Community Structure Analysis of the United States of America's Airport Network

Ana Daniel Federico Fiorini
University of Trento

The Data

The data was obtained from the Bureau of Transportation Statistics of the United States of America and it consists of information on every commercial flight made between 1987 and 2015 within the USA.

- 160+ million records
- Flights are grouped by month
- 338 files zip compressed csv's
- 827MB compressed / 18GB uncompressed

Problem

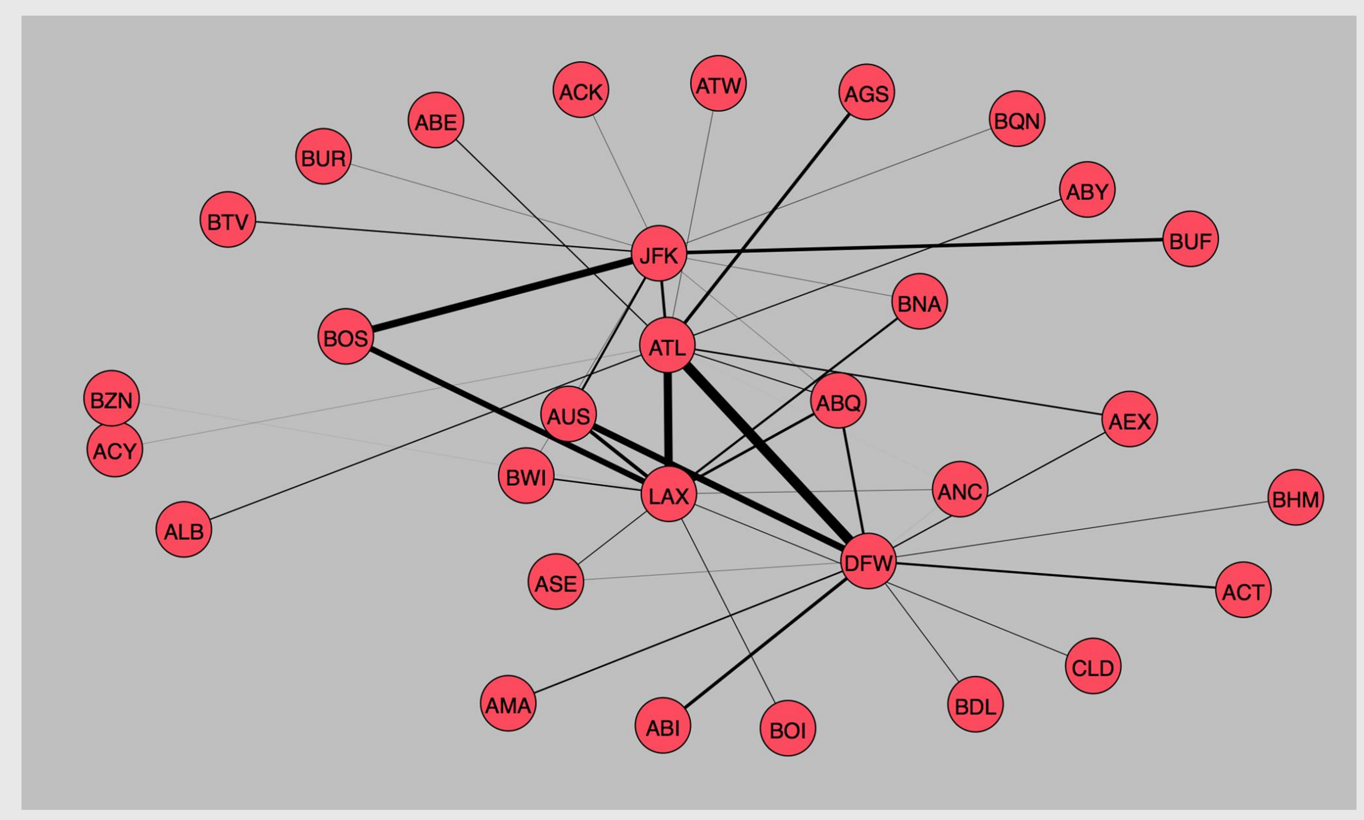
Obtain insights out of this network of airports and routes by detecting communities of cities by how well connected they are through their departing and arriving flights.

Analyse how communities of cities of the United States have changed through the years

The Graph Database

We had used Apache Spark, an open source cluster computing framework, to aggregate the 160+ million records into yearly flight routes that then were loaded into the Neo4j Graph Database.

The Neo4j database consists of 329 nodes that represent the airports of the United States of America. These nodes are connected connected by over 115,000 edges. Each edge represents the routes between 2 airports for every year and it includes information on their yearly frequency.



Detecting Communities Algorithms

The Walktrap Algorithm

Developed by Pascal Pons and Matthieu Latapy, this is an agglomerative clustering algorithm based on the logic that performing random walks in a graph will result in this walk usually getting confined in very connected areas of the graph.

Each vertex in the graph is represented as a vector of probabilities to end up in every other vertex of the graph in *t* random steps. Similar vertices get grouped into communities.

The Shortest-Path Betweenness Algorithm

Developed by M. E. J. Newman M. Girvan, this is a divisive algorithm that calculates the centrality of every edge in a graph and removes the one with the highest value in order to divide the graph into different communities.

The centrality of an edge is obtained by calculating how many shortest-paths pass through it.

The Shortest-Path Betweenness Algorithm (Inverted Weight)

An adjusted version of the shortest path betweenness algorithm in order to use properly the weight when calculating the shortest-paths. In the network of airports, the weight of an edge represents the yearly frequency of the route and a higher value in the frequency actually means that is easier to go through a route.

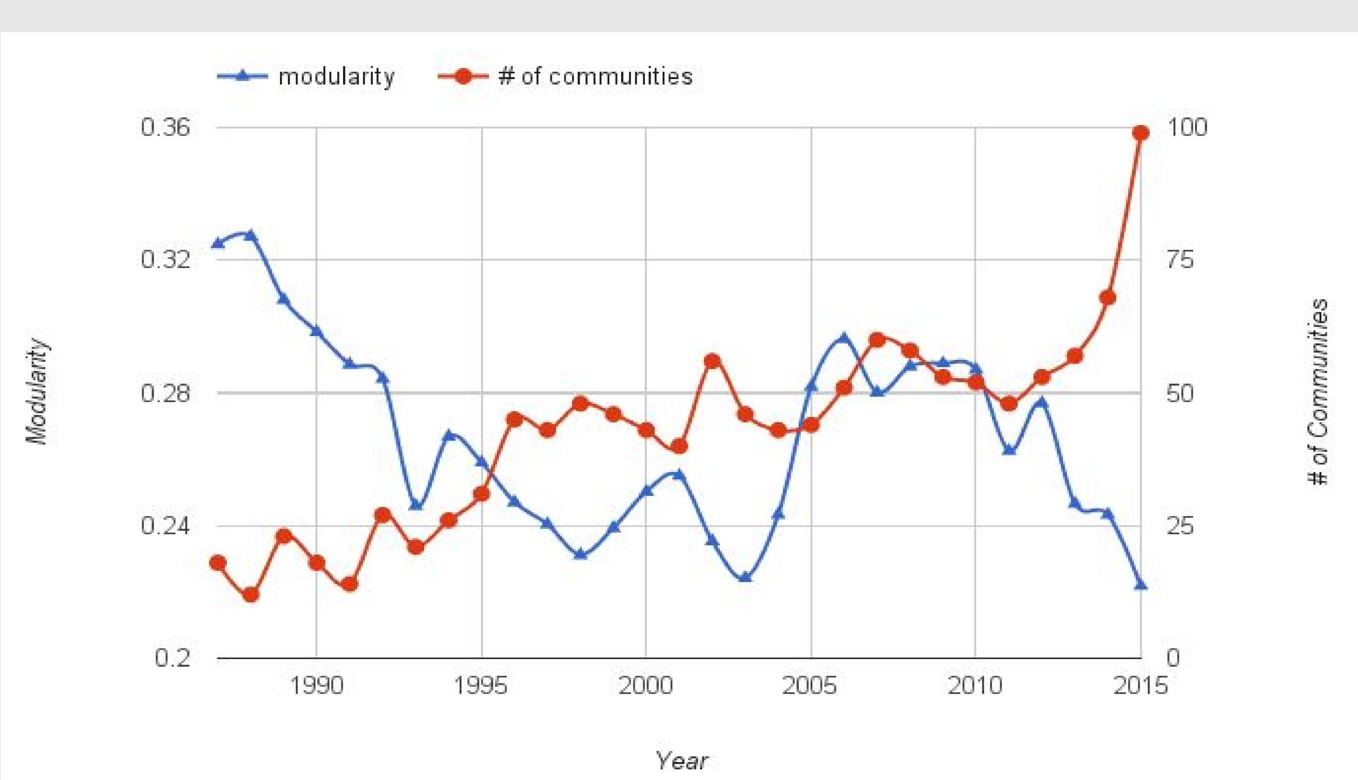
The frequency of a route is inverted to explore if this modification returns better results.

Modularity

Modularity measures the quality of a partition of a network into communities. It's obtained by comparing how many edges connect the vertices of a community to other vertices of the same community in relation to the number of the edges that are connected to other communities.

Year	Walktrap	Short-Path Betweenness	Short-Path Betweenness (Inverted Weight)
2006	**0.2963**	0.0630	0.0071
2015	**0.2218**	0.08157	0.0647

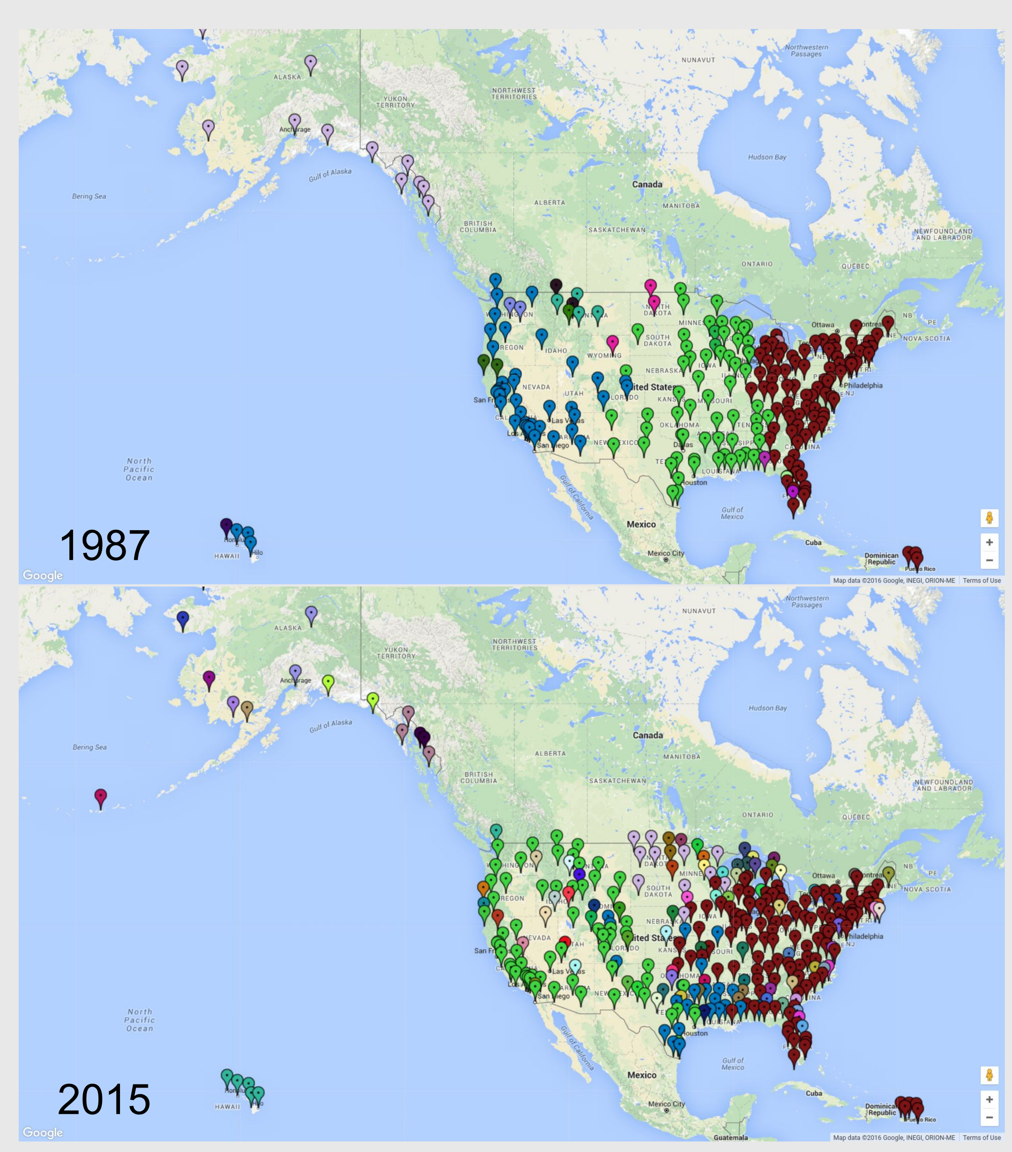
The **walktrap** algorithm performed better as it found communities with a modularity value that ranged between 0.22 and 0.29 for the communities found of the last 10 years.



The modularity of the partitions obtained by the walktrap have gradually decreased from 1987 to 2003. There was a significant increase in the modularity from 2004 to 2006, however, it started to decrease again after that period and until the present year.

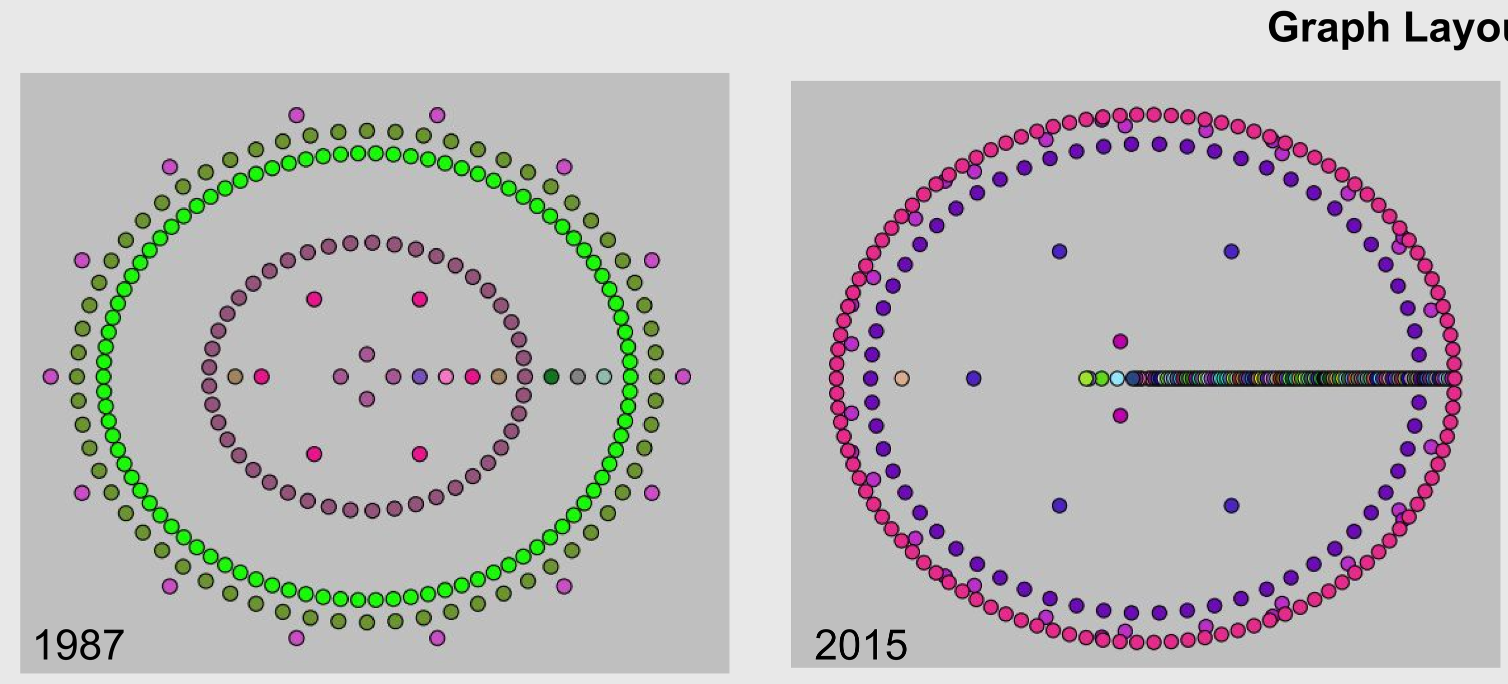
Partition Visualization

Map



Insights:
- In the 1980's, the airport community of the midwest and southwest was bigger but as time went by, many of these airports starting merging either to the West or the Northeast and Southeast regions of the United States as we can see
- In 2015 we can tell how the ones belonging to the West and to the Southeast and Northeast become less well defined as a lot of different color markers of airports belonging to other minor communities begin to appear within their areas.

Graph Layout



With *networkx*'s shell layout we can observe how in 2015 there are much more single airport communities present than in 1987. Single airport communities are evident in the plot because since there is only one airport, it is not possible to create a circle path like the shell layout is supposed to create.

Acknowledgements

The modularity of the partitions found through the years has been gradually decreasing. The fact that the walktrap algorithm is not able to obtain good quality partitions as the time advances, tells us that the network of airports of the United States of America is becoming very dense both globally and locally.

Dividing this network into good partitions is not trivial task because of the nature of the network itself. Thus, considering only the topology and the traffic is not enough for a spatial network such as the USA Airports Network.