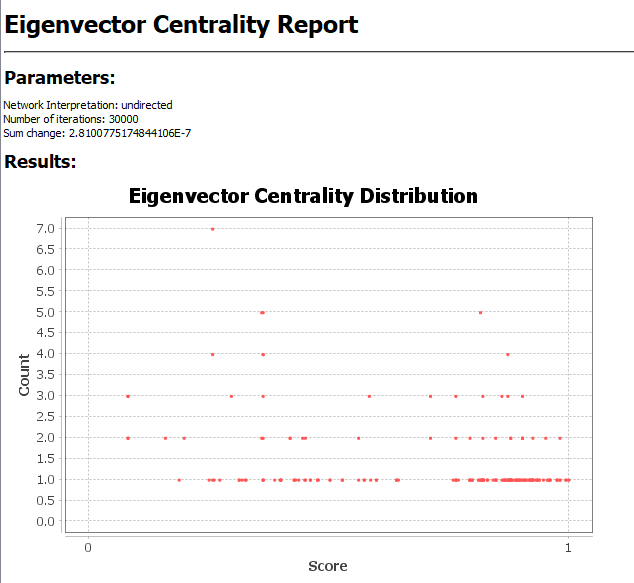


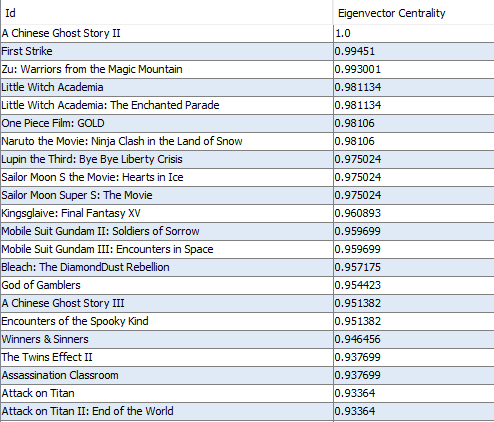
* Top 10 bộ phim xếp theo Eigienvector: là những bộ phim có liên kết đến nhiều nút quan trọng nhất.

### Sử dụng Gephi

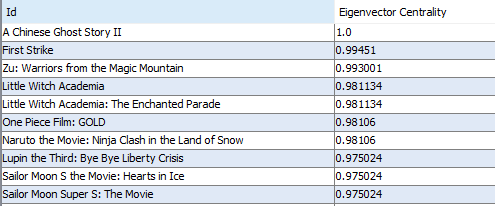
Vào thanh Statistics-> Eigenvector Centrality-> RunGraphical user interface, text, application

Description automatically generated

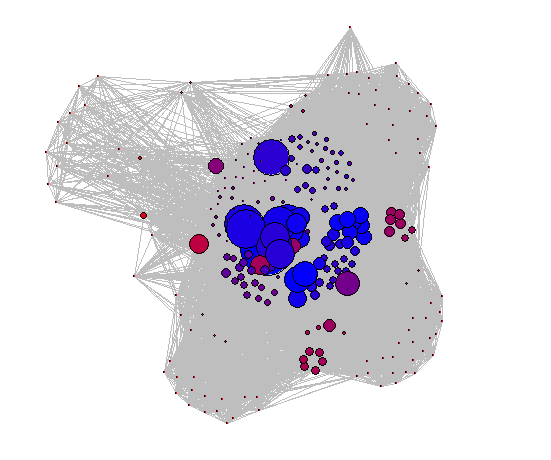
Ta được bảng phân phối và kết quả: 



* Top 10:



Biểu đồ biểu diễn node theo độ đo Eigenvector:



## Page rank

PageRank là một thuật toán đo lường ảnh hưởng bắc cầu hoặc kết nối của các nút.

Nó có thể được tính toán bằng cách phân phối lặp đi lặp lại thứ hạng của một nút (ban đầu dựa trên mức độ) cho các nút lân cận hoặc bằng cách duyệt ngẫu nhiên biểu đồ và đếm tần suất chạm vào mỗi nút trong những bước này.

Thuật toán áp dụng cho toàn bộ dữ liệu cần được truy xuất, là truy vấn độc lập.

### Sử dụng R

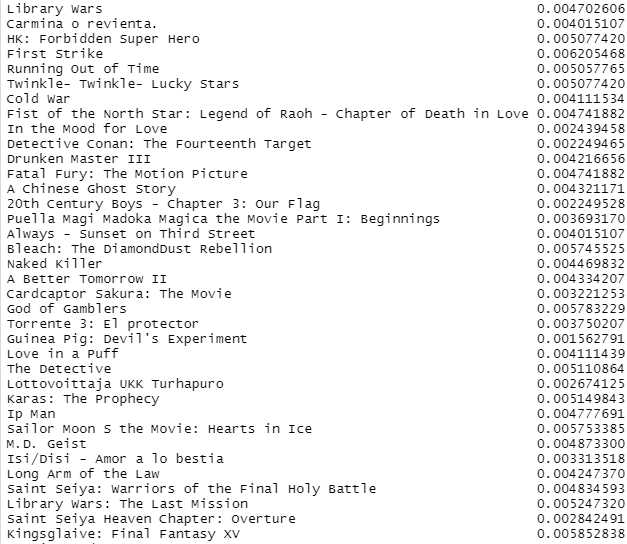
Sử dụng hàm page\_rank( graph, vids = V(graph), directed , damping) với 4 thông số:

|  |  |
| --- | --- |
| graph | Đồ thị |
| vids | Node của đồ thị |
| directed | Loại đồ thị, directed=TRUE khi đồ thị có hướng và directed =FALSE khi đồ thị vô hướng |
| damping | Hệ số giảm chấn |

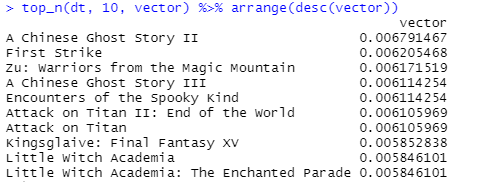
Áp dụng vào dữ liệu:

page.rank(net, vids=V(net), directed=FALSE, damping=0.85)

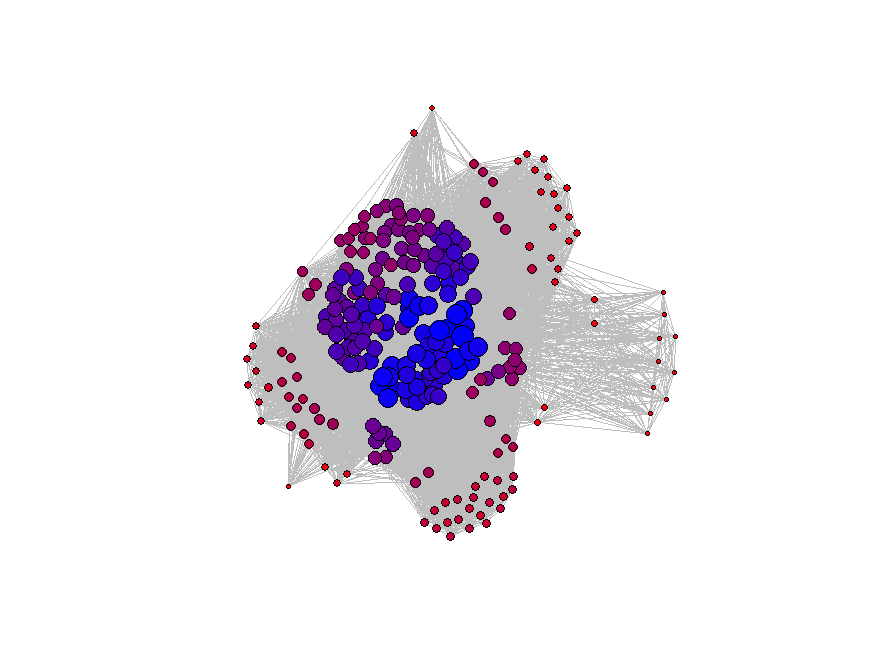
Kết quả:



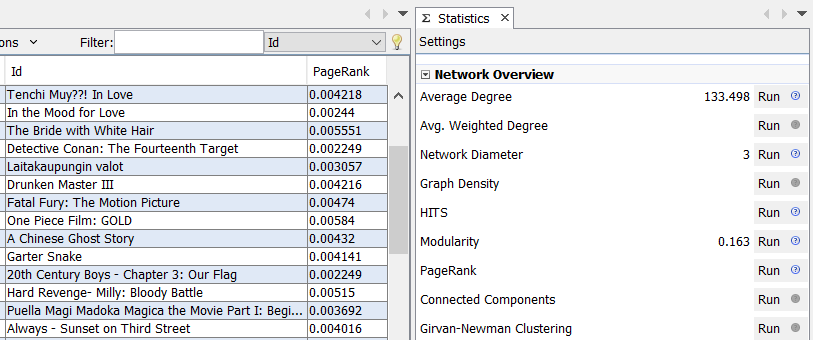
* Top 10 phim xếp loại theo Page Rank:



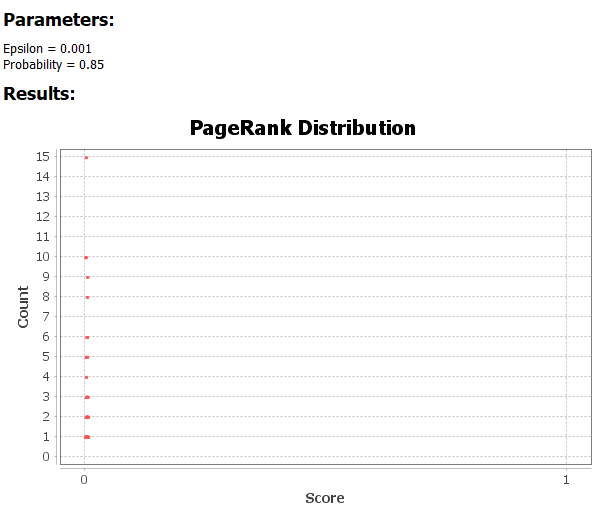
Biểu đồ:

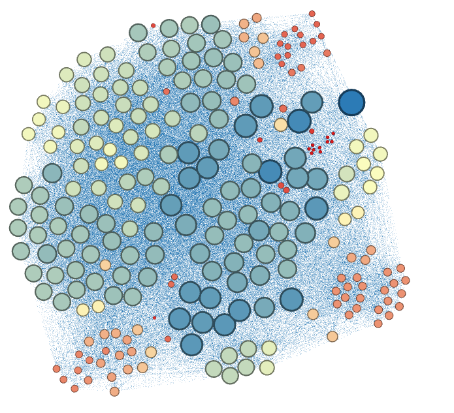


### Sử dụng Gephi

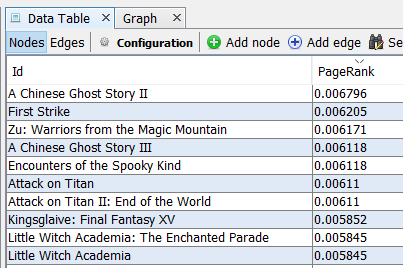
Kết quả Page Rank sau khi chạy:

Biểu đồ phân phối mật độ của độ đo Page Rank được hiển thị như sau:



Biểu đồ mối liên hệ giữa các Node theo giá trị Page Rank:

* Top 10 bộ phim xếp theo giá trị Page Rank thu được là:



# PATH-BASED MEASURES (Độ đo dựa trên đường đi)

## Betweenness centrality (Độ trung tâm ở giữa)

Là độ đo chỉ ra vị trí các nút trong mạng và khả năng kết nối các thành phần hoặc nhóm trong mạng.

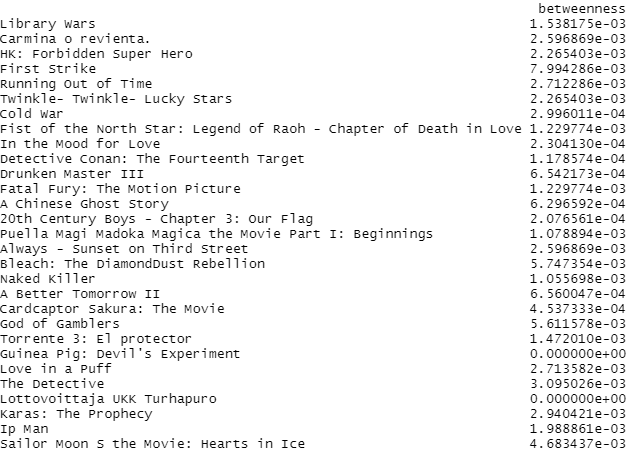
Ở đây Betweeness centrality cho biết phim có vị trí trung tâm trong mạng, nắm vai trò môi giới trong mạng, có tính trung gian cao, ảnh hưởng cao nhất đến sự kết nối của mạng.

### Sử dụng R

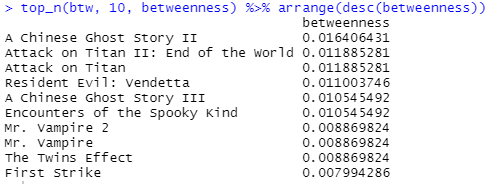
Sử dụng R Studio

betweenness(net, v = V(net), directed = FALSE, normalized = TRUE)

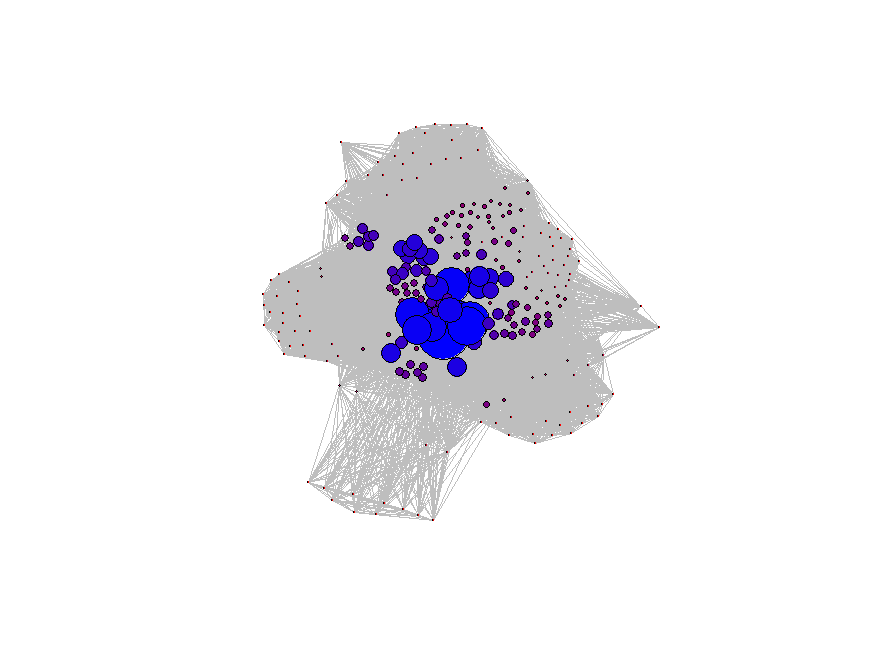
ta được kết quả như sau:



* Top 10 bộ phim xếp theo độ đo Betweenness:



Biểu đồ:

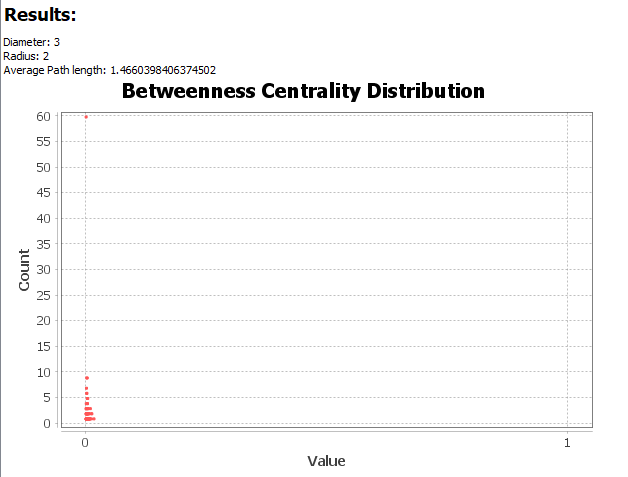


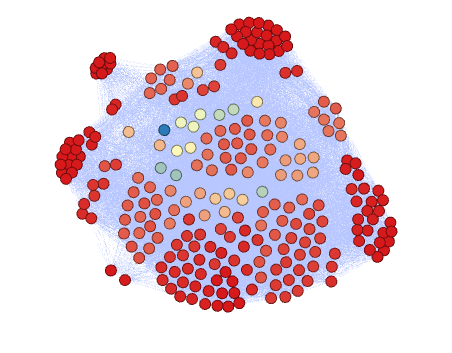
### Sử dụng Gephi

Ta chọn Statistic -> Network Diameter -> Run.

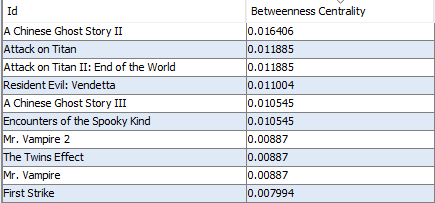
Kết quả Betweeness Centrality sau khi chạy:



Biểu đồ phân phối mật độ của độ đo Betweeness Centrality được hiển thị như sau:

Biểu đồ mối liên hệ giữa các Node theo giá trị Betweeness Centrality:

* Top 10 phim xếp theo giá trị Betweenness Centrality thu được là:



# THUẬT TOÁN GOM CỤM

## Thuật toán Louvain:

Phương pháp Louvain để phát hiện cộng đồng là một phương pháp trích xuất các cộng đồng từ các mạng lớn do Blondel và cộng sự tạo ra từ Đại học Louvain

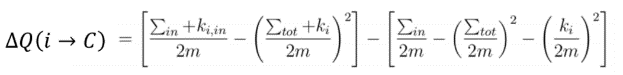
Nguồn cảm hứng cho phương pháp phát hiện cộng đồng này là việc tối ưu hóa giá trị *Modularity* khi thuật toán tiến triển. *Modularity* là một giá trị thang đo giữa [−0,5; 1] đo mật độ tương đối của các cạnh bên trong cộng đồng so với các cạnh bên ngoài cộng đồng.

Trong Phương pháp phát hiện cộng đồng của Louvain, các cộng đồng nhỏ đầu tiên được tìm thấy bằng cách tối ưu hóa cục bộ *Modularity* trên tất cả các node, sau đó mỗi cộng đồng nhỏ được nhóm thành một node và bước đầu tiên được lặp lại. Phương pháp này tương tự như phương pháp trước đó của Clauset, Newman và Moore nhằm kết nối các cộng đồng mà sự kết hợp của chúng tạo ra sự gia tăng *Modularity* lớn nhất.

Louvain sẽ sắp xếp ngẫu nhiên tất cả các nút trong mạng khi tối ưu hóa modul. Sau đó, từng nút một, nó sẽ xóa và chèn từng nút vào một cộng đồng khác nhau 𝐶 cho đến khi không có sự gia tăng đáng kể về modul (tham số đầu vào) được xác minh.

Quy trình:

* + Chọn một nút tùy ý, n trong 𝐶
  + Chọn lân cận của n mà hàm modul cực đại.
  + Lặp lại 1-2 cho đến khi tất cả các nút đã được nhóm lại thành một hàng xóm
  + Bây giờ tạo một đồ thị mới, nơi mỗi cộng đồng hợp nhất thành một nút duy nhất.
  + Lặp lại các bước 1-4 trên đồ thị mới cho đến khi modul không tăng.

Độ đo Modularity

Trong đó:

* + 𝛴𝑖𝑛 là tổng trọng số của các liên kết bên trong 𝐶
  + 𝛴𝑡𝑜𝑡 tổng trọng số của tất cả các liên kết tới các nút trong 𝐶
  + 𝑘𝑖 tổng trọng số của tất cả các liên kết trong nút 𝑖
  + 𝑘𝑖, 𝑖𝑛 tổng các trọng số của các liên kết từ nút 𝑖 đến các nút trong cộng đồng 𝐶
  + 𝑚 là tổng trọng số của tất cả các cạnh trong đồ thị.

### Sử dụng R

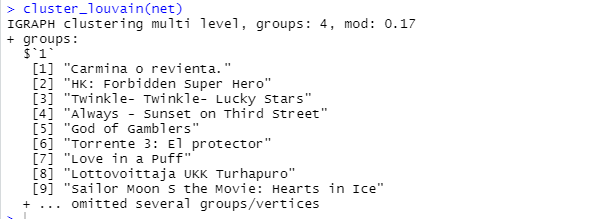
Dùng lệnh cluster\_louvain():

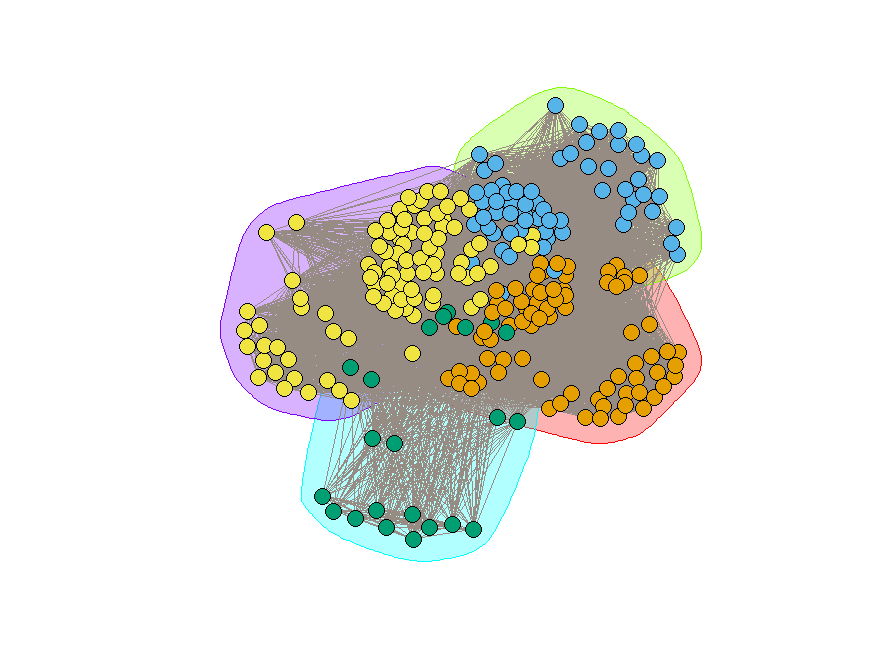
lc<- cluster\_louvain(net)

Vẽ graph bằng lệnh plot():

Plot(lc,net)

Kết quả:





Thuật toán Louvain chia cộng đồng phim thành 4 cụm với giá trị modularity=0.17

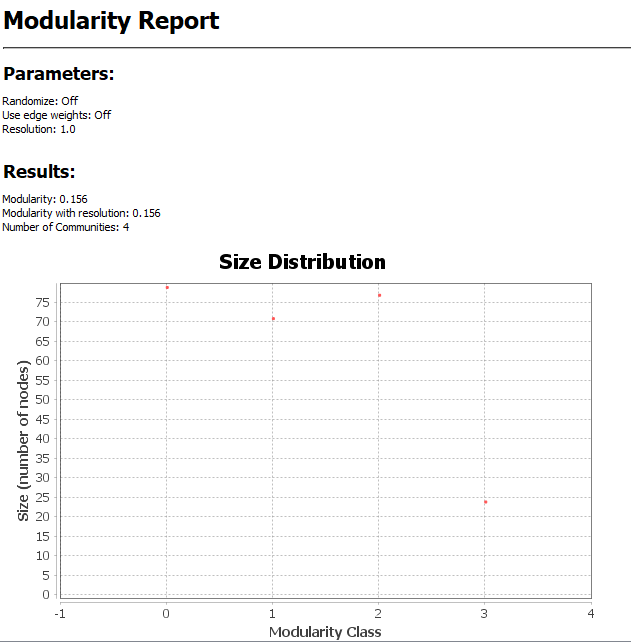
### Sử dụng Gephi

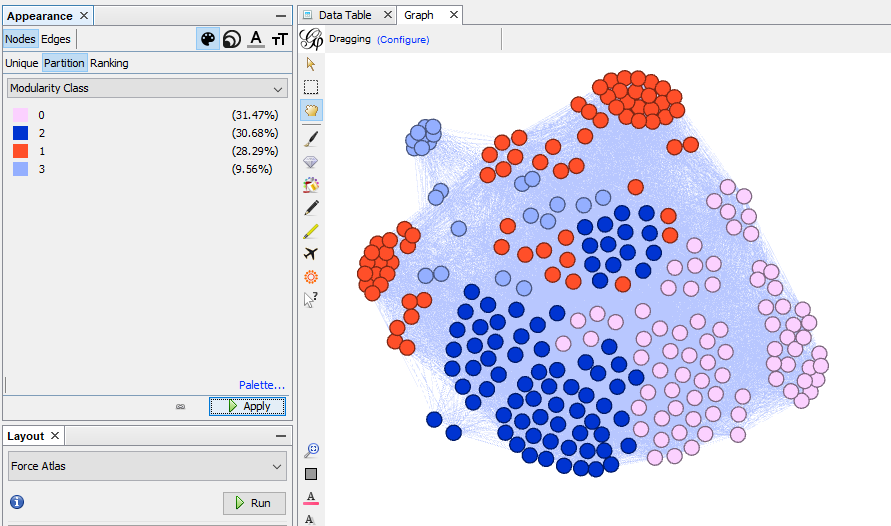
Trong hộp thoại Satistics -> Modularity -> Run

Graphical user interface, text, application, email

Description automatically generated

Bảng phân phối số cụm:



Kết quả gom cụm với tỉ lệ tương ứng 31.47%- 30.68%- 28.29%- 9.56%

## Thuật toán Girvan-Newman

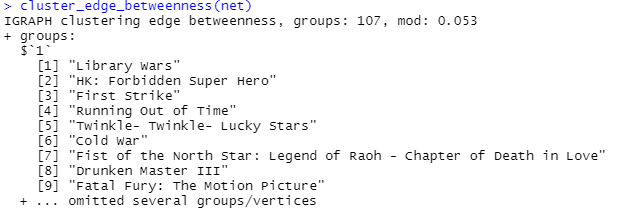
Girvan và Newman đề xuất một thuật toán phân cụm phân chia bao gồm việc lặp để xóa bỏ các cạnh từ đồ thị sử dụng độ đo trung trung gian. Điểm mấu chốt khác trong thuật toán này đó là việc tính toán lại độ đo trung gian cho tất cả các cạnh còn lại sau khi đã loại bỏ một cạnh. Bước lặp dừng lại khi độ đo chất lượng phân cụm "đơn thể" đạt giá trị tối ưu.

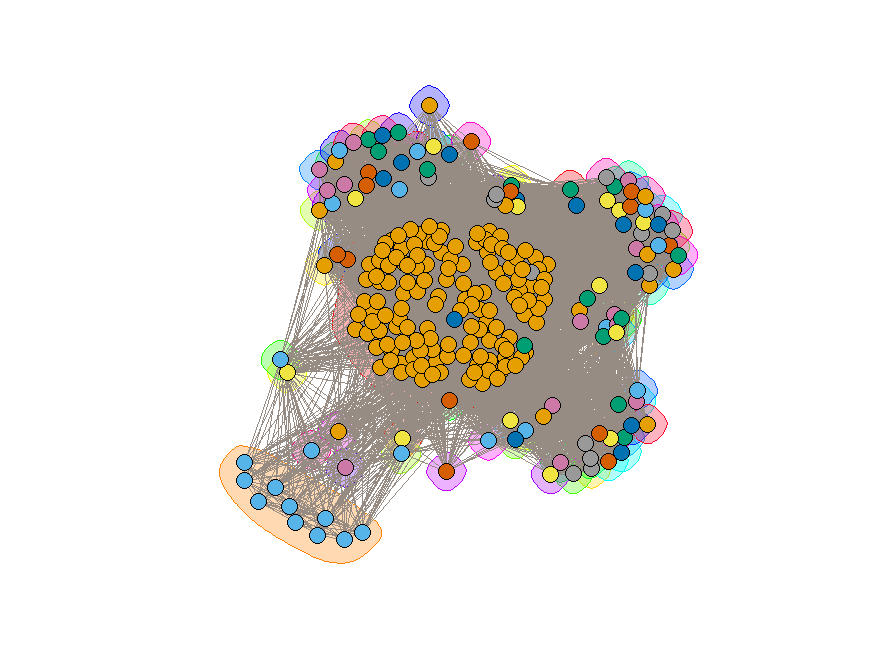
Quy trình:

* Tính khoảng giữa của tất cả các cạnh hiện có trong mạng.
* Cạnh có edge betweenness cao nhất bị loại bỏ.
* Tính toán lại giá trị edges betweenness.
* Lặp lại các bước 2 và 3 cho đến khi không còn cạnh

### Sử dụng R

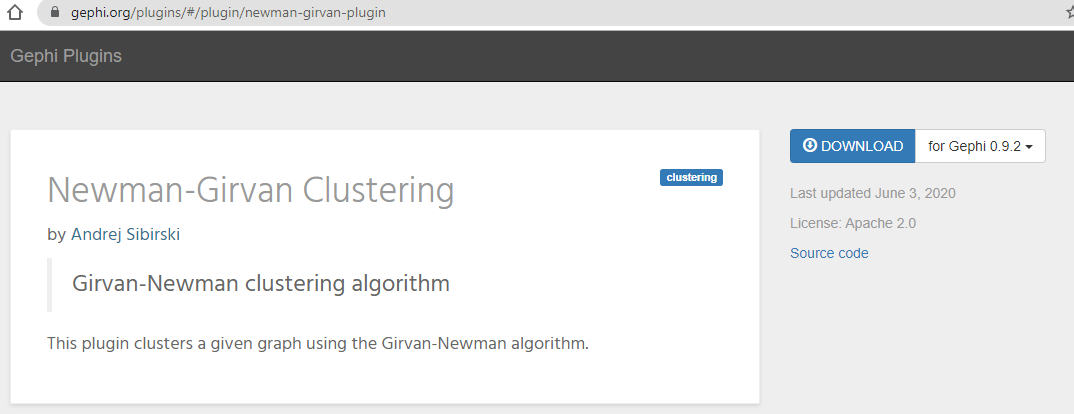
Thuật toán chia cộng đồng thành 107 cụm với độ chính xác cụm mod=0.053





### Sử dụng Gephi

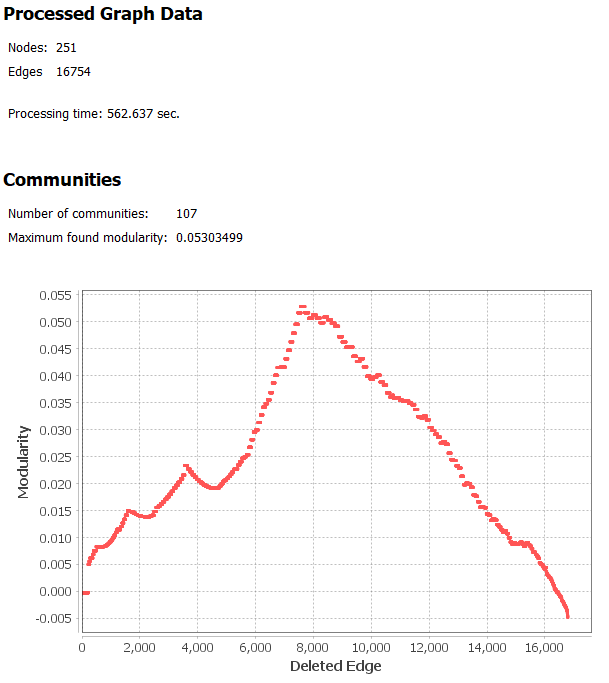
Download plugins Newman-Girvan Clustering từ website gephi.org

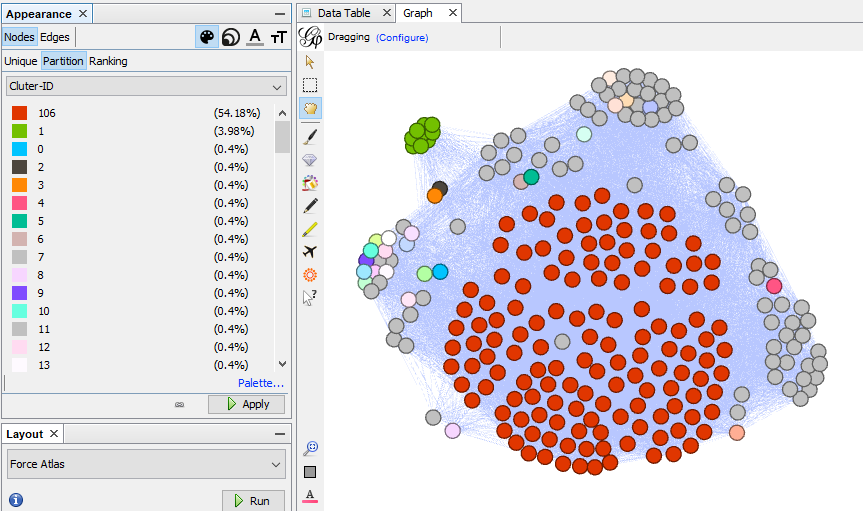


Sau khi cài đặt, hộp thoại *Stastics* xuất hiện thuật toán Girvan-Newman, nhấn chọn Run

Graphical user interface, text, application, email

Description automatically generated

Kết quả: 



# BẢNG PHÂN CÔNG CÔNG VIỆC

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Họ tên | Công việc | Chấm điểm thành viên |
| 1 | Nguyễn Minh Thư  **Sđt:** 0796803207 | - Phân tách dữ liệu  - Betweeness centrality | 10 |
| 2 | Nguyễn Thị Ngọc Hà | - Degree centrality  - Closeness centrality | 9.5 |
| 3 | Nguyễn Thị Thu Phương | - Eigenvector centrality  - Page rank | 10 |
| 4 | Trần Hoài Thanh | - Thuật toán gom cụm Louvain  - Thuật toán gom cụm Girvan Newman | 9.5 |

# TÀI LIỆU THAM KHẢO

* Slide Stanford [link](https://drive.google.com/drive/folders/1C2nIZsyVODzIewcqtkleZ1BLt9g2vXDk?usp=sharing)
* <https://www.r-bloggers.com/2018/12/network-centrality-in-r-an-introduction/>
* <https://www.datacamp.com/community/tutorials/centrality-network-analysis-R>
* <http://hoithaokhcn.tlu.edu.vn/Portals/7/2014/NewFiles/005.pdf>