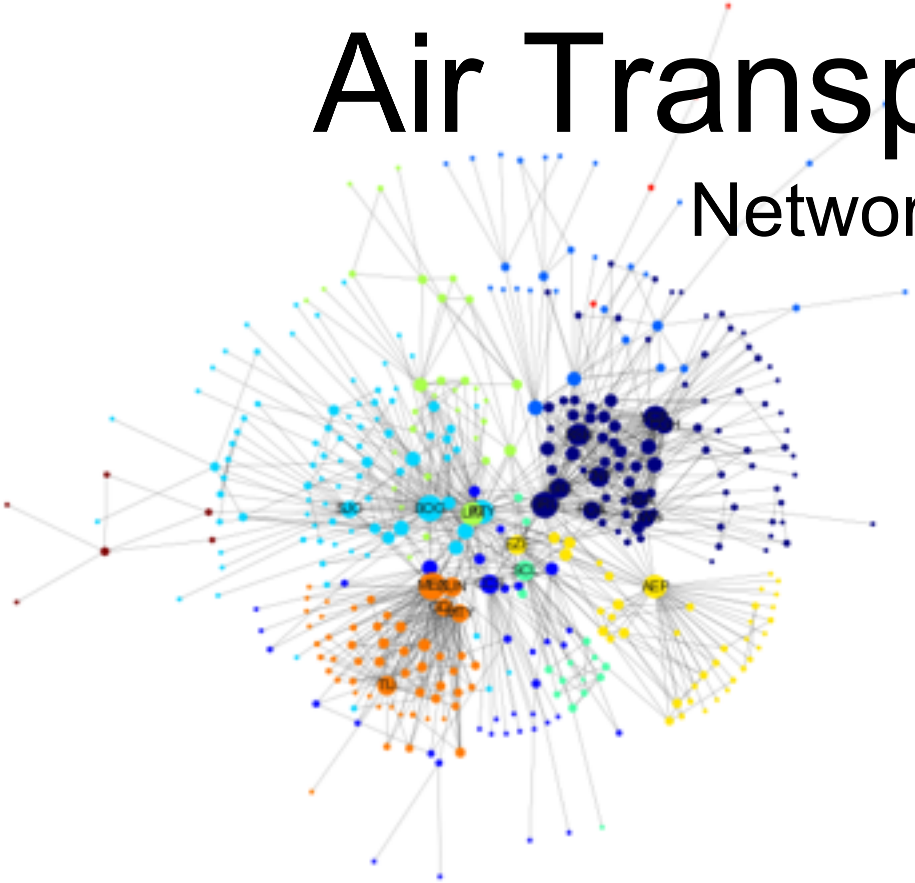
Air Transport Network

Networking Science – NS101
Final Project



Marcelo José Rovai

Universidad del Desarrollo

Master Data Science

December, 2018

Abstract

In the **Air Transport Network**, airports can be represented by nodes where flight connections between airports are the links (or edges). On this project, air transport networks will be analyzed worldwide (WW) as well as for one region (Latin America) and for individual airline companies as LATAN and RYANAIR.

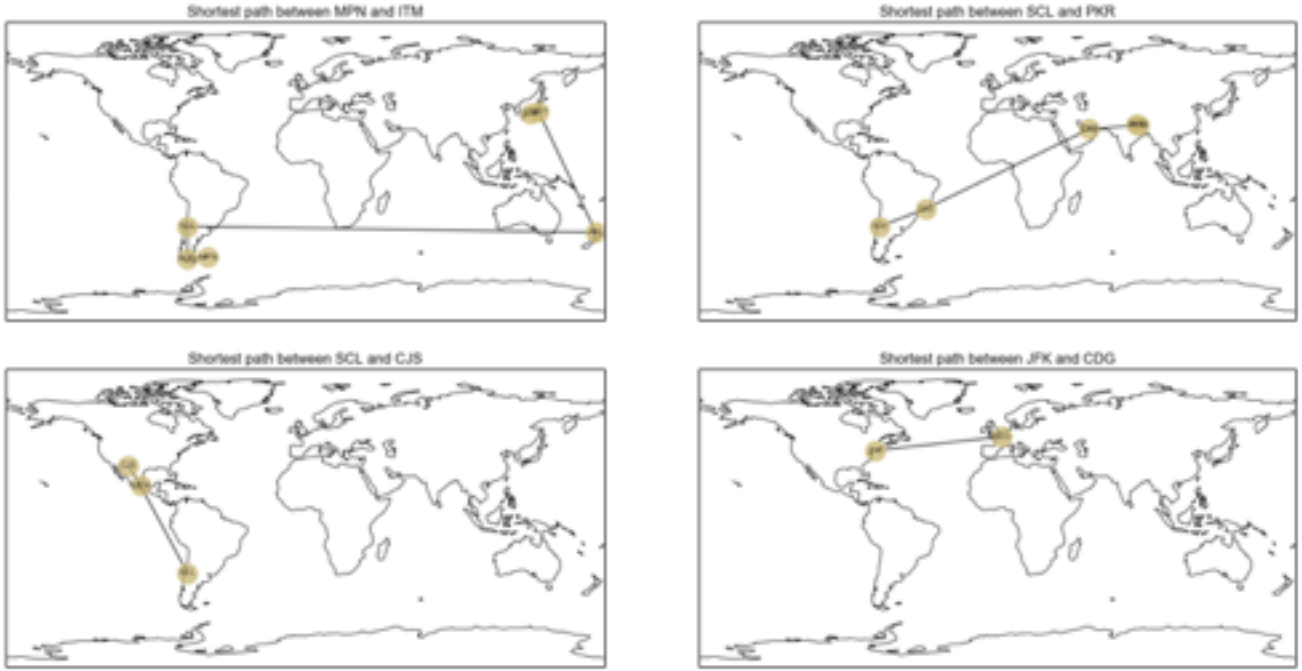
Network properties as distribution degree (k), average short path(<d>), clustering coefficient (<C>), among others will be analyzed in deep for different air network configurations, comparing them with theoretical models (as WS and BA). The main goal will to verify that air transport network, a ”complex network”, has the properties of ”small-word” and ”scale-free” networks (very similar to social networks as Facebook).

We will also explore an anomalous property of the air transport networks where nodes with relatively low degrees may have very high betweenness centrality. It is an important observation related to the robustness of the network. This characteristic, shows that critical points of the system are not necessarily the hubs (like JFK or SCL), but some other cities which uniquely provides routes to certain regions. For example, we will see more in details Alaska, that can be easily isolated from the other parts of the worldwide air transport network if a problem occurs in Anchorage.

The worldwide air transport network defines communities. These communities are mainly determined by geographical factors, however, in some cases the borders of the communities are different from the borders of geographic regions. For example, on a WW network, North of South America, Central and North America are connected, same as Cuba and Venezuela, when the regional air network is analyzed.

Motivation

- Modeling air transport networks aims airline companies to organize their routes in a cost-efficient way and therefore maximize their profits. Below examples of best route calculation using Dijkstra's algorithm:



Shortest path between MPN and ITM

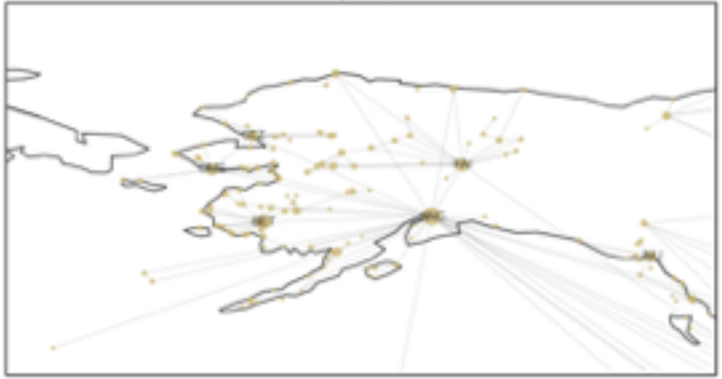
Shortest path between SCL and PKR

Shortest path between SCL and CJS

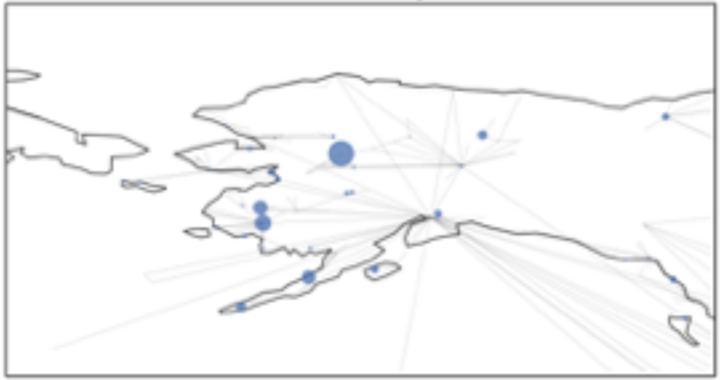
Shortest path between JFK and CDG

Motivation

- Air transport network models are also the tool to investigate system robustness. They help to determine weaknesses of the system in case of various kinds of disruptions. For example, below we can see that Alaska can be easily isolated from the other parts of the worldwide air transport network, if Anchorage (ANC) airport has a problem:



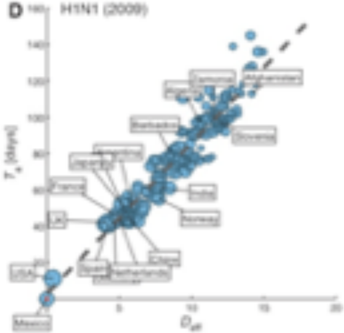
Network Degree (K) - Alaska

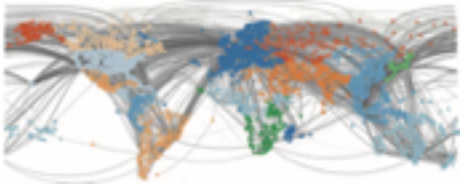


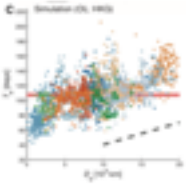
Betweenness Centrality - Alaska

Motivation

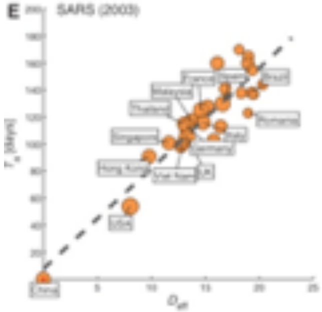
- An alternative application is modeling ”human deceases networks”. Air transport network is used by millions of people every day, therefore it plays key role in the spread of some infections, such as Influenza, H1N1 or SARS. Below, simulations of a disease spreading along air transport network:







Left figures show temporal snapshots of a dynamical system for a hypothetical pandemic with initial outbreak location (OL) in Hong Kong (HKG). [11]

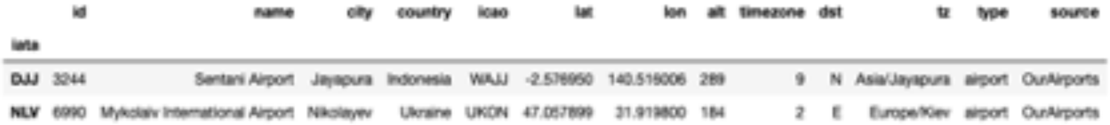


Dataset

The dataset was created from data downloaded from: https://openflights.org/data.html.

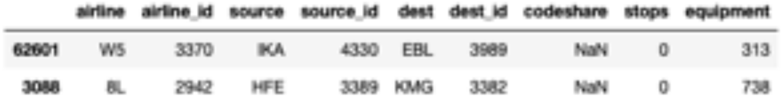
The site keeps 2 different databases:

- “OpenFlights Airports Database”, that contains over 10,000 airports (January 2017).
    - https://raw.githubusercontent.com/jpatokal/openflights/master/data/airports.dat



	id	name	city	country	icao	lat	lon	alt	timezone	dst	tz	type	source
iata													
DJJ	3244	Sentani Airport	Jayapura	Indonesia	WAJJ	-2.576950	140.516006	289	9	N	Asia/Jayapura	airport	OurAirports
NLV	6990	Mykolaiv International Airport	Nikolayev	Ukraine	UKON	47.057899	31.919800	184	2	E	Europe/Kiev	airport	OurAirports

- “OpenFlights/Airline Route Mapper Route Database”, that contains 67,663 routes between 3321 airports on 548 airlines spanning the globe (June 2014).
    - https://raw.githubusercontent.com/jpatokal/openflights/master/data/routes.dat



	airline	airline_id	source	source_id	dest	dest_id	codeshare	stops	equipment
62601	W5	3370	IKA	4330	EBL	3989	NaN	0	313
3088	8L	2942	HFE	3389	KMG	3382	NaN	0	738

Dataset Preparation and Cleaning

1. Routes dataset should be linked with Airports dataset (Using IATA codes as index).

2. Only non-stop routes will be considered.

3. Deletion of no important columns: 'airline_id’, 'source_id’, 'dest_id’, 'codeshare’ and 'stops’.

4. Creation of a new column "dist", with the aprox. distance in kilometers of any route:

    1. First create a dictionary with airport positions (LAT, LONG) starting from its IATA code

    2. Second create a function to calculate distance between airports

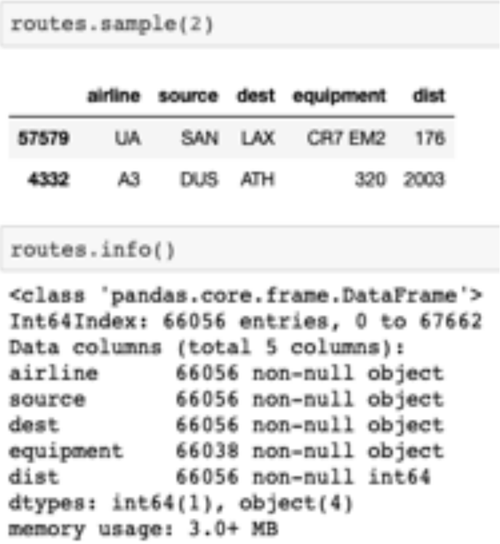
    3. Third, apply the function to dataset to calculate distance for all rows

5. Final filtering of the data to be explored:

    1. WorldWide

    2. Latin America

    3. Air Companies (LATAM and RYANAIR)



```
routes.sample(2)
```

	airline	source	dest	equipment	dist
57579	UA	SAN	LAX	CR7 EM2	176
4332	A3	DUS	ATH	320	2003

```
routes.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 66056 entries, 0 to 67662
Data columns (total 5 columns):
airline      66056 non-null object
source       66056 non-null object
dest         66056 non-null object
equipment    66038 non-null object
dist         66056 non-null int64
dtypes: int64(1), object(4)
memory usage: 3.0+ MB
```

Research Questions

- ✓ *How is the Air Transport Network (ATN) structured in terms of nodes (airports) and links (routes)? Is the ATN similar to Social Networks?*
- ✓ *How studying ATN can help on human deceases spread prevention?*
- ✓ *How to calculate the best rout between destinations?*
- ✓ *How different are the topologies of distinct Air Companies?*
- ✓ *How to hank the most important airports? How to analyze weakness?*
- ✓ *How the ATN can be split in communities?*

Creating an Air Transport Network

Let's for simplification only consider the path between 2 airports, not importing the direction of flow, i.e., we will use undirected graph. This approach is in consonance with what was described by T. Verma on [5] "Revealing the structure of the world airline network", pag.5 and 6:

"The flight data (airports, airlines, routes and geo-locations) contains some circular connections, i.e. a flight may go from A to B and not return directly to A. Instead, this flight follows a path from A to B to C and back to A. To simplify our analysis, we have made the adjacency matrices symmetric by replicating each unidirectional connection in the opposite direction. This is justified by the fact that only very small airline companies have circular connections and merely in remote parts of the world. The network is generated using the x (longitude) and y (latitude) coordinates of each airport and embedding them in a two dimensional space using an equirectangular projection of earth. The links are placed between any pair of airports if there exists a direct flight between the two."

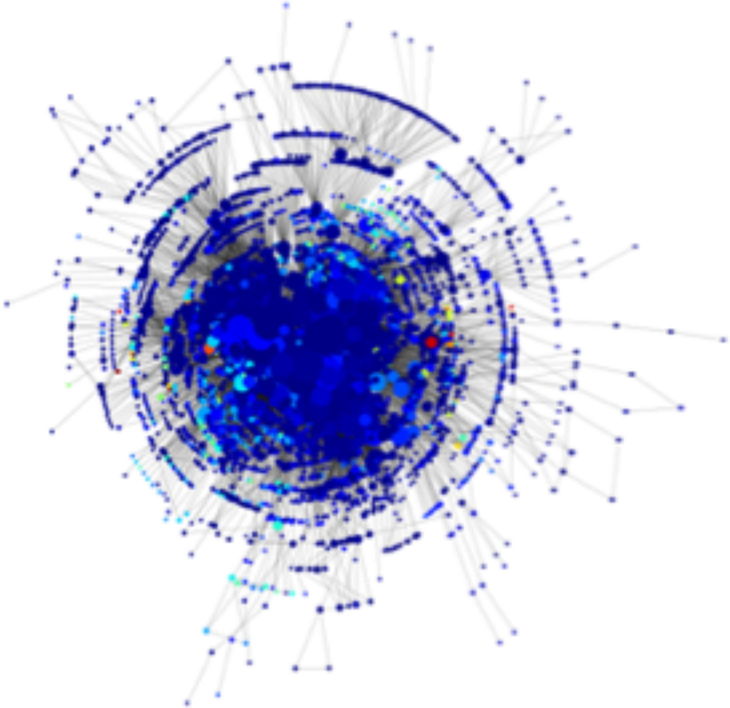
Creating an Air Transport Network

The analyses that will be developed on this project, is based on weighted projections of the air transport system, where the distance of direct connections between pairs of nodes was taken into account. On the other hand, it can be expected that "weights", as the structure of frequencies of flights, number of passengers transported, etc., may also unveil interesting information, other the ones discussed on this project. Unfortunately, the dataset available does not help with that, except for the number of companies per route that can be used as a type of weight, but I do not consider this information relevant to my work as explained next.

For example, let's take the example Santiago (SCL) to Lima (LIM) . From the dataset we can get a "weight" of "6”, what means that 6 different companies do this route. By the other rand, we can easily verify on internet, that Lima is the top destination from SCL, been 144 the number of weekly flights. Even if we take a daily average, the number will be still high, around 20. This is because companies has multiple daily flights and not all routes flight every day. Anyway, Tuan Doan Nguyen explored this possibility on a Medium article [10] and examples of different weighted routes are discussed by Massimiliano Zanin more in deep on [6], "Modeling the Air Transport with Complex Networks: a short review", pages 10 and 11.

Creating an Air Transport Network graph

Once the Graph is create, we realized that it is not connected, having a giant component that contains more than 99% of all airports. So, on our analysis only the giant component will be considered.



The graph nodes (3,154) are airports and the links (18,593), the possible routes between 2 airports.

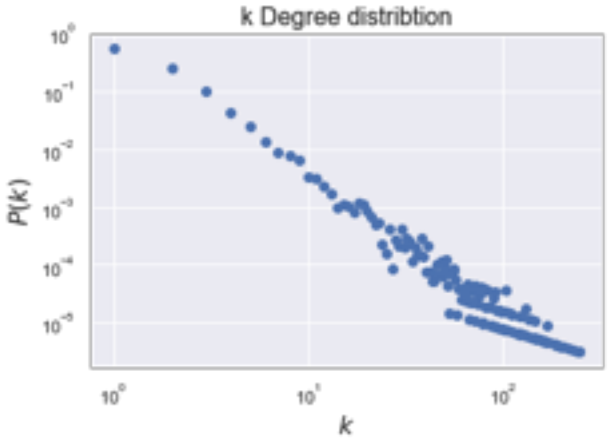
The Links (or Edges) attributes are: airline, distance and aircraft used.

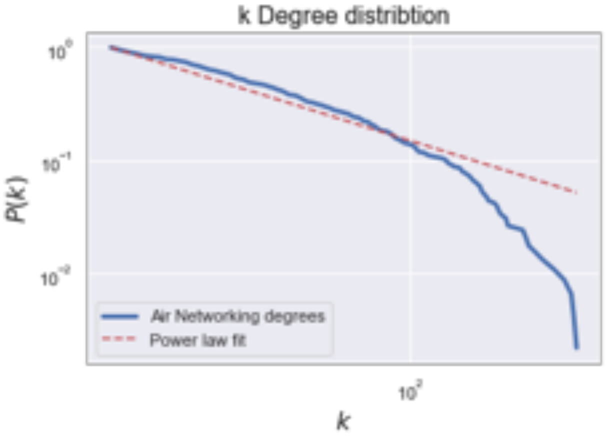
The number of different routes (k) that is related to each airport (node), varies from 1 to 246, being the global average number of routes per airport, <k>, 12.

The graph has a Diameter of "8", what means that this is the maximum distance (in paths) between 2 airports in the network, taking the shortest possible path. The average short path, <d>, is 4, what suggest that this network can be modeled using a "Small World" Model.

The average clustering coefficient, <C>, is 0.49, what is very high!

Networking Analysis

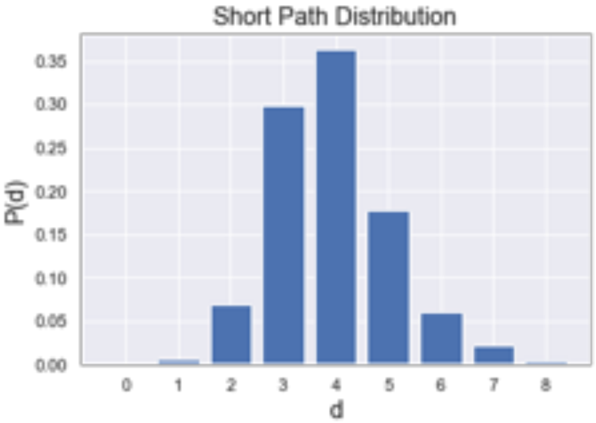




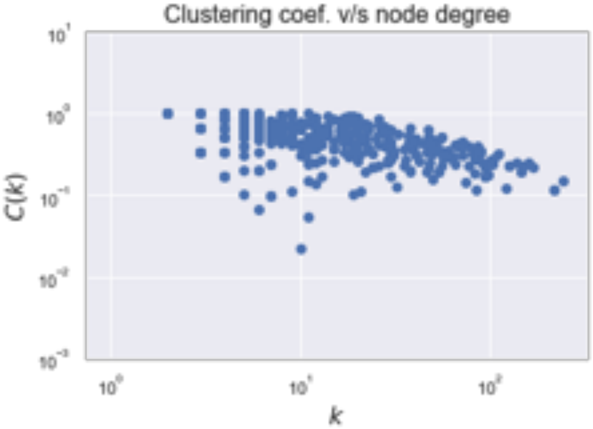
Plotting the degree distribution on a logarithm scale, we can see that some few HUBs (airport with high number of connections) appears and the pattern follows a "Power Law", for great part of distribution (k > 200). As stated by Massimiliano Zanin on [6] "Modeling the Air Transport with Complex Networks: a short review", page 7, "One important aspect of flight networks is the fact that they show the scale-free feature. This implies the presence of few hubs with a very high number of connections, confirming the predominance of a hub-and-spoke topology". This can suggest that the Air Network can be considered a "Scale-Free Small-World Network.

Also, Barabasi on [1] ”Network Science", Chapter 4, page 12, says: "A scale-free network looks like the air-traffic network, whose nodes are airports and links are the direct flights between them. Most airports are tiny, with only a few flights. Yet, we have a few very large airports, like Chicago or Los Angeles, that act as major hubs, connecting many smaller airports. Once hubs are present, they change the way we navigate the network. For example, if we travel from Boston to Los Angeles by car, we must drive through many cities. On the airplane network, however, we can reach most destinations via a single hub, like Chicago."

Networking Analysis



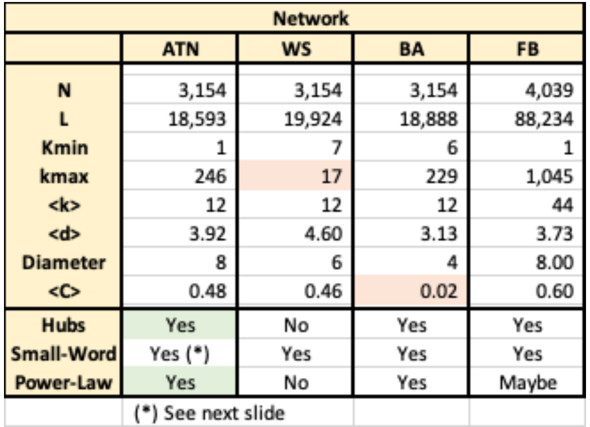
More than 1/3 of all routes has 3 or less legs, being 4 the average.



The clustering coefficient C, is defined as the probability that two cities that are directly connected to a third city also are directly connected to each other. We find that C is typically larger for the air transportation network than for a random graph and that it decreases with K size, but slower than d, for example. These results are consistent with the expectations for a small-world network but not with those for a random graph. For the Air Worldwide Network, <C> is 0.48. As also discussed by R. Guimera on [4] "The worldwide air transportation network: Anomalous centrality, community structure, and cities global roles", page 2, "Therefore we conclude that the air transportation network is, as expected, a small-world network".

Networking Analysis

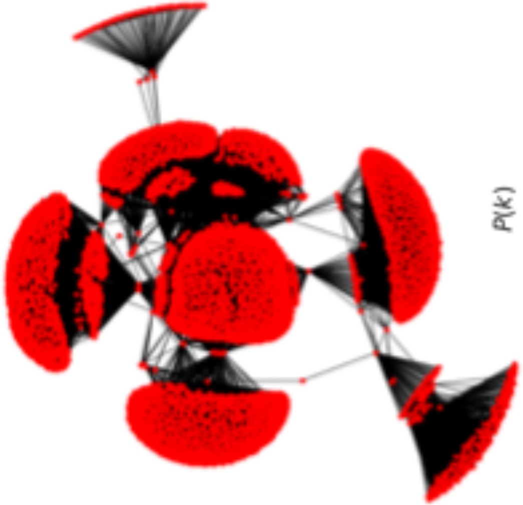
Comparing models

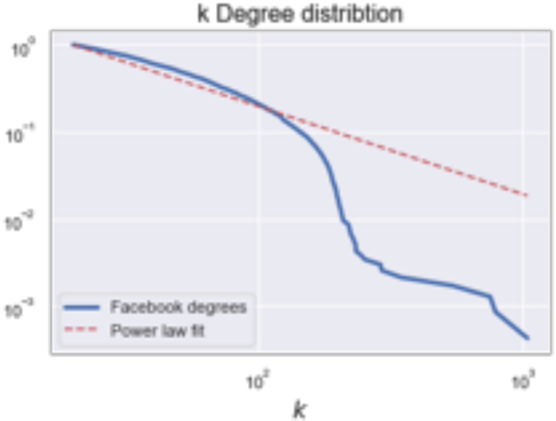


	Network			
	ATN	WS	BA	FB
N	3,154	3,154	3,154	4,039
L	18,593	19,924	18,888	88,234
Kmin	1	7	6	1
kmax	246	17	229	1,045
<k>	12	12	12	44
<d>	3.92	4.60	3.13	3.73
Diameter	8	6	4	8.00
<C>	0.48	0.46	0.02	0.60
Hubs	Yes	No	Yes	Yes
Small-Word	Yes (*)	Yes	Yes	Yes
Power-Law	Yes	No	Yes	Maybe
	(*) See next slide			

The BA model is better than the WS model at reproducing the degree distribution (Power-Law). Also the average path length,<d>, is 3.13, which is even more “small world” than the actual network, which has <d> = 3.92. On the other hand, the clustering coefficient, <C>, is 0.02, not even close to the value in the dataset, 0.48. Concluding, The WS model captures the Small-World characteristics, but not the degree distribution and the BA model captures the degree distribution and the average path length, but not C

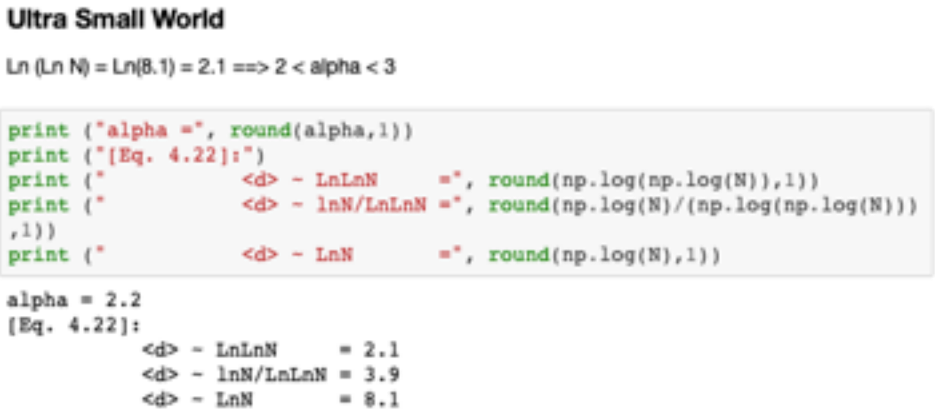
On [2] "Think Complexity", Allen B. Downey performed same analysis with a different dataset, a set of Facebook users and their friends from the Stanford Network Analysis Project (SNAP). The result found by Downey is very close with what we found on the WorldWide Air Transport Network and suggests that this network is really a "Free-Scale Small World Network.





Networking Analysis

Ultra-Small-World Network



Ultra Small World

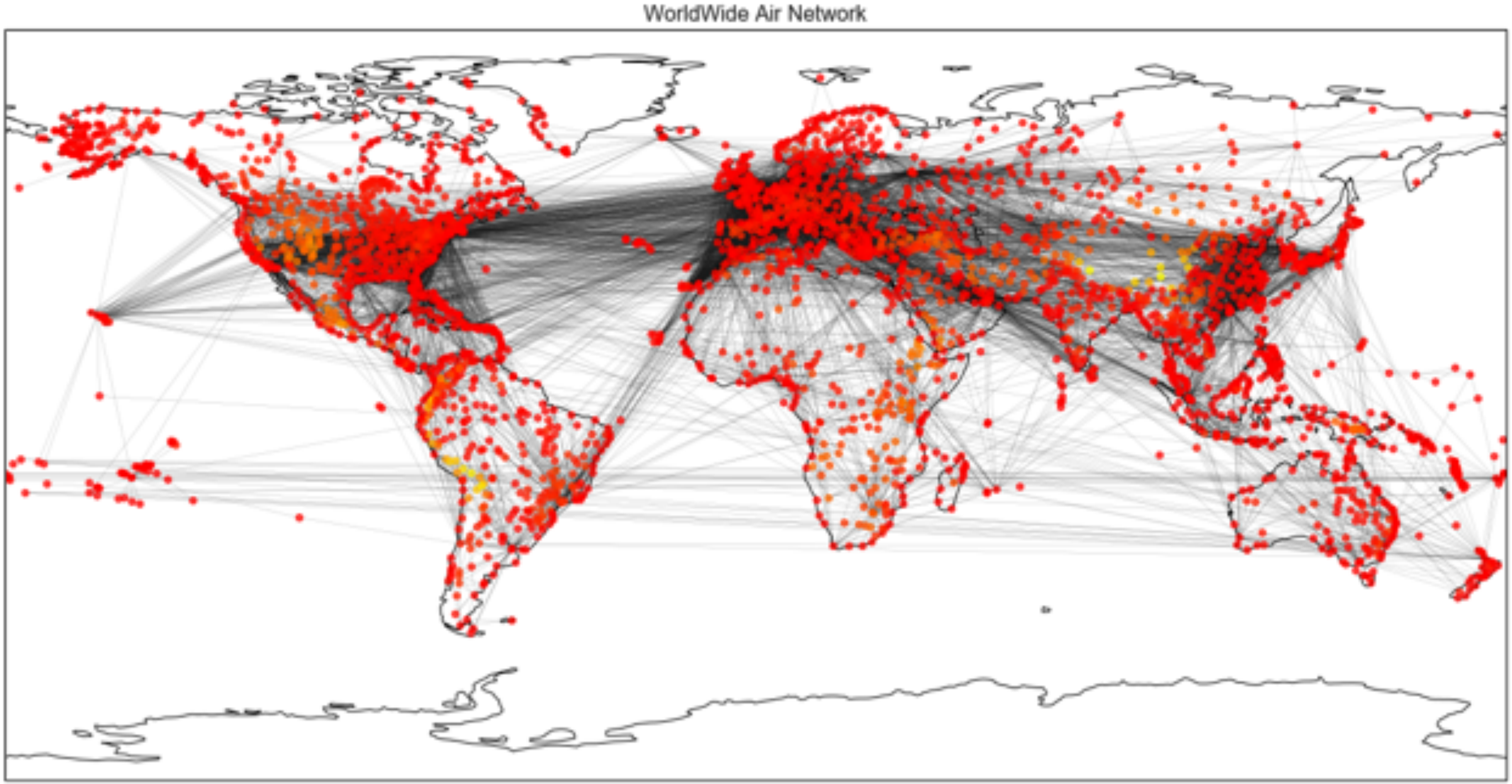
Ln (Ln N) = Ln(8.1) = 2.1 ==> 2 < alpha < 3

```
print ("alpha =", round(alpha,1))
print ("[Eq. 4.22]:")
print ("            <d> ~ LnLnN     =", round(np.log(np.log(N)),1))
print ("            <d> ~ lnN/LnLnN =", round(np.log(N)/(np.log(np.log(N)))
,1))
print ("            <d> ~ LnN       =", round(np.log(N),1))
```

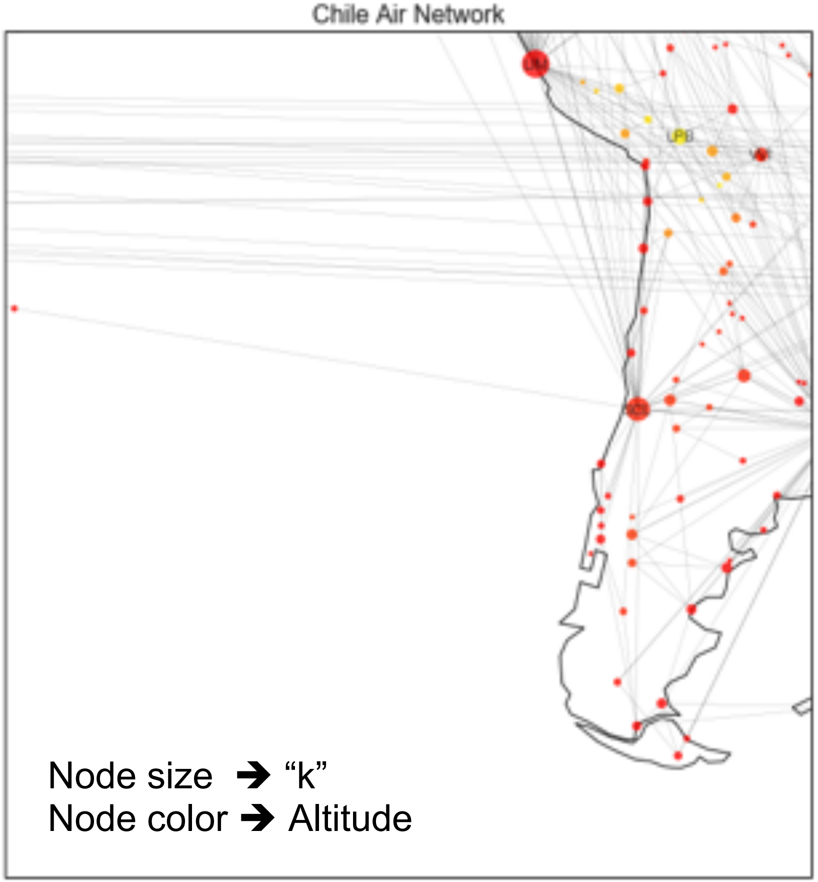
```
alpha = 2.2
[Eq. 4.22]:
            <d> ~ LnLnN     = 2.1
            <d> ~ lnN/LnLnN = 3.9
            <d> ~ LnN       = 8.1
```

According Barabasi, the presence of hubs in scale-free networks affect the small world property. "Airlines build hubs precisely to decrease the number of hops (stops) between two airports. The calculations support this expectation, finding that distances in a scale-free network are smaller than the distances observed in an equivalent random network.” According him, the average distance increases as lnlnN, a significantly slower growth than the lnN derived for random networks. He calls networks in this regime ultra-small, as the hubs radically reduce the path length. They do so by linking to a large number of small-degree nodes, creating short distances between them."

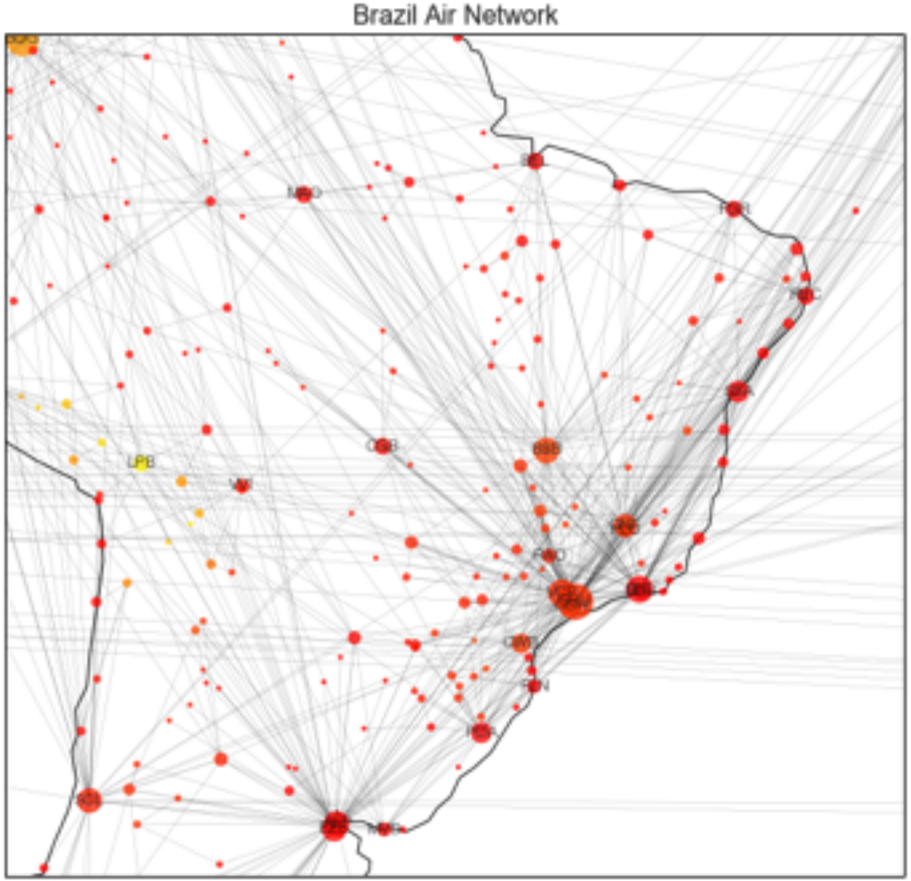
“Geo mapping” the Air Network



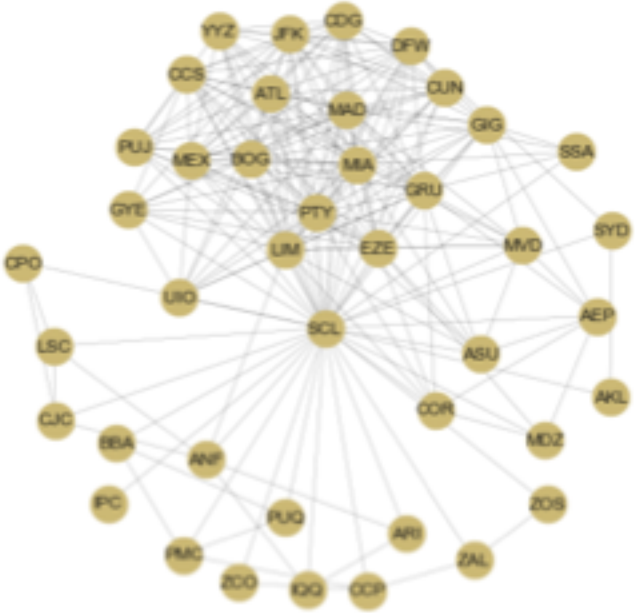
“Geo mapping” the Air Network

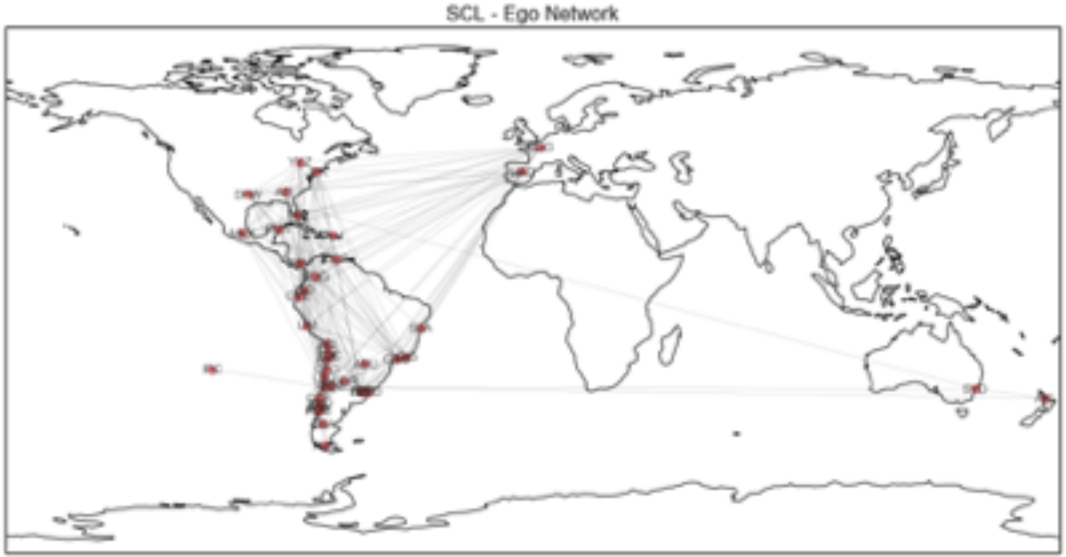


Node size ➔ “k”
Node color ➔ Altitude



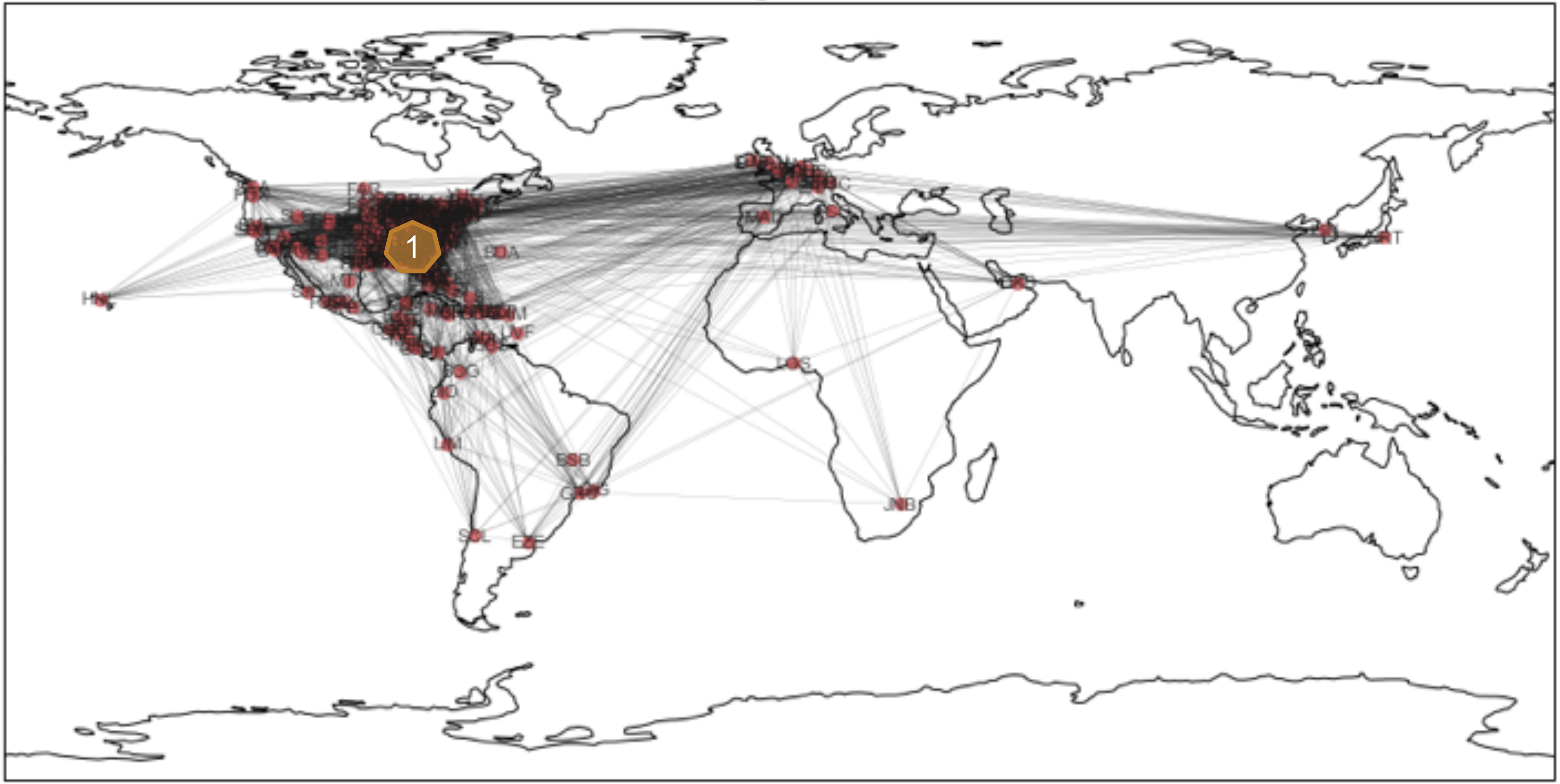
“Geo mapping” the Air Network – “Ego”



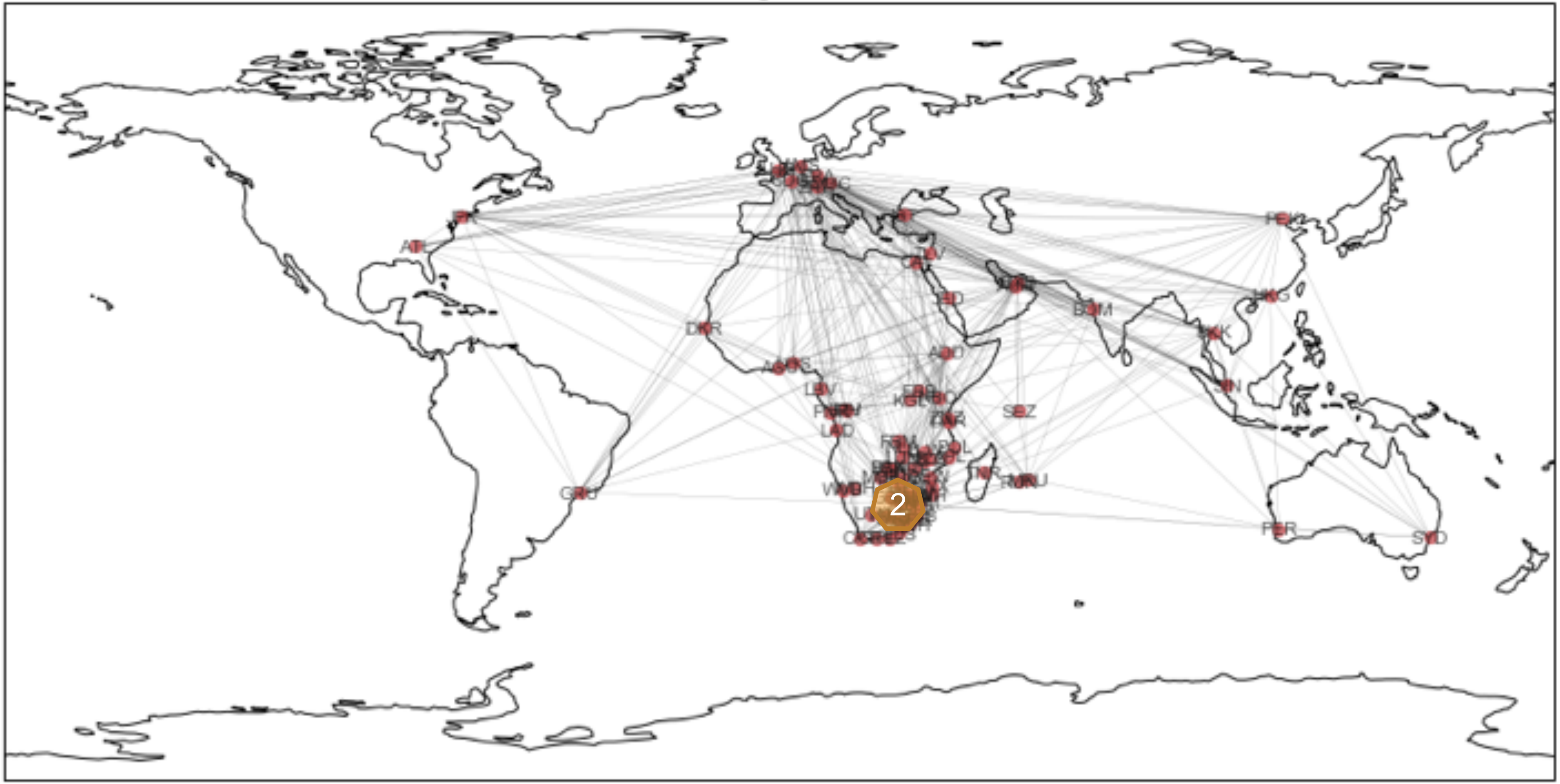

SCL - Ego Network

Ego networks consist of a focal node ("ego") and the nodes to whom ego is directly connected to (these are called "alters") plus the ties, if any, among the alters. This is a very important tool on Air Transport Network analysis, for example to study how human decease can spread along the network

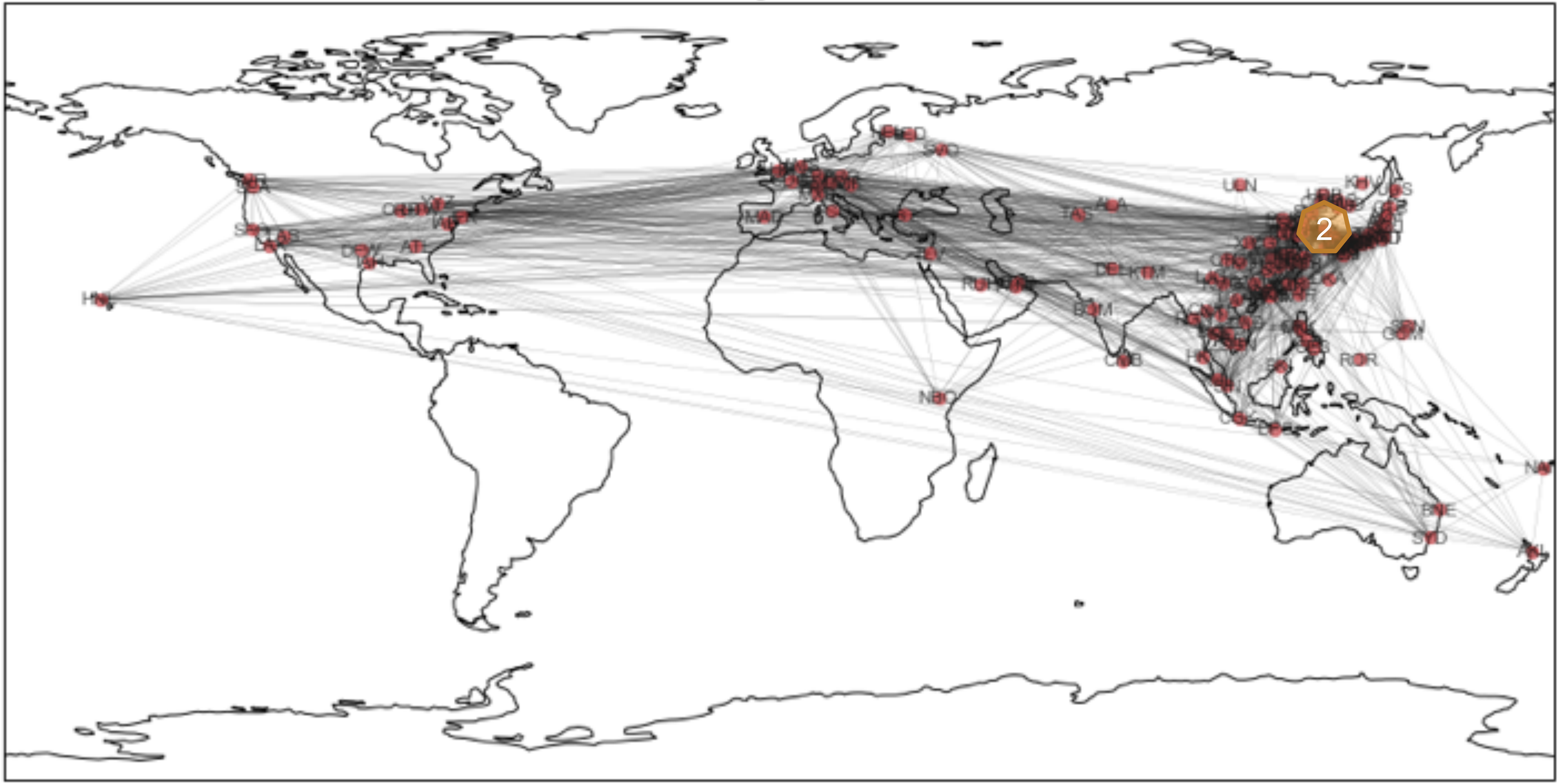
Decease outbreak, having Atlanta as starting point



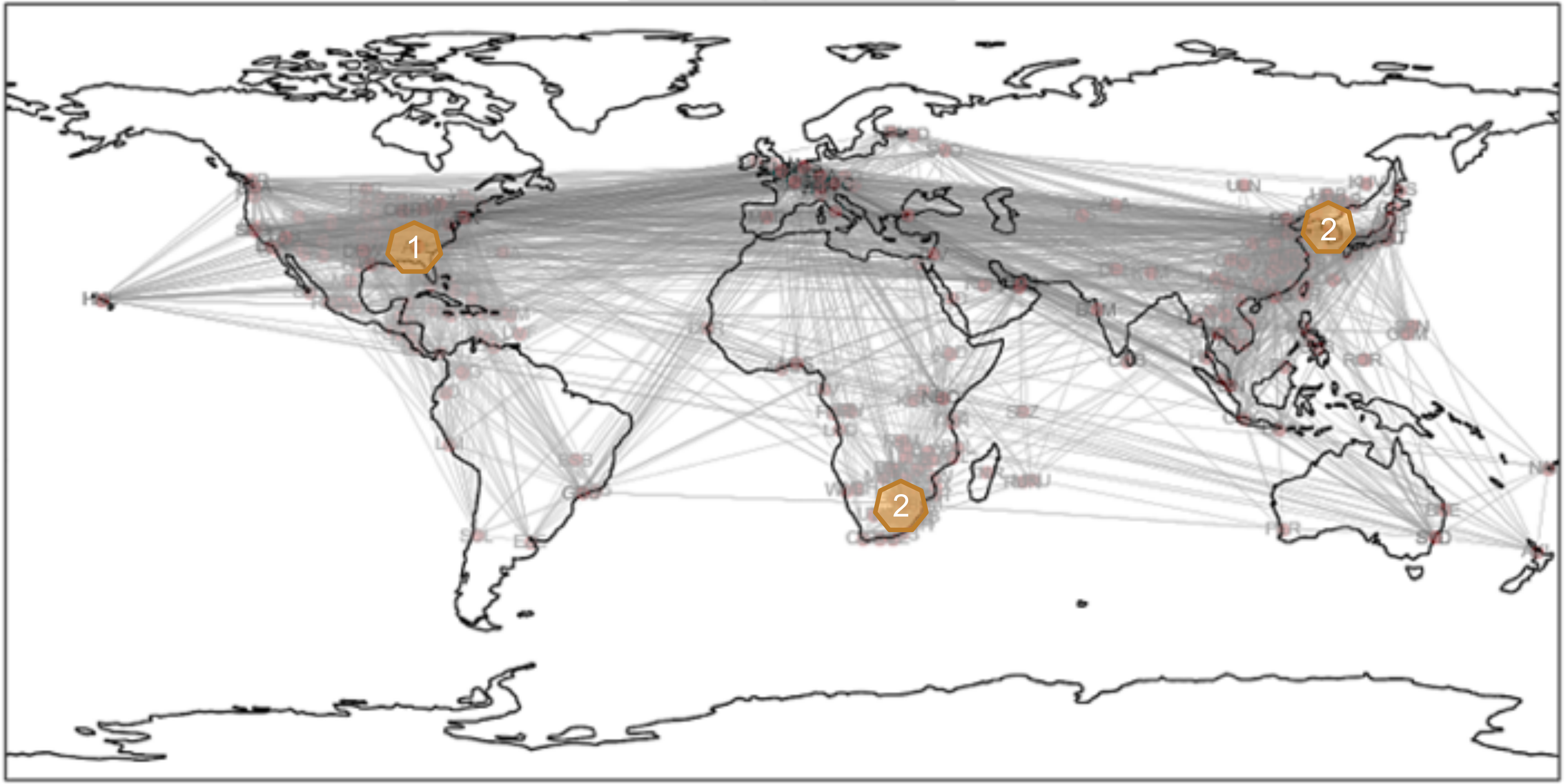
Decease outbreak, having Atlanta as starting point and Johannesburg as next stop



Decease outbreak, having Atlanta as starting point and Seoul as next stop



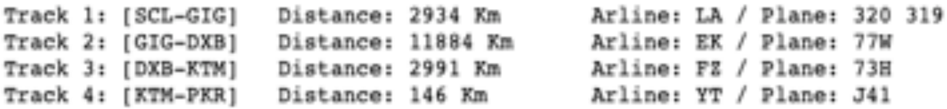
JNB ← ATL → ICN



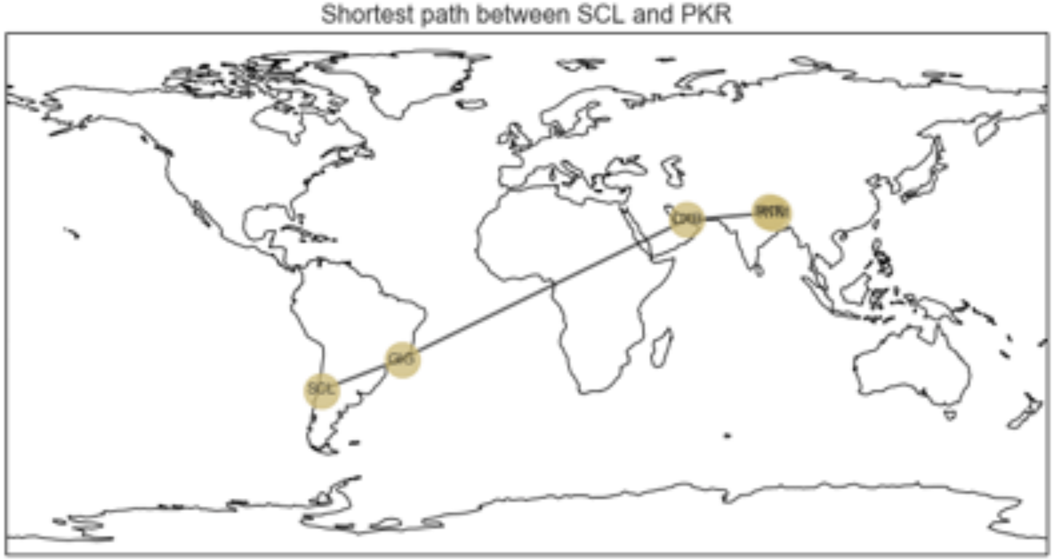
Calculating the best route

Using Dijkstra’s algorithm

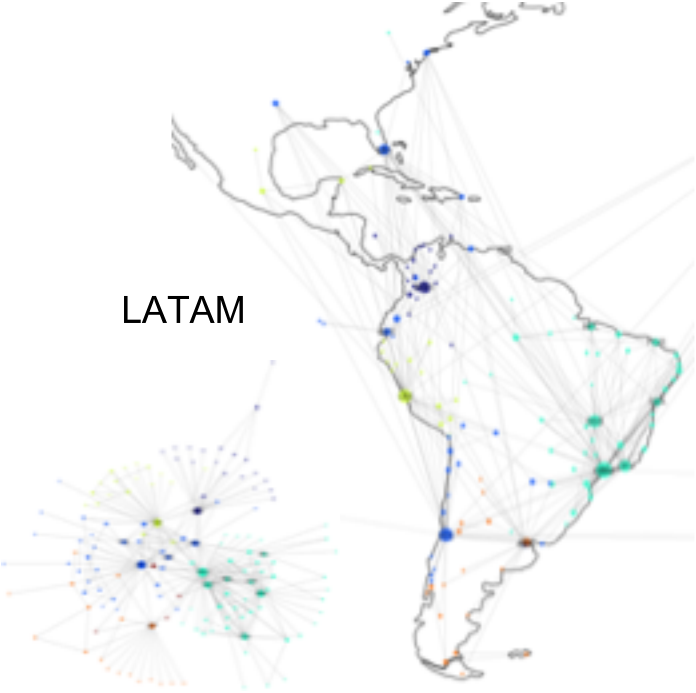
Shortest Path from Santiago, Chile to Pokhara, Nepal



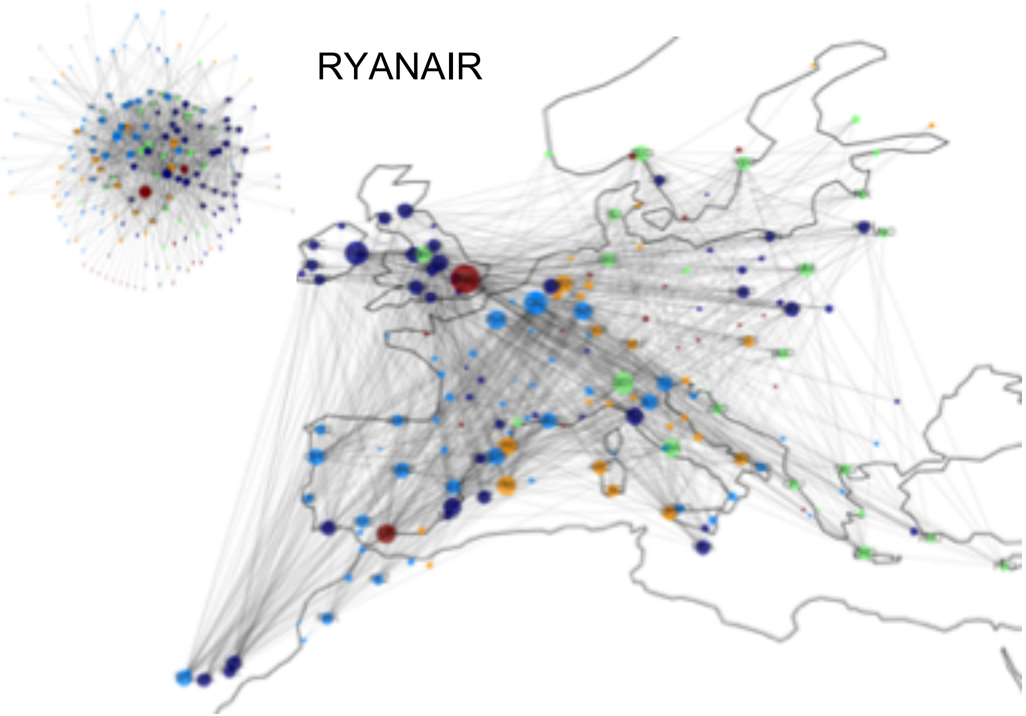
```
Track 1: [SCL-GIG]   Distance: 2934 Km       Arline: LA / Plane: 320 319
Track 2: [GIG-DXB]   Distance: 11884 Km      Arline: EK / Plane: 77W
Track 3: [DXB-KTM]   Distance: 2991 Km       Arline: FZ / Plane: 73H
Track 4: [KTM-PKR]   Distance: 146 Km        Arline: YT / Plane: J41
```



Comparing Air Companies Topologies



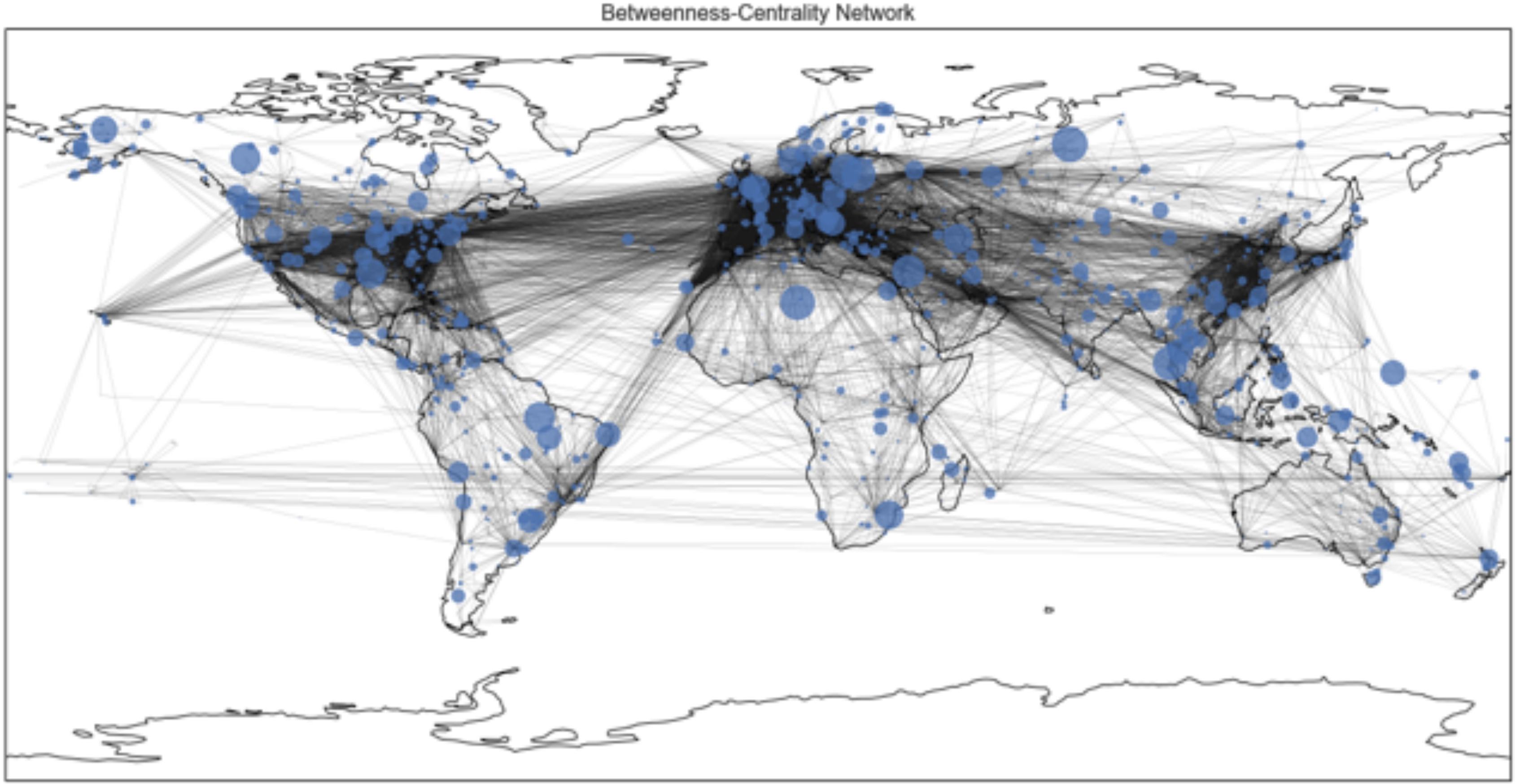
LATAM



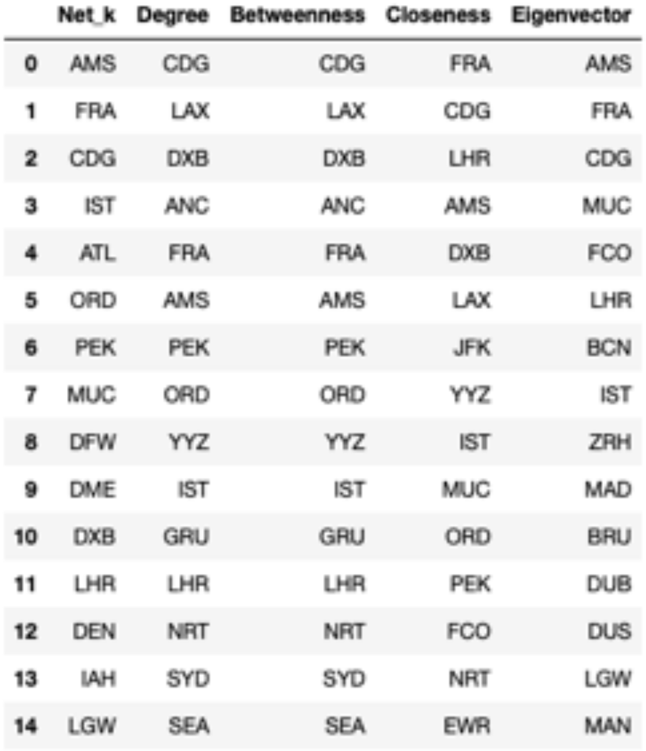
RYANAIR

Looking on both air companies network distribution, we can verify what is described on [6], that the structure of of low-cost airlines, as Ryanair, the biggest European low-cost company, has a densely connected core, in opposition to companies as LATAM or other big companies.

Centrality



Centrality



	Net_k	Degree	Betweenness	Closeness	Eigenvector
0	AMS	CDG	CDG	FRA	AMS
1	FRA	LAX	LAX	CDG	FRA
2	CDG	DXB	DXB	LHR	CDG
3	IST	ANC	ANC	AMS	MUC
4	ATL	FRA	FRA	DXB	FCO
5	ORD	AMS	AMS	LAX	LHR
6	PEK	PEK	PEK	JFK	BCN
7	MUC	ORD	ORD	YYZ	IST
8	DFW	YYZ	YYZ	IST	ZRH
9	DME	IST	IST	MUC	MAD
10	DXB	GRU	GRU	ORD	BRU
11	LHR	LHR	LHR	PEK	DUB
12	DEN	NRT	NRT	FCO	DUS
13	IAH	SYD	SYD	NRT	LGW
14	LGW	SEA	SEA	EWR	MAN

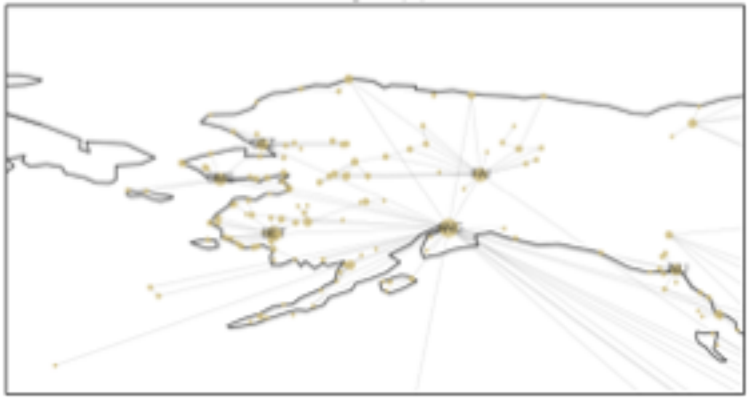
We can see that the criteria used to calculated the "most centralized airports", have high influence in the order and also the results are not in the same order that the ones with bigger "k" ("Net_k"). Note that exist cases where some airports are very "centralized", but low connected. This can be considered an anomaly, as explained by R. Guimera on [4] "The worldwide air transportation network: Anomalous centrality, community structure, and cities global roles”:

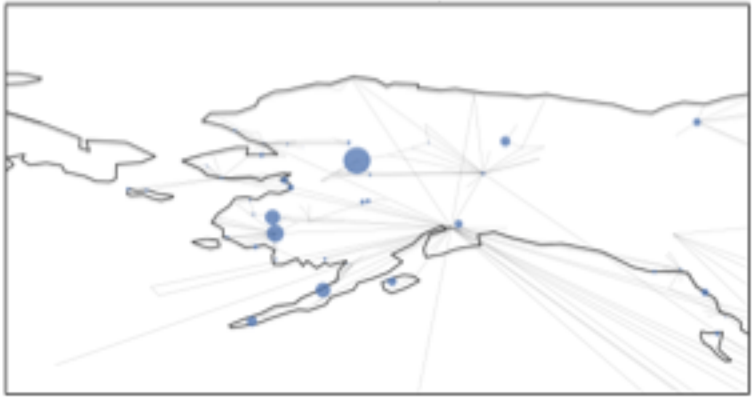
Focusing on "Betweenness", the relevant question is, however, "what general and plausible mechanism would give rise to scale-free networks with the obtained anomalous distribution of betweenness centralities?”

To answer this question, it is useful to consider a region such as Alaska. For that, let’s see their maps in detail:

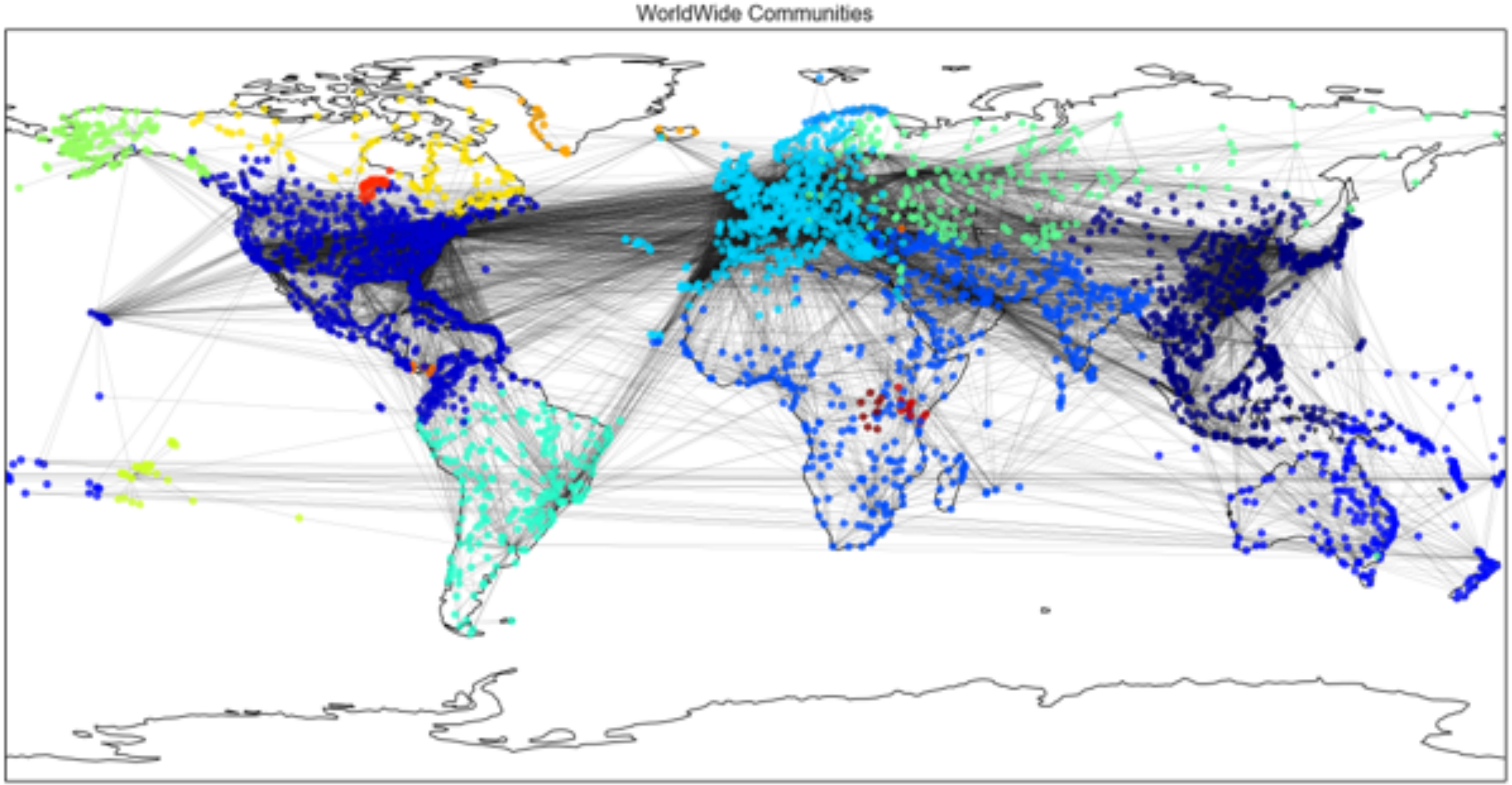
Centrality

Alaska is a sparsely populated, isolated region with a disproportionately large, for its population size, number of airports. Most Alaskan airports have connections only to other Alaskan airports. This fact makes sense geographically. However, distance-wise, it also would make sense for some Alaskan airports to be connected to airports in Canada's Northern Territories, but these connections are, however, absent. Instead, a few Alaskan airports, singularly Anchorage, are heavily connected to the continental U.S. The reason is clear: the Alaskan population needs to be connected to the political centers, which are located in the continental U.S., whereas there are political constraints making it difficult to have connections to cities in Canada, even to ones that are close geographically. It is now obvious why Anchorage's centrality is so large. Indeed, the existence of nodes with anomalous centrality is related to the existence of regions with a high density of airports but few connections to the outside. The degree-betweenness anomaly is therefore ultimately related to the existence of communities in the network. This can be verified on the map where Alaska's region is detailed:



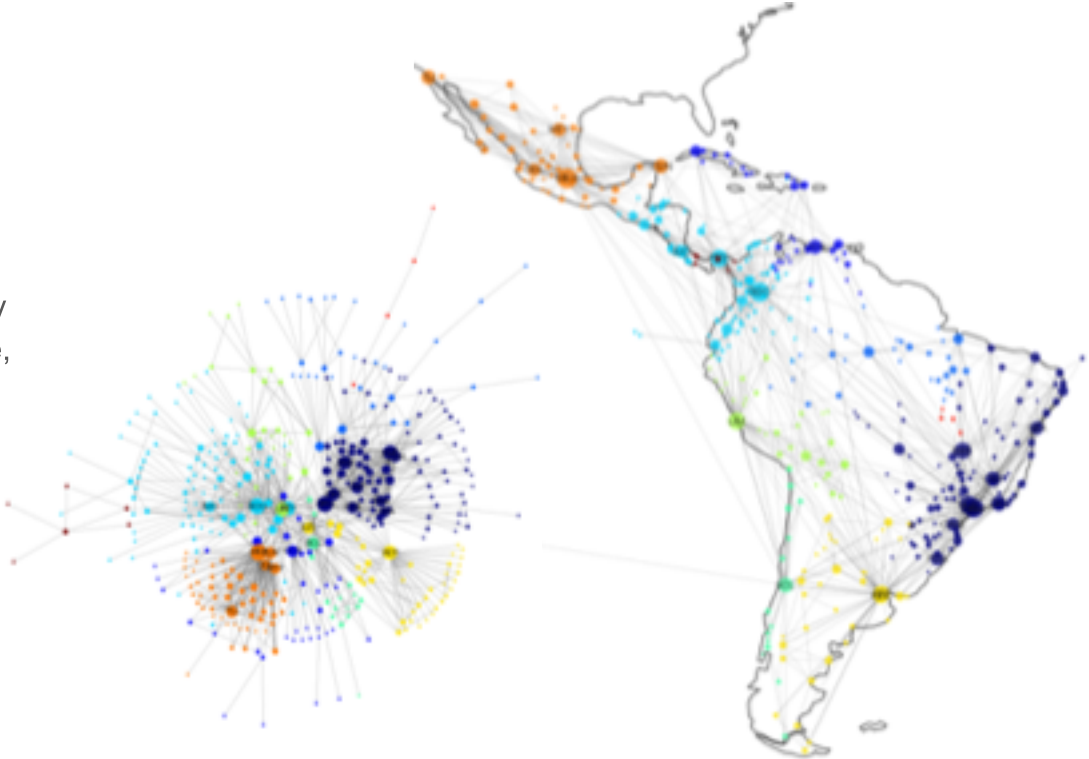


Community detection



Community detection

On previous map is possible to confirm that the airports are connected in "communities" with some geographic correlation, but not only this. Another significant result is that even though geographical distance plays a clear role in the definition of the communities, the composition of some of the communities cannot be explained by purely geographical considerations. For example, the community that contains most cities in Europe also contains most airports in Asian Russia. Similarly, Japanese cities are mostly grouped with cities in the other countries in Oceania, but India is mostly grouped with the Arabic Peninsula countries and with countries in Northeastern Africa. These facts are consistent with the important role of political factors in determining community structure, as we can see on the Latin America case (Cuba & Venezuela):



Conclusion

On this project, the main idea, was to analyze the structure of the world-wide air transportation network, understanding how we could define the best possible routes between two destinations, not only in economic terms, but also to prevent the spread of human deceases.

The study enables us to unveil a number of significant results. The worldwide air transportation network is a "Small-World" network (*), similar to social networks as Facebook, in which (i) the number of nonstop connections from a given city and (ii) the number of shortest paths going through a given city have distributions that are scale-free. Surprisingly, the nodes with more connections are not always the most central in the network. We hypothesize that the origin of such a behavior is the metacommunity structure of the network. We find the communities in the network and demonstrate that their structure can only be understood in terms of both geographical and political considerations.

(*) According Barabasi, the presence of hubs in scale-free networks affect the small world property. "Airlines build hubs precisely to decrease the number of hops (stops) between two airports". Barabasi calls networks in this regime "Ultra-Small, as the hubs radically reduce the path length.

Resources

1. Networking Science
2. Think Complexity
3. IPython Cookbook, Second Edition
4. The worldwide air transportation network
5. Revealing the structure of the world airline network
6. Modelling the Air Transport with Complex Networks: a short review
7. Reliability Analysis for Aviation Airline Network Based on Complex Network
8. Finding The Shortest Path, With A Little Help From Dijkstra
9. An Introduction to Graph Theory and Network Analysis
10. Catching that flight: Visualizing social network with Networkx and Basemap
11. The Hidden Geometry of Complex, Network-Driven Contagion Phenomena

Universidad del Desarrollo

Thank you

Marcelo José Rovai

Santiago, Chile

December 29th 2018