CMSC 734 Homework-2 : Analysis of Global Flight Route Data using NodeXL

Varun Manjunatha
November 12, 2013

1 Introduction

NodeXL [1] is a powerful software that is installable as a plugin for Microsoft Excel. It’s strengths are in handling any kind of data that can be represented as a network or graph. Therefore, a natural dataset for NodeXL to analyse is any bus, train or flight route data - because each journey has a clear-cut “source” and “destination”, which can form vertices and can be connected by the edge of a graph. In this work, we come up with three significant findings after analyzing flight-route data.

2 Data

We used the data from [2], which contains data in a series of Excel files. First, we have data for flights between different countries. We also have data for flights between cities, including frequency of flights between any two given cities. Finally, we have a list of all cities that have airports with latitude and longitudinal data (which can be used for geographical overlay). I had to append continent information to the data, and this was done based on geography and not sovereignty (for example, Greenland belongs to North America even though it is a Danish colony). Almost 3000 airports in 226 countries/territories have been listed - it is not clear whether this is an exhaustive list of every airport in the world or not. To test, I searched and found some extremely obscure and infrequently used airports in India - this makes me believe that the list is fairly comprehensive.

3 Network Analysis

We did two types of analysis - one was a country level analysis and the other was a city level analysis. The country level analysis was useful to answer questions like “In terms of air traffic, which countries are the busiest?”, or “If I have to travel from point A to point B, what are the countries I am likely to land in?”. The city level analysis answers questions like “What are the most well connect cities in the world by air?”. To answer questions of this nature, we computed a common statistical measure of connectivity - which is the betweenness centrality [3]. The betweenness centrality of a given vertex measures the number of shortest paths from all vertices of the graph to all other vertices that pass through the given vertex.

4 Headlines

4.1 Headline 1 : Air Traffic Hubs Require Both Infrastructure and Strategic Location

We were able to isolate each continent and do an analysis for each country in the continent separately (Figure 1). For example (Figures 2 and 3), in Asia, the airline hubs are found to be Singapore and United Arab Emirates, in addition to cities in India and China. In Africa, the countries with busy air traffic are Kenya, South Africa and Morocco. In Europe - France, Germany and United Kingdom, in North America - United States and Canada, in Oceania - Australia, New Zealand and Fiji and in South America - there was no country that had high betweenness centrality (which means global flights generally do not land in South America). We can conclude from the above findings that to be an airline hub, a nation must not only have the infrastructure to handle thousands of flights daily, but also be located strategically (unlike the cities in South America). By using the dynamic filters in NodeXL, we could also search for countries for the most flights between them. We found that United States and Canada exchanged the most flights between each other making them the two most mutually connected nations in the world.
Figure 1: Flight connections for every country in the world. Size of the dot is proportional to betweenness centrality

4.2 **Headline 2 : The Surprise Package is Anchorage, Alaska**

We performed a detailed analysis on a global, continental and country-wise scale of betweenness centrality in cities, and a geographical map of this is given in Figure 4. Remember : this is not an airport ranking based on number of passengers - it is a ranking based on how connected the airport is to other air-routes. Our top 10 was :

1. London
2. Anchorage
3. Paris
4. Moscow
5. Chicago
6. Frankfurt
7. Seattle
8. Dubai
9. Tokyo
10. Toronto

The presence of Anchorage, Alaska is very surprising because it is generally not associated as a major air traffic hub. However, we must remember that Anchorage also connects a large number of more remote airports in North America, and therefore, its betweenness centrality is quite high. A detailed analysis of each continent revealed the following airports with high betweenness centrality (Figures 5 and 6) :

1. **Asia** : Beijing, Shanghai, Delhi, Calcutta, Bombay, Jakarta, Tokyo, Manila, Kuala Lumpur, Seoul, Bangkok, Taipei, Dubai, Singapore, Hong Kong

2. **Africa** : Johannesburg, Cairo, Nairobi, Abuja Algiers, Addis Ababa, Luanda, Casablanca
(a) Betweenness Centrality in Africa - circle layout. Note Morocco, Kenya and South Africa.

(b) Betweenness Centrality in Asia - circle layout. Note United Arab Emirates and Singapore, China and India.

(c) Betweenness Centrality in Europe. Note United Kingdom, Germany and France.

Figure 2: Betweenness Centrality in a Circle Layout - Africa, Asia, Europe
(a) Betweenness Centrality in North America - circle layout. Note United States and Canada.

(b) Betweenness Centrality in South America - circle layout. Note the lack of large dots (low betweenness centrality)

(c) Betweenness Centrality in Oceania. Note Australia, New Zealand and Fiji.

Figure 3: Betweenness Centrality in a Circle Layout - Americas and Oceania
3. **North America**: Chicago, Atlanta, Anchorage, New York, Dallas Fort-Worth, Honolulu, Seattle, Miami, Minneapolis, Los Angeles, San Francisco, Toronto, Montreal, Vancouver, Mexico City, Cancun, Panama City

4. **South America**: Sao Paulo, Rio De Janeiro, Santiago, Buenos Aires, Caracas, Bogota, Lima

5. **Oceania**: Sydney, Brisbane, Auckland, Melbourne, Papeete (Tahiti)


4.3 **Headline 3**: On average, any two airports in the world are connected by four flights

NodeXL can compute metrics of data other than measures of centrality. For example, it can measure average geodesic distance and maximum geodesic distance. On our data-set, the average geodesic distance is 3.93. This means, to travel between any two points on the graph, one must take 3.93 paths on average. This metric isn’t exact because paths where source and destination are the same are included in the metric. Even so, we can interpret this by saying: on average, one needs to take four flights to go from any airport to any other airport in the world. The maximum geodesic distance is 12 - this means that there are two airports in the world which required twelve flights to connect them.

5 **A Critique on NodeXL**

Overall, I feel that NodeXL is a very powerful tool but it is not very intuitive to use. It’s advantages and disadvantages are listed below:

5.1 **Advantages**

- NodeXL can work with Microsoft Excel which is a very popular and widely used program for analysis of data.
- It is extremely easy to analyze social network data (Twitter, Facebook, Gmail) with NodeXL.
(a) Betweenness Centrality for cities in Africa. Points arranged on a rectangular grid.

(b) Betweenness Centrality for cities in Asia. Points arranged on a rectangular grid and arranged country-wise.

(c) Betweenness Centrality for cities in Europe. Dynamically filtered out unimportant airports and dragged out the important airports for greater visibility.

Figure 5: Betweenness Centrality for cities in Africa, Asia, Europe
Figure 6: Betweenness Centrality for cities in the Americas and Oceania

(a) Betweenness Centrality for cities in North America - grid layout. Dynamically filtered out unimportant airports.

(b) Betweenness Centrality for cities in South America - grid layout.

(c) Betweenness Centrality for cities in Oceania.
The Grouping algorithms can easily form subsets of the data for more minute analysis. One aspect I especially liked was that we can form sub-groups from the groups (for example, group by continent first, followed by group by country).

• Easy to export images and files.
• The Dynamic Filter feature is very useful to make inspections on the data.
• The vertices are automatically listed when edge-edge columns are filled up on the edges worksheet.

5.2 Disadvantages

• It requires an introductory tutorial - it is very hard to remember what menu goes where.
• It does not have an undo option. The authors of the software acknowledge this, but admit that there is no easy fix.
• When one changes the visibility of a node to “hide” or “skip”, the only way to change it back to “show” is to change each row in the visibility column. Clearly, this is a problem when there are thousands of rows.
• When we dynamically filter, we sometimes filter out many of the nodes leading to huge empty spaces. I might be wrong, but there was no way to get the nodes to occupy the rest of the space without manually placing the nodes.
• To the best of my knowledge, it is not possible to do a geographical group in a box.

References

