**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**



**ĐỒ ÁN CUỐI KỲ**

**MÔN MẠNG XÃ HỘI**

**Giảng viên hướng dẫn:**  Ths Nguyễn Thị Kim Phụng

**Sinh viên thực hiện :**  Lê Văn Long -19521783

TP.HCM Tháng 11 Năm 2022

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

* 1. Giới thiệu

Từ xưa đến nay, nhu cầu đi lại của con người luôn là một yếu tố thiết yếu trong cuộc sống, các phương tiện đi lại ngày càng được phát triển và chi phí đi lại cũng giảm đi đáng kể qua từng năm.

Sau 2 năm cả thế giới hứng chịu đại dịch Covid 19, các đường giao thông quốc tế được mở trở lại. Để di chuyển vòng quanh thế giới, chúng ta có thể lựa chọn nhiều hãng hàng không khác nhau, các hãng hàng không cũng đưa ra nhiều ưu đãi giảm giá để cạnh tranh với nhau nhằm đem lại lợi nhuận cho riêng mình.

Môn học mạng xã hội có thể phân tích “mạng lưới phủ sóng” của top 15 hãng hàng không năm 2022, qua đó tính được độ khó/dễ khi di chuyển qua các nước bằng đường hàng không qua các thuật toán xếp hạng, phân cụm.

Xác định bài toán

Input: Tập dữ liệu Airline ban đầu trên nguồn dữ liệu Data World đã qua xử lí tiền dữ liệu.

Output: Đưa ra độ đo, đưa ra cộng đồng phục vụ cho việc phân tích mạng xã hội “AirTravelNetwork”

* 1. Tập dữ liệu

1.3.1 các tập dữ liệu

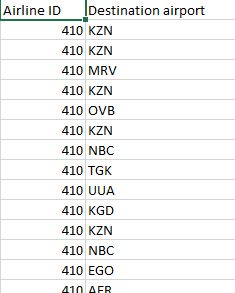
Dữ liệu về top 15 hãng hàng không năm 2022: <https://www.businessinsider.com/top-20-best-airlines-in-world-none-in-us-qatar-2022-9>

Dữ liệu về các quốc gia phủ sóng của các hãng hàng không:

<https://data.world/datasets/flight-routes>

1.3.2 Giải thích ý nghĩa về tập dữ liệu:

Data được tổng hợp từ 3 file csv như sau

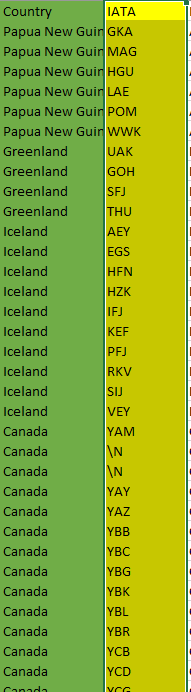


*Hình: file Route.csv gồm có Airline ID (Mã hãng hàng không) và*

*Destination airport (Mã IATA) của sân bay mà máy bay của hãng hàng không đó bay tới.*

**

*Hình: File Airlines.csv, gồm có Airline ID (mã hãng hàng không) và Name (tên hãng hàng không). Cột màu xanh lá là thông tin được lấy ra, cột màu vàng dùng để làm khóa ngoại với cột Airline ID trong file Route.csv*

**

*Hình: File Airports.csv, cột màu vàng dùng làm khóa ngoại với cột Destination Airport trong file Route.csv.*

Sau khi kết data từ 3 file csv như trên, ta được file csv chứa dữ liệu cuối cùng như sau:

**

*Hình: file csv. chứa data cuối cùng*

|  |  |  |
| --- | --- | --- |
| Tên cột | Kiểu dữ liệu | Ý nghĩa |
| Airline | String | Tên hãng hàng không |
| Country | String | Tên quốc gia mà hãng hàng không đi đến |

Bảng dữ liệu mẫu

|  |  |
| --- | --- |
| Airline | Country (destination) |
| AirFrance | United States |
| Air France | Cote d'Ivoire |
| Air France | Nigeria |
| … | … |

Bài toán cần giải quyết: Dựa trên dataset trên, chọn ra các đỉnh, cạnh và áp dụng các thuật toán mạng xã hội để tìm ra quốc gia/ vùng tự trị nào là dễ đi đến bằng đường hàng không nhất.

Đề xuất mô hình xử lý:

Lấy các quốc gia/ vùng tự trị làm **ĐỈNH**

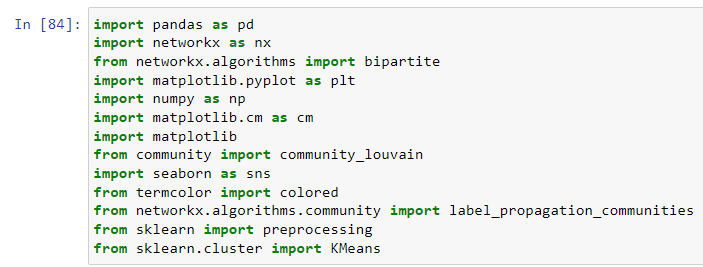
**CẠNH** là đường nối giữa hai quốc gia/vùng tự trị mà chúng cùng là điểm đến của 1 hãng hàng không

CHƯƠNG 2: XỬ LÝ VÀ PHÂN THÍCH DỮ LIỆU

2.1 Làm sạch dữ liệu

Link Source Code:

Các thư viện được sử dụng:



Hình 1: Import thư viện vào file source code

Tiến hành đọc file csv:



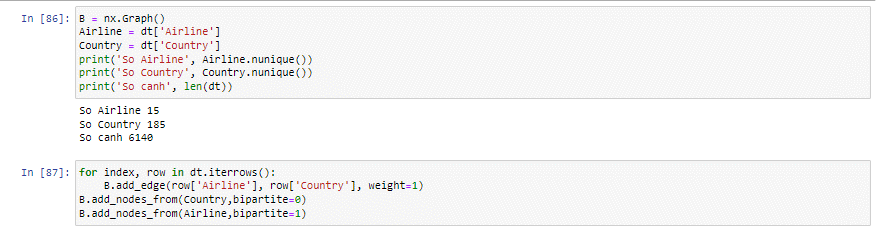
Hình 2: đọc file csv, xóa các dòng dữ liệu lặp lại và rỗng.

2.2 Chuyển đổi dữ liệu trên thành đồ thị

Đồ thị hai phía

Node: là tên của hàng không và tên các quốc gia phủ sóng của nó

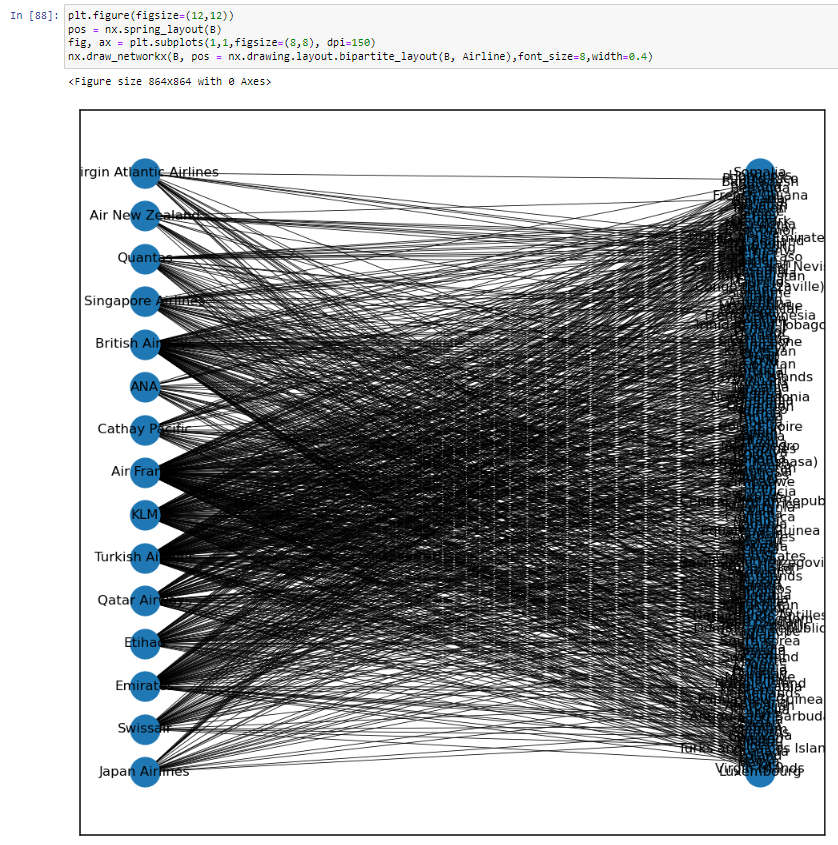
Edge: mối quan hệ giữa các quốc gia của các hãng hàng không



Hình 3: Đưa dữ liệu đã tiền xử lí vào đồ thị vô hướng

Chúng ta có thể thấy 15 hãng hàng không, 185 quốc gia/ vùng tự trị và 6140 cạnh của đồ thị.

Hiển thị đồ thị hai phía



Hình 5: Đồ thị hai phía

Nhìn vào đồ thị có thể thấy mỗi hàng hàng không phủ sóng đến rất nhiều quốc gia khác nhau và mỗi quốc gia khác nhau cũng có nhiều hãng hàng không chọn làm điểm đến.

Đồ thị một phía

Node: là các quốc gia (Country)

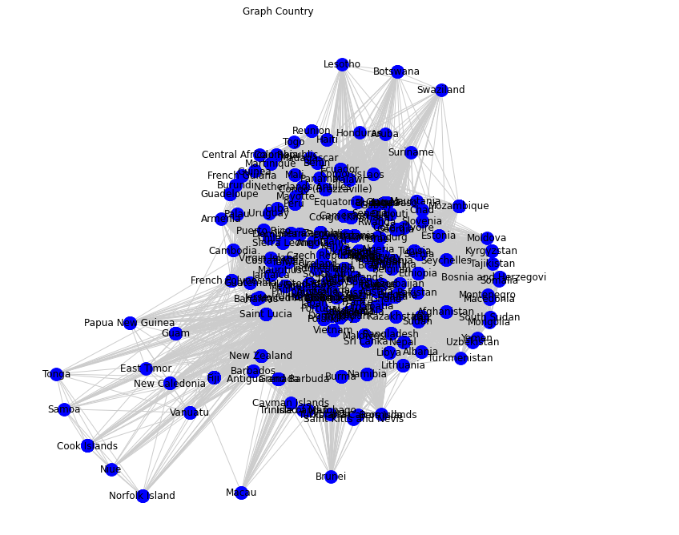
Edge: Hai quốc gia cùng là điểm đến của 1 hãng hàng không sẽ hợp thành 1 cạnh

Điều này nói lên rằng mỗi hàng hàng không vận chuyển khách tới nhiều quốc gia/vùng tự trị khác nhau, tại mỗi quốc gia/vùng tự trị có sự cạnh tranh giữa các hãng hàng không.

Weight: Trọng số là số quốc gia cùng là điểm đến của các hãng hàng không

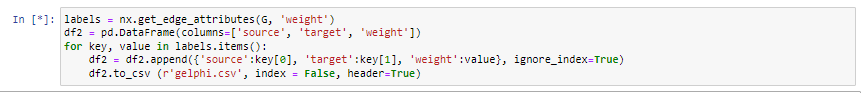


Hình 6: Code hiển thị đồ thị 1 phía

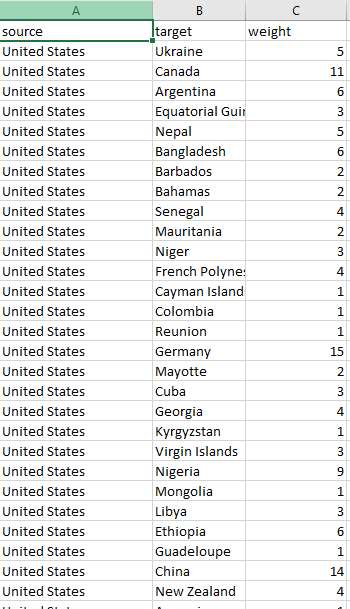


Hình 7: Đồ thị 1 phía

Xuất đồ thị 1 phía ra file csv để thực hiện trên phần mềm Gephi:



*Hình 8: Code xuất file csv để thực hiện trên Gephi*



Hình 9: File csv để thực hiện trên Gephi

Source và Target là 2 node 2 đầu của cạnh, là 2 quốc gia cùng là điểm đến của ít nhất 1 hãng hàng không.

Weight: Trọng số, số hãng hàng không chọn cả 2 nước source target làm điểm đến.

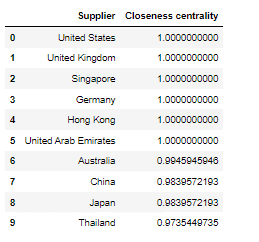
CHƯƠNG 3 CÁC THUẬT TOÁN XẾP HẠNG

3.1 Closeness Centrality

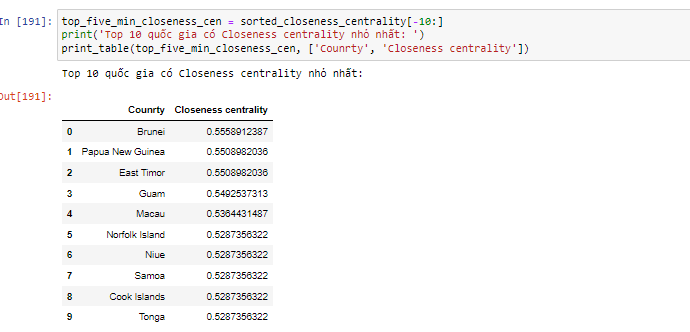
Closeness centrality là độ đo khoảng cách từ một đỉnh đến các đỉnh còn lại trong đồ thị. Đỉnh có giá trị Closenness Centrality cao nhất sẽ là đỉnh có thể truyền đạt, tiếp nhận thông tin từ các node khác trong mạng một cách nhanh nhất, ít tốn thời gian nhất.

Ý nghĩa: những quốc gia có closeness centrality cao là những quốc gia được phủ sóng bởi nhiều hãng hàng không khác nhau

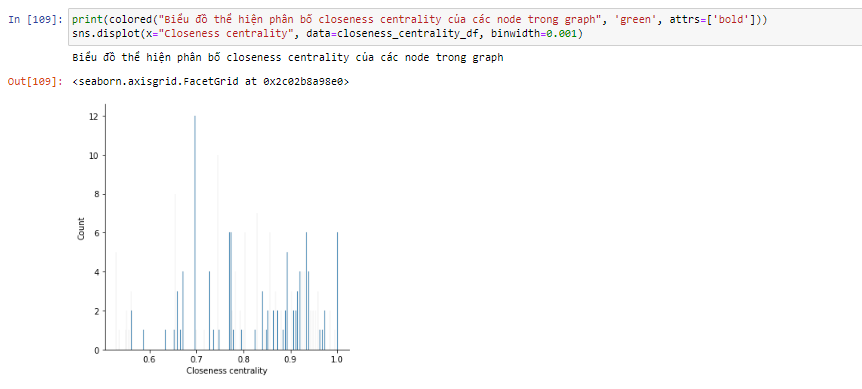
3.1.1 Sử dụng Python



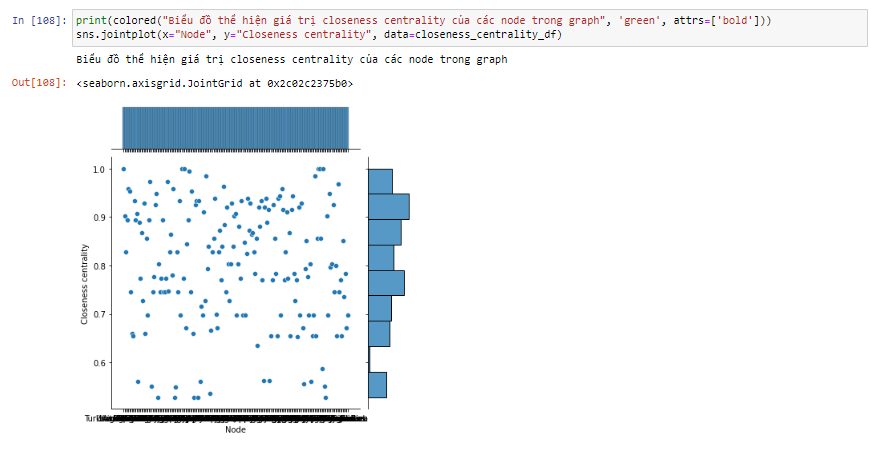
Hình 9 Top 10 quốc gia có Closeness Centrality cao nhất



Hình 10 Top 10 quốc gia có closeness centraltiy thấp nhất

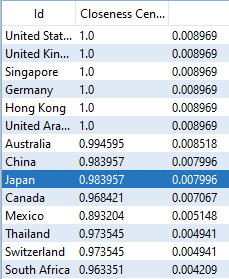


Hình 11: Biểu đồ thể hiện phân bố closeness centrality của các node trong graph

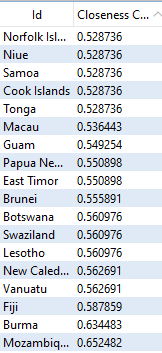


Hình 12: Biểu đồ thể hiện giá trị closeness centrality của các node trong graph

3.1.2 Sử dụng Gephi



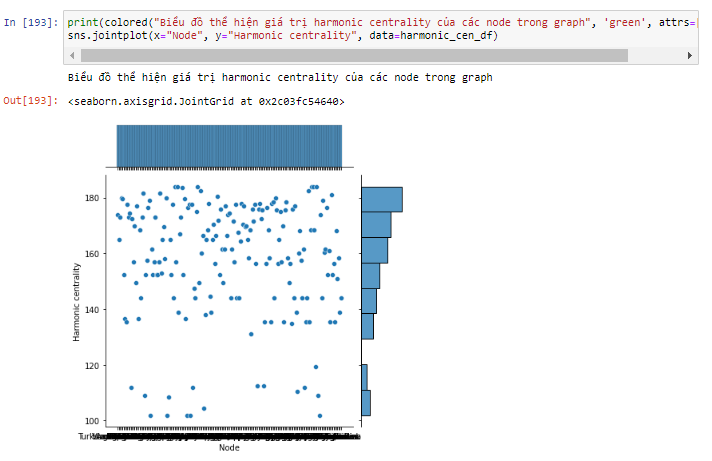
Hình 13: Top 10 quốc gia có closeness centrality cao nhất



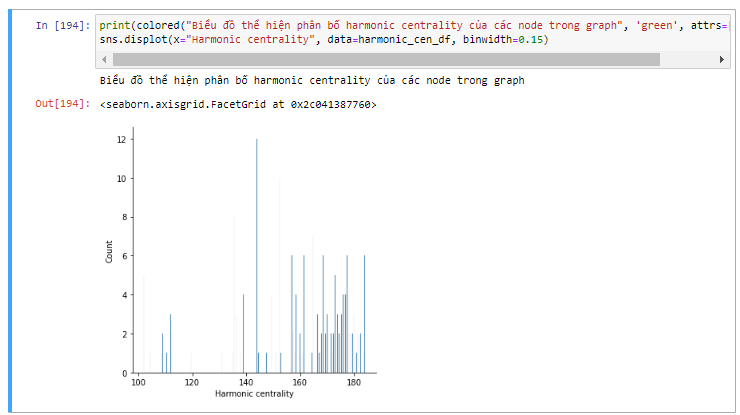
Hình 14: Top 10 quốc gia có closeness centrality thấp nhất

3.2 Harmonic Centrality

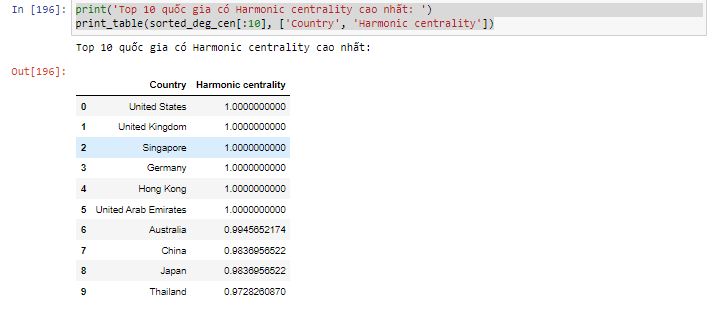
3.2.1 Sử dụng Python



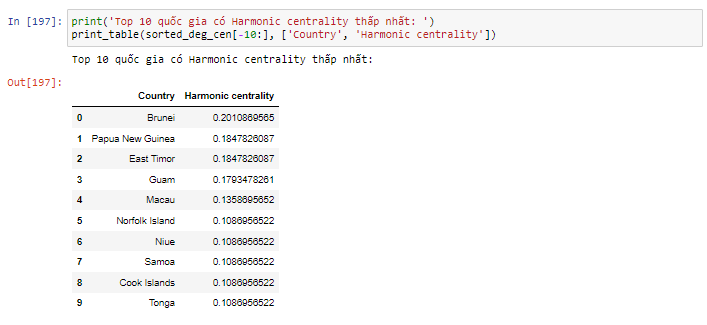
Hình 15: Biểu đồ thể hiện giá trị Harmonic centrality của các node trong graph



Hình 16: Biểu đồ thệ hiện phân bố harmonic centrality của các node trong graph

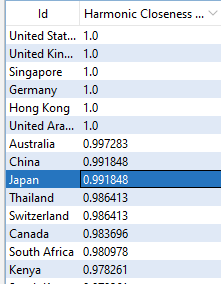


Hình 17: Top 10 quốc gia có Harmonic Centrality cao nhất

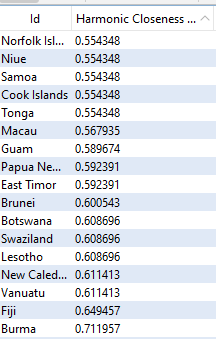
**

*Hình 18: Top 10 quốc gia có Harmonic Centrality thấp nhất*

3.2.2 Sử dụng Gephi

**

*Hình 18: Top 10 quốc gia có Harmonic Centrality cao nhất*

**

*Hình 19: Top 10 quốc gia có Harmonic Centrality thấp nhất*

3.3 Page Rank

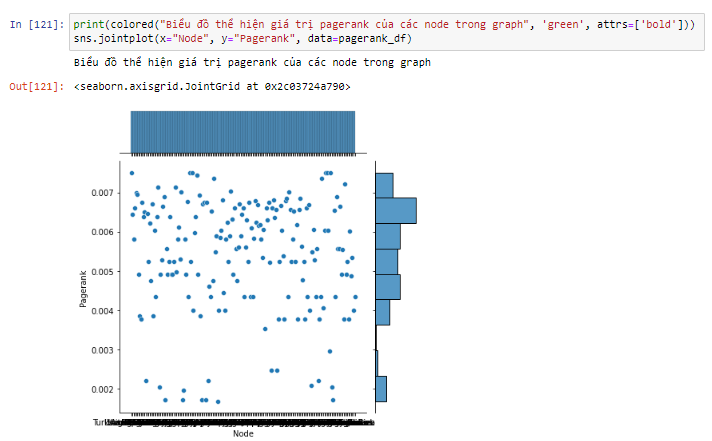
- PageRank là một thuật toán đo lường ảnh hưởng bắc cầu hoặc kết nối của các nút.

- 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.

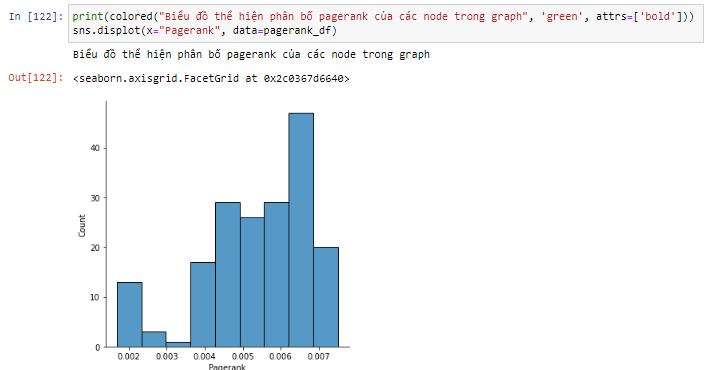
- Thuật toán áp dụng cho toàn bộ dữ liệu cần được truy xuất, là truy vấn độc lập.

Ý nghĩa: PageRank có độ đo càng cao thì các quốc gia đó càng có nhiều đường bay tới quốc gia khác trên một hãng hàng không và ngược lại.

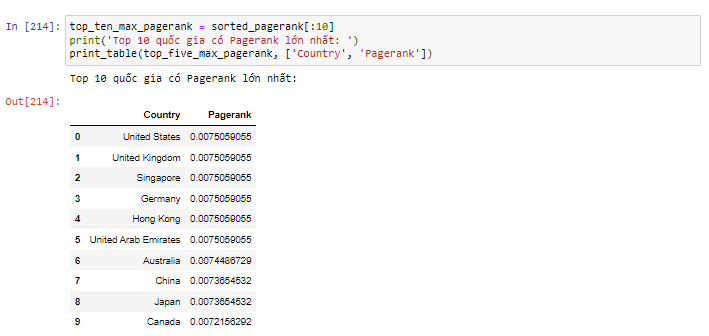
3.3.1 Sử dụng Python

**

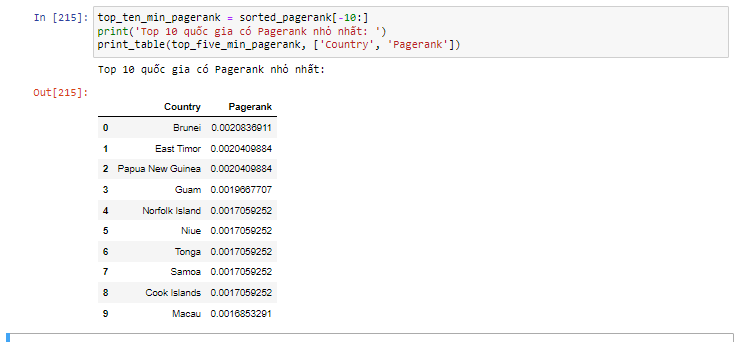
*Hình 20: Biểu đồ thể hiện giá trị pagerank của các node trong graph*

**

*Hình 21: Biểu đồ thể hiện phân bố pagerank của các node trong graph*

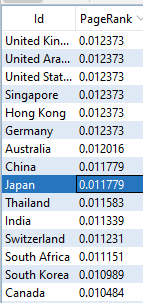
**

*Hình 22: top 10 quốc gia có page rank lớn nhất*

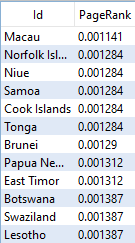
**

*Hình 23: top 10 quốc gia có page rank nhỏ nhất*

3.3.2 Sử dụng Gephi

**

*Hình 24: top 10 quốc gia có page rank lớn nhất*

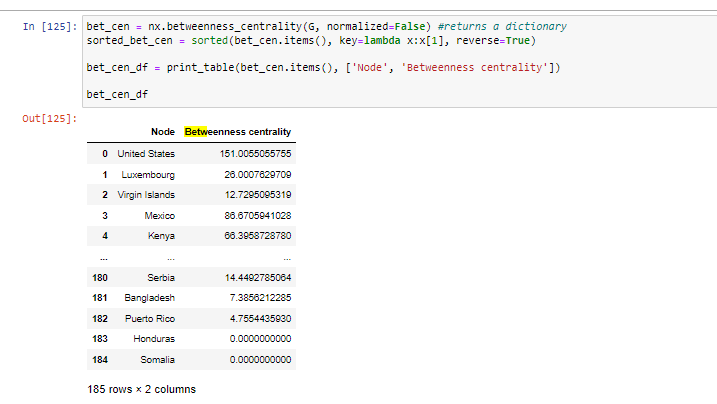
**

Hình 25: top 10 quốc gia có page rank nhỏ nhất

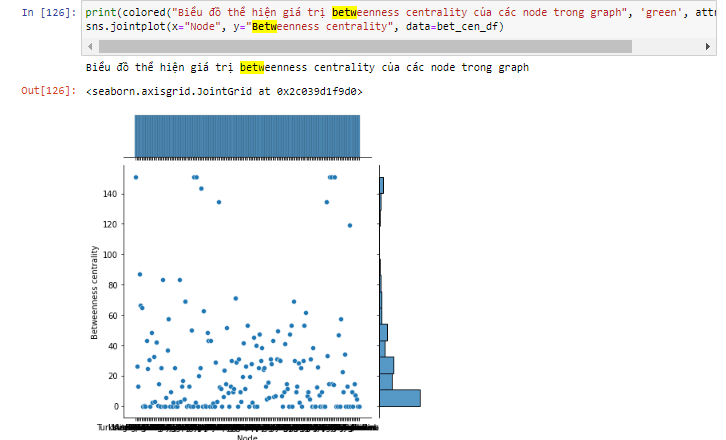
3.4.1 Betweenness Centrality

* Betweenness centrality được định nghĩa như tổng tỷ số của các đường đi ngắn nhất từ một node tới một node khác đi qua một node cho trước. Độ đo này dung để xem xét khả năng chi phối các quan hệ giữa các nút khác trong mạng. Node có giá trị này càng lớn thì node đó sẽ có sự ảnh hưởng càng lớn đến việc phân bố cấu trúc các cụm hay nhóm trong mạng càng lớn.
* Ý nghĩa: Khi một người chọn hãng hàng không để bay ra nước ngoài, họ sẽ biết tới những hãng hàng không khác qua các nước có betweness cao

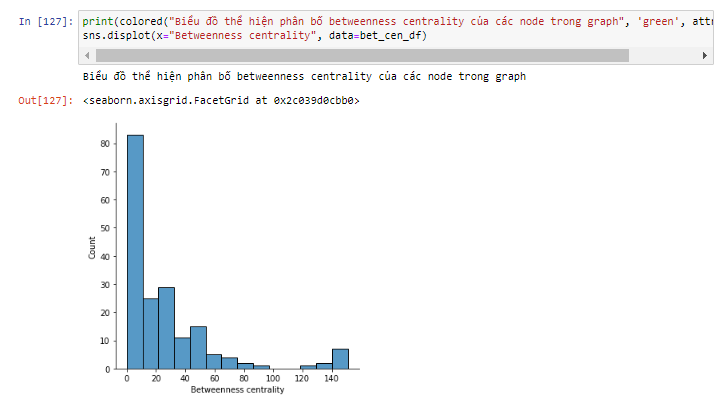
### 3.4.2 Sử dụng Python



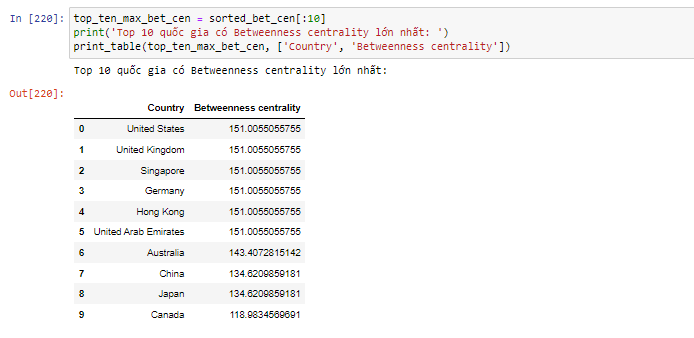
*Hình 26 Kết quả giá trị Betweenness Centrality*



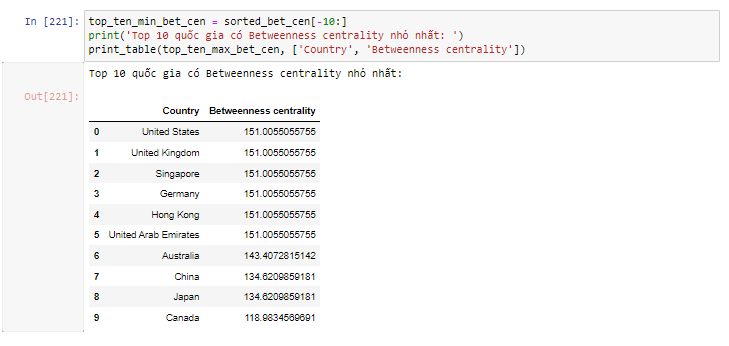
*Hình 27 Biểu đồ thể hiện giá trị Betweenness Centrality của các node trong đồ thị*

**

*Hình 28 Biểu đồ thể hiện phân bố Betweenness Centrality của các node trong đồ thị*

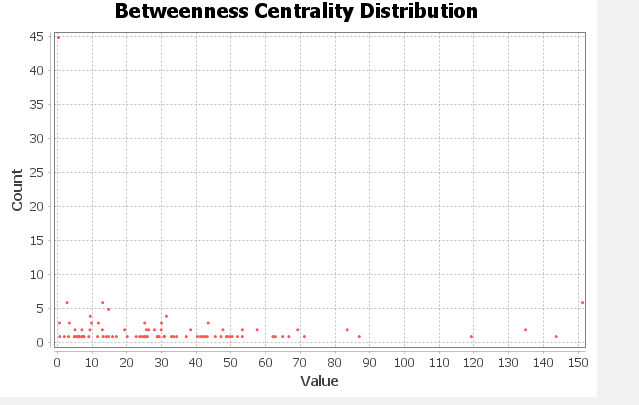


*Hình 29 Top 10 cao nhất giá trị Betweenness Centrality*



*Hình 30 Top 10 thấp nhất theo giá trị Betweenness Centrality*

### 3.4.3 Sử dụng Gephi



*Hình 31 Biểu đồ phân phối theo giá trị Betweennes Centrality Distribution*



*Hình 32 Top 10 cao nhất theo giá trị Betweeness Centrality*



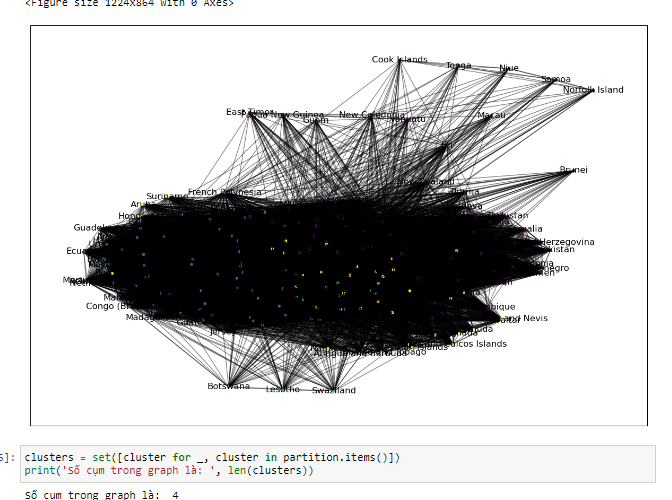
*Hình 33 Top 10 thấp nhất theo giá trị Betweeness Centrality*

CHƯƠNG 4: CÁC THUẬT TOÁN GOM CỤM

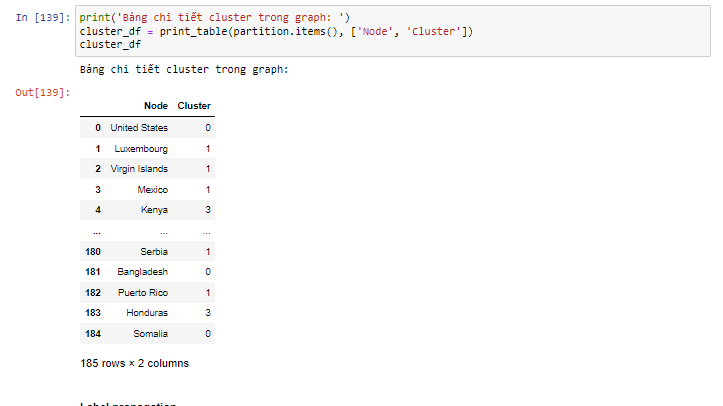
4.1 Thuật toán Louvain

Thuật toán Louvain là một phương pháp phân chia cộng đồng và thực hiện lặp đi lặp lại việc phân chia cộng đồng nhiều lần để có được mô đun tối đa của toàn bộ mạng.

4.1.1Sử dụng Python

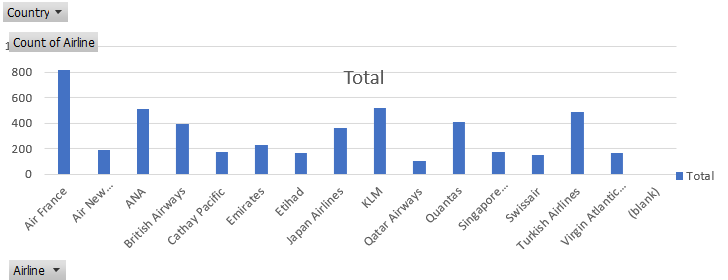


Hình 34: Kết quả chạy Louvain bằng python

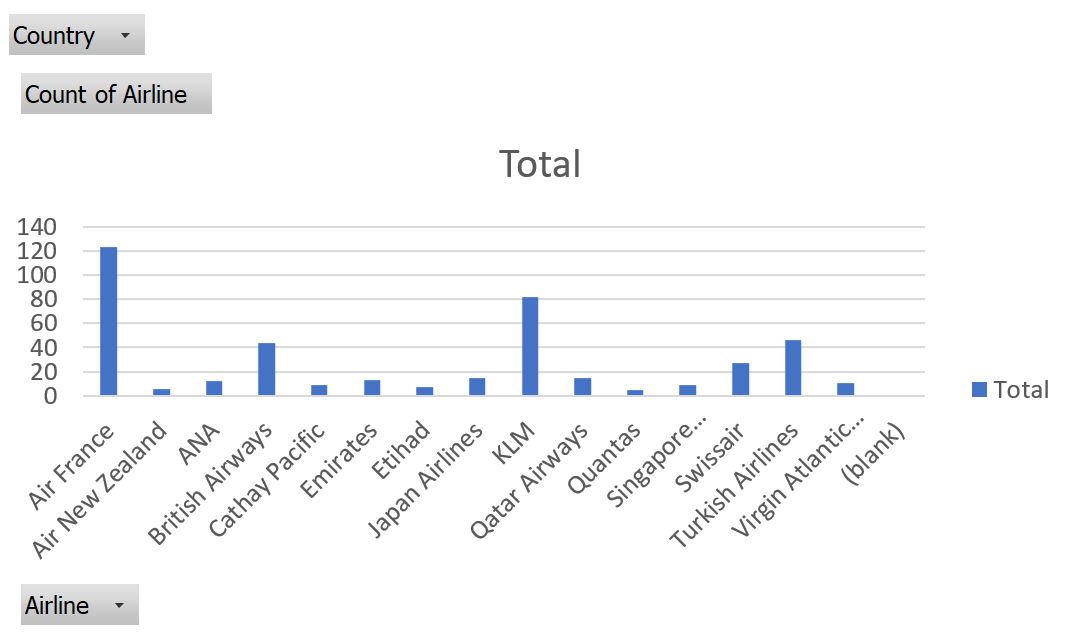


Hình 35 Bảng chi tiết các cụm Louvain

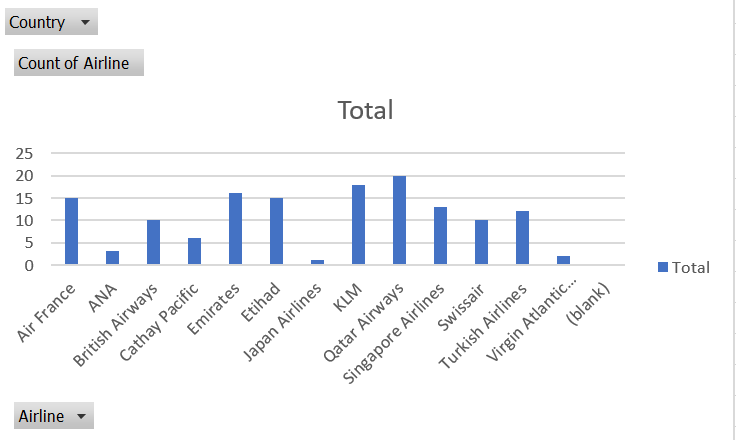
|  |  |  |  |
| --- | --- | --- | --- |
| Cụm 1 | Cụm 2 | Cụm 3 | Cụm 4 |
| United States | Luxembourg | Botswana | Kenya |
| Poland | Virgin Islands | India | Grenada |
| Vietnam | Mexico | Kuwait | Turks and Caicos Islands |
| Sudan | Haiti | Swaziland | Nigeria |
| Mongolia | Malta | Brazil | Ghana |
| Thailand | Finland | Uganda | Austria |
| Papua New Guinea | Comoros | Egypt | Antigua and Barbuda |
| Saudi Arabia | Latvia | Lesotho | Netherlands |
| Norfolk Island | Azerbaijan |  | Czech Republic |
| Switzerland | Guinea | Burundi |
| South Korea | Peru | Kazakhstan |
| Niue | Bahamas | Barbados |
| Guam | Martinique | Russia |
| South Sudan | Uruguay | Namibia |
| United Kingdom | Bulgaria | Aruba |
| Singapore | Armenia | Cyprus |
| Australia | Togo | Zimbabwe |
| Samoa | Mayotte | Qatar |
| Jordan | Palau | Spain |
| Cook Islands | Jamaica | Morocco |
| Afghanistan | Georgia | Algeria |
| Bosnia and Herzegovina | Laos | Romania |
| Iraq | Dominican Republic | Cayman Islands |
| China | Guadeloupe | Albania |
| Macau | Netherlands Antilles | Portugal |
| Seychelles | Croatia | Libya |
| Ireland | Colombia | Isle of Man |
| South Africa | Norway | Hungary |
| Iran | Malawi | Lithuania |
| Maldives | Slovenia | Trinidad and Tobago |
| Indonesia | Taiwan | Belgium |
| Uzbekistan | Equatorial Guinea | Argentina |
| Sri Lanka | Costa Rica | Italy |
| Moldova | Angola | Suriname |
| Montenegro | Mali | Saint Kitts and Nevis |
| Ethiopia | Mauritania | Greece |
| Burma | Rwanda | Gibraltar |
| New Caledonia | Central African Republic | Bermuda |
| France | Saint Lucia |  |
| Vanuatu | Chad |
| Oman | Israel |
| Kyrgyzstan | Congo (Kinshasa) |
| Turkey | Panama |
| Sweden | Philippines |
| Nepal | Zambia |
| Mozambique | Gabon |
| Yemen | Cote d'Ivoire |
| Brunei | Liberia |
| Estonia | Cuba |
| Turkmenistan | Tunisia |
| Tajikistan | Cameroon |
| Japan | Cambodia |
| New Zealand | Tanzania |
| Germany | Niger |
| Hong Kong | Sierra Leone |
| Fiji | Ukraine |
| United Arab Emirates | Ecuador |
| East Timor | Senegal |
| Tonga | French Polynesia |
| Malaysia | Madagascar |
| Macedonia | Congo (Brazzaville) |
| Denmark | Mauritius |
| Pakistan | Guatemala |
| Canada | Burkina Faso |
| Bangladesh | Djibouti |
| Somalia | Chile |
|  | Jersey |
| Reunion |
| French Guiana |
| Benin |
| Serbia |
| Puerto Rico |



Cụm 1



Cụm 2



Cụm 3

Cụm 4

Nhận xét: Thuật toán Louvain không phù hợp để gom cụm với dataset này, bởi đặc tính riêng của từng cụm không thật sự nổi bật

4.2 Phân cụm bằng Girvan Newman

Girvan Newman là thuật toán tìm và loại bỏ các cạnh nối các cộng đồng

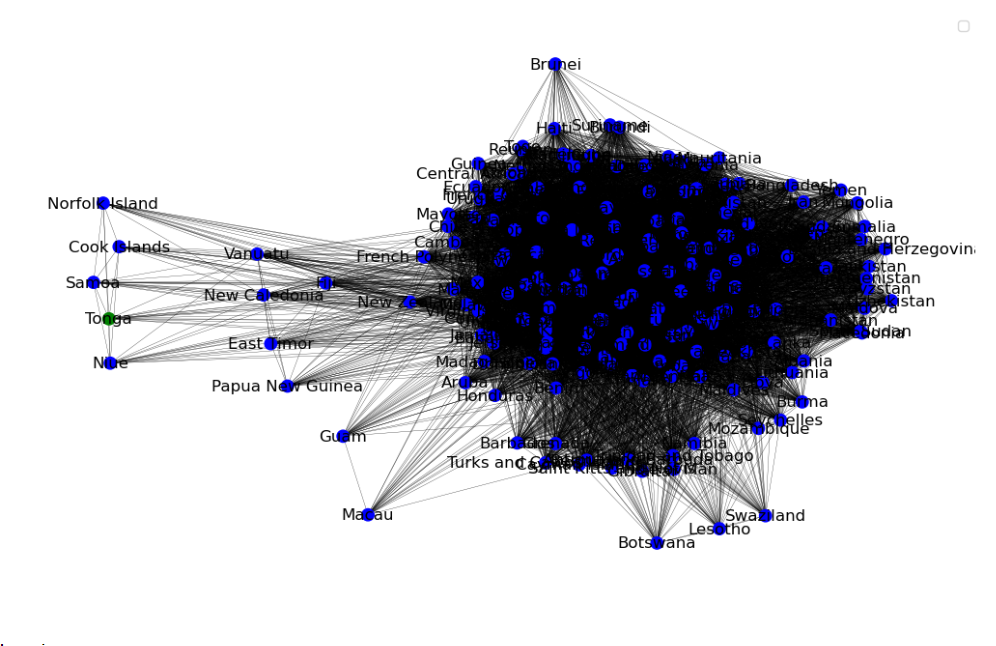
trong mạng, các cạnh đóng vai trị ‘trung gian’ trong mạng lưới. Một đường đi giữa 2

đỉnh trong mạng thuộc 2 cộng đồng khác nhau nhất thiết phải đi qua ít nhất một trong

số các cạnh như vậy, vì thế nếu ta thiết lập các đường đi giữa tất cả các đỉnh trong

mạng và xác định được cạnh nào trong đồ thị được đi qua nhiều nhất, ta có thể loại bỏ

nó để phân chia mạng thành cộng đồng riêng biệt.



Đồ thị phân cụm bằng Girvan Newman

Dự vào phân cụm bằng Girvan Newman cho thấy, cụm 1 gồm Tonga (một đảo quốc rất nhỏ gần Aurstralia và NewZealand), rất khó khăn để di chuyển tới đây bằng máy bay.

Cụm 2 gồn những quốc gia còn lại, các hãng hàng không có thể tới đó dễ dàng.

PHẦN 4: BÀI TẬP TÍNH TAY CÁC ĐỘ ĐO, GOM CỤM TRÊN 10 NODE NGẪY NHIÊN