TREBALL DE FI DE CARRERA

TÍTOL DEL TFC: “Un model per a la xarxa global d’aeroports basat en el grau d’intermediació d’enllaços.”

TITULACIÓ: Enginyeria Tècnica Aeronàutica, especialitat Aeronavegació

AUTOR: Francesc Segura Revilla

DIRECTOR: Francesc Comelles, Alicia Miralles

DATA: 25 de juliol de 2011
Títol: “Un model per a la xarxa global d'aeroports basat en el grau d'intermediació d'enllaços.”

Autor: Francesc Segura Revilla

Director: Francesc Comellas, Alicia Miralles

Data: 25 de juliol de 2011

Resum

Moltes xarxes reals com la World Wide Web, la xarxa telefònica, l'elèctrica o sistemes de transport (incloent les rutes aèries), biològics o socials pertanyen a una categoria que s'anomena "món petit invariants d'escala" (small-world scale-free). Les característiques d'aquestes xarxes són una gran concentració local de nodes o clustering (els nodes tenen molts veïns comuns) i al mateix temps un diàmetre petit (màxima distancia entre qualsevol parella de nodes) i també el fet que el nombre d'enllaços dels nodes segueix una llei potencial. L'objectiu d'aquest treball és estudiar les característiques de la xarxa global d'aeroports i modelar la xarxa a través d'un procés d'optimització d'una malla amb un nombre similar de nodes i enllaços. Per a realitzar l'estudi farem servir el programa Python en la seva versió 2.7 i el paquet NetworkX (v 1.4) especialitzat en l’estudi de xarxes. En la primera part d'aquest TFC analitzem les propietats de la xarxa d'aeroports i verifiquem que es tracta d'una xarxa món petit invariant d'escala. A la segona part del TFC considerem models que s'obtenen a partir d'una malla toroïdal canviant aleatòriament els seus enllaços a mesura que s'optimitza una certa funció de cost (basada en la distància i el tràfic). El resultat de les proves efectuades han estat xarxes semblants a les aleatòries i amb una distribució d'enllaços que no es correspon a l'esperada. Com a conclusió podem dir que modelar la xarxa d'aeroports requereix un estudi més complex amb una funció de cost segurament més complicada i la consideració d'altres paràmetres que permetin reflectir el procés d'evolució que ha seguit la xarxa original.
Overview

Many real networks as the World Wide Web, the telephone and electrical networks and transportation (including air routes), biological or social systems belong to a category called “small-world scale-free”. The features of these networks are a huge local node concentration, also called clustering (this means that the nodes have lots of common neighbours) and at the same time they have a small diameter (the maximum distance between two pair of nodes); and also the fact that the number of links in a node follows a power law.

In this TFC we study the characteristics of the global network of airports and we model this network through an optimization process starting from a lattice network with a similar number of nodes and links. For this study we use Python (version 2.7) and the package NetworkX (v 1.4) designed for the study of networks. In the first part of this TFC we analyze the properties of the airports network and verify that it is a small world scale free network. In the second part we consider models obtained from a toroidal grid by randomly changing links while optimizing a certain cost function (based on distance and traffic). The results of the tests show that the networks obtained do not follow the expected degree distribution and are similar to random networks.

In conclusion we can say that modeling the airports network requires a more complex study with a more elaborated cost function and the inclusion of other parameters accounting for the evolution processes that has followed the original network.
INDEX

INTRODUCTION .................................................................................................................. 1

CHAPTER 1. GRAPH THEORY .......................................................................................... 2
1.1. What is a graph? .................................................................................................. 2
  1.1.1. Graph types .............................................................................................. 3
1.2. Basic properties ................................................................................................. 4
  1.2.1. Diameter .................................................................................................. 5
  1.2.2. Mean distance ......................................................................................... 5
  1.2.3. Clustering coefficient ............................................................................. 5
  1.2.4. Centrality .............................................................................................. 7
  1.2.5. Degree distribution .............................................................................. 9
1.3. Random network models .................................................................................. 10
  1.3.1. The Erdős–Rényi model ..................................................................... 10
  1.3.2. The Barabási-Albert model and its variants .......................................... 12
  1.3.3. The Watts-Strogatz model ................................................................. 14
1.4. Complex Networks ............................................................................................ 15
  1.4.1. Real-world Networks .......................................................................... 15
  1.4.2. Properties of real-world Networks ...................................................... 18
1.5. Optimization algorithms .................................................................................... 21
  1.5.1. Threshold method .............................................................................. 21
  1.5.2. Simulated annealing (SA) .................................................................. 21

CHAPTER 2. GLOBAL AIRLINE NETWORK ANALYSIS ............................................. 23
2.1. Basic properties .................................................................................................. 23
  2.1.1. Asia and Middle East airline network .................................................. 23
  2.1.2. North America airline network ............................................................ 25
  2.1.3. Global airline network ....................................................................... 28
2.2. Simulation of the global airline network .......................................................... 30
  2.2.1. Generation of the model ....................................................................... 31
  2.2.2. Cost function ......................................................................................... 32
  2.2.3. Optimization using the threshold method ............................................ 33
  2.2.4. Optimization algorithm using simulated annealing ............................. 37
  2.2.5. Add, remove and change with a given probability ............................... 39
  2.2.6. Conclusions ......................................................................................... 41
2.3. Bibliography and references .............................................................................. 42
INTRODUCTION

Networks are a very important part in the actual world that surrounds us, from the router connections and the World Wide Web until the roads of a country, the global airline network and the electrical power grid. We can go away from the technological field and focus on the social one and also we can discover networks, such as the friendship between people (a network strongly exploited by companies as Facebook), the working relations and the common hobbies. There are also biological networks such as the food chain and the interaction between proteins in our bodies. The physical world is also rich in networks such as interactions between atoms in the matter. Graphs are used for describing mathematical concepts in networks. This tool analyzes the networks treating them as a collection of edges and nodes. For example in the airline network, the airports are the nodes and the connections between them are the edges. This method is so powerful because it can treat different networks with the same mathematical principles and show if their features are similar.

All the networks have special attributes and the graph theory lets us study this features for achieve an optimal network design. There are important factors such as the cost of the different ways to go from one node to another and the load on every node that will determine if a network is optimal for its purposes and optimal as a network itself, for example, if that net can keep on working if some nodes fail.

The first part of this work will be dedicated to the explanation of the easiest concepts of the graph theory in order to trying to make understandable the second part, where we will analyze three different airline networks, the Asia and Middle East network, the North America network and finally the global airline network. After this, in the second part we will try to simulate a network with similar features to some airline network using the Python package Networkx and some different solution methods to see which one is the optimal.
CHAPTER 1. GRAPH THEORY

Graph theory is the most important tool to study the topology, the behavior and the properties of every complex network. Every network in the real life can be represented by a bunch of nodes and edges; for example, in the Internet world every website has links to other ones, the websites in a graph would be the nodes and the links between them would be the edges; in the global airline network the nodes would represent the airports and the flights between them would be represented by the edges of the graph.

The beginning of the graph theory starts in 1736 with the paper written by Leonard Euler on the Seven Bridges of Königsberg. This is a historically notable problem that tried to find if it was possible to walk through the city only crossing its seven bridges once and only once. Euler proved himself that the problem has no solution. This problem shows how the graph theory was developed in the early days, it focused on small nets and tried to study the nodes and the edges in an isolated way. Nowadays, due to the incredible technological advances, the networks studied are larger and they are studied in a more global way to find its properties and behavior. These large networks are important because they are the ones what are found on the real life and could have hundreds and thousands nodes and edges. The study of these graphs in a mathematical or graphical way could be hard and graph theory has moved to statistical ways for achieve the comprehension of its properties. So, in addition to know which node is the better connected or which is the shortest path to go from node A to node B, it is also important to know which is the parameter that measures better the graph connectivity in overall and how the graph will be affected if we replace some edges for other ones.

In order to explain these concepts, this chapter will try to define what a graph is and its types, the basic concepts of the graph theory and the statistical properties that build their characteristics.

1.1. What is a graph?

A graph is an abstract representation of a set of objects where some pairs are connected by links. The interconnected objects are represented by mathematical abstractions called vertices or nodes and the links that connect that nodes are called edges. Visually a graph is pictured as a bunch of dots that represent the nodes, joined by lines that represent the edges. Mathematically, a graph \( G = (V, E) \) is a pair of sets \( V \) and \( E \), where \( V \) is a set of vertices (the nodes of the graph), and \( E \) is a set of edges, that denote the link between the vertices.

Here we have a little glossary of common used terms:
• Two vertices are *adjacent* if there is an edge which links them.
• The *order* of a graph is the number of vertices it has: \( n = |V| \).
• The *size* of the graph is the number of edges it has: \( m = |E| \).
• The *degree* \( k_i \) of a vertex \( i \) is the number of edges that are adjacent to it. The *degree* \( \Delta \) of a graph is the maximum degree of all of its vertices: \( \Delta = \max(k_i) \).
• A *path* is the succession of edges that link two vertices of a same graph. The *length* of the path is defined by the number of edges that this path crosses.
• A *geodesic path* between a couple vertices is the shortest path that links them, and it is due to definition the *distance* between that couple vertices. It can be more than one geodesic path between two nodes.
• A *giant component* of a graph is a connected subgraph that contains majority of the entire graph’s nodes.

### 1.1.1. Graph types

#### 1.1.1.1. Graph types given its special distribution of nodes or edges

We can have special graphs given its vertex and edge distributions. In the first place we find the definition of a regular graph. In this graph each vertex has the same number of neighbors, what means that every vertex has the same degree; mathematically it is called \( \Delta\text{-regular} \), where the degree of all of its nodes is \( \Delta \). We can also find a complete graph, named as \( K_d \), where \( d \) is the order of the nodes, what is a graph that each vertex is connected with an edge to all the others. Also a graph can be finite and infinite, this means that a graph with \( G=(V, E) \) such that \( V \) and \( E \) are finite, is a finite graph, otherwise it will be infinite. Most commonly in graph theory it is implied that the graphs discussed are finite.
Graphs can also be classified in terms of connectivity. Two vertices $u$ and $v$ are called connected if $G$ contains a path from one to the other. Otherwise, they are called disconnected. A graph is connected if every pair of distinct vertices in the graph are connected; otherwise, it is called disconnected. A graph is called $k$-vertex-connected ($k$-edge-connected) if no set of $k-1$ vertices (edges) exists that, when removed, disconnects the graph. For example a directed graph is called weakly connected, but if replacing all of its directed edges with undirected edges it will produce a strongly connected graph.

1.2. Basic properties

Now we will define all the main properties that we need to use in order to improve the graph characteristics. These terms will be the diameter of the graph, the mean distance of the graph and of the node, the clustering coefficient, the centrality measurements and the degree distribution.
1.2.1. Diameter

The distance between two vertices $i$ and $j$ of a graph, $d(i,j)$, is defined as the number of edges that the shortest path between these vertices has.

The diameter of a graph is the maximum eccentricity of a graph. The eccentricity of a vertex, $v$, is the maximum distance between that vertex $v$ to all other vertices in $G$. The radius is the minimum eccentricity of the graph.

1.2.2. Mean distance

The mean distance of a graph is defined as the average of the distance for all pairs of vertices:

$$\bar{d} = \frac{1}{n(n-1)} \sum_{i,j \in V} d(i,j)$$ (1.1)

1.2.3. Clustering coefficient

The clustering coefficient of a graph measures the degree of vertices that are more connected between them than to the others. The usual definition of clustering is related to the number of triangles in the network. The clustering is high if two vertices sharing a neighbor have a high probability of being connected to each other. For a graph, clustering coefficient values near to 0 shows that many vertices are adjacent to another one but not among them, and...
values next to 1 show that there is a connection between almost every pair of vertices.

There are two common definitions of clustering. The first is global,

\[ C = \frac{3 \times \text{the number of triangles in the network}}{\text{the number of connected triples of vertices}} \]  (1.2)

where a “connected triple” means a single vertex with edges running to an unordered pair of other vertices.

A second definition for clustering is based on the average of the clustering for single nodes. The clustering for a single node is the fraction of pairs of its linked neighbors (triangles) out of the total number of pairs of its neighbors (triples).

\[ C_i = \frac{\text{the number of triangles connected to vertex } i}{\text{the number of triples centered on vertex } i} \]  (1.3)

For vertices with degree 0 or 1, for which both numerator and denominator are zero, we use \( C_i = 0 \). Then the clustering coefficient for the whole network is the average:

\[ C = \frac{1}{n} \sum C_i \]  (1.4)

In both cases the clustering is in the range \( 0 \leq C \leq 1 \).

Now we will see an easy example of how it the clustering coefficient works. The network in (Fig. 1.4) has one triangle and eight connected triples, and therefore has a clustering coefficient \( 3 \times 1 / 8 = 3 / 8 \) according to the global equation (1.2). The individual vertices have local clustering coefficients, Eq. 1.3, of 1, 1, \( \frac{1}{6} \), 0 and 0, for mean value, Eq. 1.4, of \( C = 13/30 \).

Fig. 1.4 Example of a simple network [8].
The last specification for the clustering coefficient is in directed graphs. In these graphs it is necessary to calculate clustering coefficient in the entrance and in the exit. The mechanism will be the same for both.

In a larger scale we can see a visual example:

Fig. 1.5 Two graphs both of which have 84 vertices and 358 edges: the graph on the left is a uniform random graph; we can see clearly that the clustering coefficient for the graph on the right will be larger than the coefficient for the randomly created graph [18].

1.2.4. Centrality

The centrality is the relative importance of a vertex within a graph (for example, how important a person is within a social network or how well-used a road is within an urban network).

There are different measures of centrality that are widely used in network analysis: degree centrality, betweenness, closeness, and eigenvector centrality. We will focus on the first two.

1.2.4.1. Degree centrality

The first, and simplest, is degree centrality. Degree centrality is defined as the number of edges incident upon a node over the total number of vertices. In directed graphs we usually define two separate measures of degree centrality, namely indegree and outdegree.

For a graph $G: = (V, E)$ with $n$ vertices, the degree centrality $C_D(v)$ for vertex $v$ is:
A model for the global airport network based on the intermediation degree of the edges

\[ C_D(v) = \frac{\deg(v)}{n-1} \]  

(1.5)

We can also define the degree centrality for a graph \( G \) as,

\[ C_D(G) = \frac{\sum_{i=1}^{n}[C_D(v^*) - C_D(v_i)]}{(n-1)(n-2)} \]  

(1.6)

where \( v^* \) is the node with the highest degree centrality in \( G \).

1.2.4.2. Betweenness centrality

This centrality measurement focuses on the number of shortest paths that crosses a vertex or an edge (there are both measurements).

For a graph \( G:=(V,E) \) with \( n \) vertices, the betweenness \( C_B(v) \) for vertex \( v \) is computed:

1. For each pair of vertices \( (i,j) \), compute all shortest paths between them.
2. For each pair of vertices \( (i,j) \) determine the fraction of shortest paths that pass through the vertex in question (here, vertex \( v \)).
3. Sum this fraction overall pair of vertices \( (i,j) \).

In a mathematical way:

\[ C_B(v) = \sum_{i \neq v \neq j \in V} \frac{\sigma_{ij}(v)}{\sigma_{ij}} \]  

(1.7)

where \( \sigma_{st} \) is the number of shortest paths from \( i \) to \( j \), and \( \sigma_{st}(v) \) is the number of shortest paths from \( i \) to \( j \) that pass through vertex \( v \).

To compute the betweenness centrality of an entire graph we have to normalize the values of the separated vertices, dividing through the number of pairs of vertices not including \( v \), which is \( (n - 1)(n - 2) \) for directed graphs and \( (n - 1)(n - 2)/2 \) for undirected graphs.
1.2.5. Degree distribution

The degree distribution of a graph is usually represented by the number of vertices (on the vertical edge) that have a $k$ degree (on the horizontal edge). The degree distribution can be described by a probability function $P(k)$ that shows the probability of choosing a random vertex it will have $k$ degree.

For a random graph, the degree distribution follows a Poisson distribution:

$$P(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!} \quad (1.10)$$

where $\lambda$ is the mean degree of the graph.

In the real networks with a high number of vertices we usually find a power law distribution. This distribution has a probability function:

$$P(k) = ck^{-\gamma} \quad (1.11)$$

where $\gamma$ is the slope of the line if we draw the edges in a logarithmic way:

$$\log(P(k)) = -\gamma \log(k) + \log(a) \quad (1.12)$$
The networks that present this kind of behavior are called scale-free networks; the most of the real networks present this distribution because in every real network are nodes that are more important than others and they have a high number of edges connected to them, but there are few of these.

1.3. Random network models

In the early days of the network study, the general believe was that networks where randomly created and since the 1960 several models for the generation of networks were developed, that models where important to study the properties of the networks and to compare them to the models of the real ones.

We can distinguish between two types of model. The first models take a bunch of nodes and we add edges between them following a given or desired probability, for example, we can compare these models to blind dates with lots of people or with singles reunions to meet new partners, we have a bunch of nodes that will represent the people and the creating edges will be the possible friendship engaged between them. The second models are growing network models, these models are based in the addition of more nodes that are connected to the two of the original ones. This is a model much more similar to the real life because it is more dynamic and real networks tend to behave in their expansion as this second type of models. For example, if we have the airport global network and a new airport is built, their flights will connect to other existing airports.

1.3.1. The Erdős–Rényi model

This is the easiest model and was developed by Paul Erdős and Alfréd Rényi in the 1960, and is used to generating random graphs using statistical information.
There are two graph ensembles or variants of this random graph model. The first one is the $G(n, M)$ model, a graph is chosen uniformly at random from the collection of graphs which have $n$ nodes and $M$ edges. The second one is the $G(n, p)$ model where the graph is constructed by connecting nodes randomly. Each edge is included in the graph with the probability $p$ independent from every other edge. Equivalently, all graphs with $n$ nodes and $M$ edges have equal probability of:

$$P(G(n, m)) = p^M (1 - p)^{(\binom{n}{2}) - M} \quad (1.13)$$

The expected number of edges in $G(n, p)$ model is $\binom{n}{2} p$ and any graph in $G(n, p)$ will almost surely have approximately this many edges. These two families are known to be similar if $M = \binom{n}{2} p$, so long as $p$ is not too close to 0 or 1, or when $pn^2$ tends to infinity.

The mean degree of this graph model is:

$$\bar{k} = \frac{2M}{n} = p(n - 1) \approx pn \quad (1.14)$$

The distribution of the degree of any particular vertex is:

$$P(deg(v) = k) = \binom{n - 1}{k} p^k (1 - p)^{n - 1 - k} \quad (1.15)$$

Fig. 1.10 Three generated networks by the model of Erdős and Rényi for different probabilities of the edge existence [21].

This network has a Poisson degree distribution.
1.3.2. The Barabási-Albert model and its variants

This is the easiest model for creating graphs with power laws as degree distribution and it was first applied by Derek de Solla Price in 1976, but its name comes from Albert-László Barabási and Réka Albert who rediscovered the process independently in 1999 and applied it to degree distributions on the web.

The Barabási-Albert model is based on two simple, but existing widely in real networks, assumptions regarding network evolution:

- **Growth**: new nodes are added to the network, where each new node is connected to \( m \) existing nodes. In the real life this means that nodes in the network increases over time.
- **Preferential attachment**: this is the heart of the model. Each new node is connected to existing nodes with a probability proportional to its existing degree. In plain words, the more connected a node is, the more likely it is to receive new links. Intuitively, the preferential attachment can be understood if we think in terms of social networks connecting people. Heavily linked nodes represent well-known people with lots of relations, which are represented by the edges. When a newcomer enters the community, he is more likely to become acquainted with one of those more visible people rather than with a relative unknown. Similarly, on the web, new pages link preferentially to hubs, for example very known sites such as Google or Youtube, rather than to pages that hardly anyone knows.

The network begins with an initial network of \( m_0 \) nodes, \( m_0 \geq 2 \) and the degree of each node in the initial network should be at least 1, and otherwise it will always remain disconnected from the rest of the network.

New nodes are added to the network one at a time. The probability of connecting to an existing node, \( i \), \( \Pi(i) \) is given by:

\[
\Pi(i) = \frac{k_i}{\sum_j k_j}
\]  

where \( k_i \) is the degree of node \( i \).

The initial core of the graph, which is frequently a few nodes, is usually assumed to be connected, but its structure has only a small effect of the final result.

Preferential attachment is an example of a positive feedback cycle where initially random variations (one node initially having more links or having started accumulating links earlier than another) are automatically reinforced, thus greatly, magnifying differences. Out of this field it is called also the Matthew effect, “the rich get richer”.

As it was said before, the degree distribution from the BA model is scale free, in particular, it is a power law of the form

\[ P(k) \sim k^{-3} \quad \text{(1.17)} \]

The average path length increases approximately logarithmically with the size of the network:

\[ l \sim \frac{\ln N}{\ln(\ln N)} \quad \text{(1.18)} \]

This model has a shorter average path length than the ER model but in the other hand, its clustering coefficient is five times higher than the one in a random graph with comparable degree and size and it decreases with the degree of the network with the proportion \( C \propto n^{-0.75} \).

![Graph showing degree distribution](image)

**Fig. 1.11** The degree distribution of the BA model, which follows a power law [24].

In this model we can study separately both basic assumptions. In the first case the growth is retained but not the preferential attachment. The resulting degree distribution in this limit is exponential, indicating that growth alone is not sufficient to produce a scale-free structure. The second case retains the preferential attachment but eliminates growth. The model begins with a fixed number of disconnected nodes and adds links, preferentially choosing high degree nodes as link destinations. In the early simulation the degree distribution looks scale-free but as the time goes by it becomes nearly Gaussian as the network reaches saturation, so preferential attachment is not sufficient to produce scale free-networks.
This model should be viewed as a simplification of reality, since it does not take into account many properties of evolving real-world networks. Due to this, the BA model has some variants to try to improve it to achieve a better matching to the real world.

### 1.3.3. The Watts-Strogatz model

The WS model is a random graph generation model that produces graphs with small-world properties, including average path lengths and high clustering. This model was proposed in 1998 by Duncan J. Watts and Steven Strogatz due to two important properties observed in many real-world networks that in the classic ER model are not present:

- They do not generate local clustering. Instead because they have a constant, random, and independent probability of two nodes being connected, ER graphs have a low clustering coefficient.
- They do not account for the formation of hubs. Formally, the degree distribution of ER graphs converges to a Poisson distribution, rather than a power law observed in many real-world scale-free networks.

The Watts and Strogatz model was designed as the simplest possible model that addresses the first of the two limitations. It takes into account the clustering coefficient retaining the short average path length of the ER model. It does so by interpolating between an ER graph and a regular ring lattice. Consequently, the model is able to at least partially explain the “small-world” phenomena in a variety of networks, such as the network of movie actors.

The major limitation of the model is that it produces an unrealistic degree distribution. In contrast, the real networks are often scale-free inhomogeneous in degree, having hubs and a scale-free degree distribution. Such networks are better described in that respect by the preferential attachment family models, such as the BA model. On the other hand, the BA model fails to produce the high levels of clustering seen in real networks.

![Diagram](image)

**Fig. 1.12** The variation of the basic circular lattice in order to increase the probability from the lowest value to the highest [30].
1.4. Complex Networks

During the late 1990s, the advancement of computers and their availability and power made possible to gather huge databases of network structures and to analyze them quickly and efficiently. This allowed, for the first time, the comparison of real network data with the existing models and to demonstrate that a lot of real networks that have been studied don’t have the random structure that we used to think they had.

1.4.1. Real-world Networks

Below, we present a few examples of real-world networks, many of which are well approximated by a scale-free degree distribution; we will see their most important properties to see which model they match.

1.4.1.1. Computer networks and the Internet

The computer world has lots of study networks that are connected in so many ways. The connections may be physical using a cable such as copper or optical fiber, and it may be also done by satellites or wireless systems. The range of these networks may be so variable too; it can go from LAN small networks with a few computers to WAN networks of entire cities. However, today, the most computers are isolated and connected by a common network, the Internet. This network is in charge of exchanging packages of information between the digital computers called routers.

1.4.1.2. Technological networks

This type of networks have been designed by people for distribute the resources of a country. Types of technological networks include the electrical power grid, the phone network, and the transport networks – roads, airline connections between airports, rail-roads, and subway networks. These networks are strongly correlated to the geographical and topological surface that they are attached, as the surface of the earth or the subsoil of a city, so the distance between two nodes becomes important in this field, due to this fact and to the need of high amount of money to develop this kind of networks, they don’t show the dynamicity of another networks such as the WWW.
1.4.1.3. Virtual technological networks and the WWW

Another type of networks are based not on physical connections, else in the logical connections. One of these networks is the World Wide Web, the network of HTML pages that are usually viewed using a browser. Each page is a node in the network, and if a link exists between pages, then a directed edge exists between them. This network is huge; in 2009 it was estimated to contain hundreds of billions of pages. One of the properties of this network is its dynamic behavior because pages can be created and links between them can change in seconds. Another network related to this one is the email network, where every person is a node and is linked to all other people in their address book.

Another example of this type of networks is the phone call graph. This is a graph created by phone network operators. Each node represents a phone number, and a directed link exists between nodes if the source of the link initiated a call to the destination within a certain time frame.

1.4.1.4. Social Networks

An important class of networks is the class of networks of social interactions between individuals. These may consist of networks of friendship, working relations or sexual relations. Besides their importance to social studies, understanding the structure of those networks is also important for example for the vaccination periods and the obligation to prevent diseases such as flues and sexually transmitted diseases.

Some other studied networks in this field are the actor network, where every actor is a node connected to every other actor with whom he or she has appeared in a movie and the network between citations in scientific articles. From the actor network, a curious experiment has been the “Six Degrees of Kevin Bacon”, an experiment that links an actor with another with the previous criterion, if they have starred in the same film; creators say that every actor is connected to Kevin Bacon in less than or in six degree which are other actors in movies.

1.4.1.5. Economic networks

Economic networks can be viewed as a special type of social network. However, the nodes may not represent individuals, but rather companies, countries or industries. The edges represent trading relations between companies or countries, companies sharing employees or stock-holding relationships.
1.4.1.6. Biological networks

One of the most important and well studied classes of networks are the biological networks. This category may contain several types of networks such as the one which represents the interactions between proteins, between genes or between proteins and genes. Other biological networks may be the food chain and the predator-prey network, where nodes are species, and a directed link represents a species that feeds with another. Despite these two networks are logical, we can also see physical networks in the biological field, for example, nervous system in the animals, the neurons in the brain and the network of blood vessels in an organism. Recently, the structure of neural networks has also been studied and has been shown to be scale free.

Table 1.1. Basic parameters of the networks previously mentioned [8]

<table>
<thead>
<tr>
<th>Network</th>
<th>Type</th>
<th>( n )</th>
<th>( m )</th>
<th>( z )</th>
<th>( l )</th>
<th>( \gamma )</th>
<th>( C(1) )</th>
<th>( C(2) )</th>
<th>( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>film actors</td>
<td>undirected</td>
<td>449913</td>
<td>25516482</td>
<td>113.43</td>
<td>3.48</td>
<td>2.3</td>
<td>0.2</td>
<td>0.78</td>
<td>0.208</td>
</tr>
<tr>
<td>company directors</td>
<td>undirected</td>
<td>7673</td>
<td>55392</td>
<td>14.44</td>
<td>4.6</td>
<td>-</td>
<td>0.59</td>
<td>0.88</td>
<td>0.276</td>
</tr>
<tr>
<td>math coauthorship</td>
<td>undirected</td>
<td>253339</td>
<td>496489</td>
<td>3.92</td>
<td>7.57</td>
<td>-</td>
<td>0.15</td>
<td>0.34</td>
<td>0.12</td>
</tr>
<tr>
<td>physics coauthorship</td>
<td>undirected</td>
<td>52909</td>
<td>245300</td>
<td>9.27</td>
<td>6.19</td>
<td>-</td>
<td>0.45</td>
<td>0.56</td>
<td>0.363</td>
</tr>
<tr>
<td>biology coauthorship</td>
<td>undirected</td>
<td>1520251</td>
<td>11803064</td>
<td>15.53</td>
<td>4.92</td>
<td>-</td>
<td>0.088</td>
<td>0.6</td>
<td>0.127</td>
</tr>
<tr>
<td>telephone call graph</td>
<td>undirected</td>
<td>47000000</td>
<td>80000000</td>
<td>3.16</td>
<td>2.1</td>
<td></td>
<td>0.005</td>
<td>0.001</td>
<td>-0.029</td>
</tr>
<tr>
<td>email messages</td>
<td>directed</td>
<td>59912</td>
<td>86300</td>
<td>1.44</td>
<td>4.95</td>
<td>1.5/2.0</td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>email address books</td>
<td>directed</td>
<td>16881</td>
<td>57029</td>
<td>3.38</td>
<td>5.22</td>
<td></td>
<td>0.17</td>
<td>0.13</td>
<td>0.092</td>
</tr>
<tr>
<td>student relationships</td>
<td>undirected</td>
<td>573</td>
<td>477</td>
<td>1.66</td>
<td>16.01</td>
<td></td>
<td>0.005</td>
<td>0.001</td>
<td>-0.029</td>
</tr>
<tr>
<td>sexual contacts</td>
<td>undirected</td>
<td>2810</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Information</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWW nd.edu</td>
<td>directed</td>
<td>269504</td>
<td>1497135</td>
<td>5.55</td>
<td>11.27</td>
<td>2.1/2.4</td>
<td>0.11</td>
<td>0.29</td>
<td>-0.029</td>
</tr>
<tr>
<td>WWW Altavista</td>
<td>directed</td>
<td>2.04E+08</td>
<td>2.13E+09</td>
<td>10.46</td>
<td>16.18</td>
<td>2.1/2.7</td>
<td>0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>citation network</td>
<td>directed</td>
<td>783339</td>
<td>6716198</td>
<td>8.57</td>
<td></td>
<td></td>
<td>0.07</td>
<td>0.082</td>
<td>-0.016</td>
</tr>
<tr>
<td>Roget’s Thesaurus</td>
<td>directed</td>
<td>1022</td>
<td>5103</td>
<td>4.99</td>
<td>4.87</td>
<td></td>
<td>0.13</td>
<td>0.15</td>
<td>0.157</td>
</tr>
<tr>
<td>word co-occurrence</td>
<td>undirected</td>
<td>460902</td>
<td>17000000</td>
<td>70.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technological</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Internet</td>
<td>undirected</td>
<td>10697</td>
<td>31992</td>
<td>5.98</td>
<td>3.31</td>
<td>2.5</td>
<td>0.035</td>
<td>0.39</td>
<td>-0.189</td>
</tr>
<tr>
<td>power grid</td>
<td>undirected</td>
<td>4941</td>
<td>6594</td>
<td>2.67</td>
<td>18.99</td>
<td></td>
<td>0.01</td>
<td>0.08</td>
<td>-0.003</td>
</tr>
<tr>
<td>train routes</td>
<td>undirected</td>
<td>587</td>
<td>19603</td>
<td>66.79</td>
<td>2.16</td>
<td></td>
<td>0.06</td>
<td>0.69</td>
<td>-0.333</td>
</tr>
<tr>
<td>software packages</td>
<td>undirected</td>
<td>1439</td>
<td>1723</td>
<td>1.2</td>
<td>2.42</td>
<td>1.6/1.4</td>
<td>0.07</td>
<td>0.082</td>
<td>-0.016</td>
</tr>
<tr>
<td>software classes</td>
<td>directed</td>
<td>1377</td>
<td>2213</td>
<td>1.61</td>
<td>1.51</td>
<td></td>
<td>0.033</td>
<td>0.012</td>
<td>-0.119</td>
</tr>
<tr>
<td>electronic circuits</td>
<td>directed</td>
<td>24097</td>
<td>53248</td>
<td>4.34</td>
<td>11.05</td>
<td>3</td>
<td>0.01</td>
<td>0.03</td>
<td>-0.154</td>
</tr>
<tr>
<td>peer-to-peer network</td>
<td>undirected</td>
<td>880</td>
<td>1296</td>
<td>1.47</td>
<td>4.28</td>
<td>2.1</td>
<td>0.012</td>
<td>0.011</td>
<td>-0.366</td>
</tr>
<tr>
<td>Biological</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>metabolic network</td>
<td>undirected</td>
<td>765</td>
<td>36886</td>
<td>9.64</td>
<td>2.56</td>
<td>2.2</td>
<td>0.09</td>
<td>0.67</td>
<td>-0.24</td>
</tr>
<tr>
<td>protein interactions</td>
<td>undirected</td>
<td>2115</td>
<td>2240</td>
<td>2.12</td>
<td>6.8</td>
<td>2.4</td>
<td>0.072</td>
<td>0.071</td>
<td>-0.156</td>
</tr>
<tr>
<td>marine food web</td>
<td>directed</td>
<td>135</td>
<td>598</td>
<td>4.43</td>
<td>2.05</td>
<td></td>
<td>0.16</td>
<td>0.23</td>
<td>-0.263</td>
</tr>
<tr>
<td>freshwater food web</td>
<td>directed</td>
<td>92</td>
<td>997</td>
<td>10.84</td>
<td>1.9</td>
<td></td>
<td>0.2</td>
<td>0.087</td>
<td>-0.326</td>
</tr>
<tr>
<td>neural network</td>
<td>directed</td>
<td>307</td>
<td>2359</td>
<td>7.68</td>
<td>3.97</td>
<td></td>
<td>0.18</td>
<td>0.28</td>
<td>-0.226</td>
</tr>
</tbody>
</table>
To end this part, we have included a table that shows some of the commented networks with all his basic parameters. The properties measured are as follows: the type of graph, directed or undirected, the total number of nodes $n$, the total number of edges $m$, the mean degree $\bar{z}$, the mean-vertex distance $\bar{l}$, the exponent $\gamma$ of the degree distribution if the distribution follows a power law (or “-” if not; in-out degree exponents are given for directed graphs), the clustering coefficient $C^{(1)}$ given from the global equation, the clustering coefficient $C^{(2)}$ given from the equation calculated by the mean clustering coefficient of the nodes and the correlation coefficient $r$, that we will talk later.

### 1.4.2. Properties of real-world Networks

The most of real networks share some common properties between them, in this part we will try to explain this features and the repercussion that they have in the entire graph.

#### 1.4.2.1. Degree distribution

Several real networks have a special degree distribution; this type of distribution is called scale-free.

The scale-free concept is applied to that networks that for more edges we add, they don't change the scale of their degree distribution. The most part of the scale-free networks show a power law degree distribution. In the early years of the graph study, the ER model was so extended than there were no other models and all the scientists took for granted that all the networks were random and followed a Poisson degree distribution. But nearly forty years later, when the BA model was released all the conception of the networks in the real world changed. Their study about the WWW showed that it was based on few nodes with a high link number; more than the 80% of the webs that they mapped had less than four links and less than a 0.01% of all the nodes had more than a thousand. Analyzing that information, they demonstrated that the resultant histogram had a potential distribution, the probability of a node being connected to $k$ others was proportional to $ck^\gamma$ where the $\gamma$ was approximately 2. In a widely way, the most of real networks have a $\gamma$ value between 2 and 3.

One of the ways to explain this behavior is the mechanism of preferential attachment and the fitness model, what is a similar mechanism to the preferential attachment but taking in account that the newer nodes will have another term that will decide their probability of attachment, not only letting the older nodes to be hubs of the network.
The power law distribution highly influences the network topology. It turns out that the major hubs are closely followed by smaller ones, these ones, in turn, are followed by other nodes with an even smaller degree and so on. This hierarchy allows for a fault tolerant behavior. Since failures occur at random and the majority of nodes have a relatively small degree, the probability of that failure will occur to a hub is too small, suggesting that such topologies could be useful for security. Even if such event happens, the network will be connected, which is guaranteed by the remaining hubs. On the other hand, if we choose a few major hubs and take them out of the network, it simply falls apart and is turned into a set of rather isolated graphs. Due to this, scale-free networks are so vulnerable to targeted attacks, which will destroy the connectedness very quickly, and it is shown that the security of the hubs is a vital part for the survival of the network. According to the experiments done by Barabási, deleting the 5% of the hubs of a scale-free network the distance for crossing the graph will be duplicated. If we remove between the 5% and the 15% of the better connected nodes the network will fall apart and the giant component of the graph will be reduced significantly.

According to the previous argument, hubs are important for the speed of the network, but it can also be harmful if a disease or a virus infects a hub, because it will reach the entire network so fast. Immunization of a random number of nodes will not be successful because the network will be connected to the hubs, so the solution will be immunize a high percentage of hubs; but this solution presents some problems like identifying the hubs and the ethical problem of only immunization some specific nodes.
Another important characteristic of scale-free networks is the clustering coefficient distribution, which decreases as the node degree increases. This distribution follows also a power law. That means that the low-degree nodes belong to very dense sub-graphs and those sub-graphs are connected to each other through hubs.

1.4.2.2. Small-world phenomenon

These hubs are also the responsible of the small-world phenomenon, in this case, where it seems that we have a disordered network, the average distance between two nodes is very small relative to a highly ordered network. Normally, the values that a distance between to nodes can have is about 4, 5 or 6, this means that specific links can connect vertices that are so far away one from the other.

This average distance has been proved in some studies and it originated the “six degrees of separation concept” which states that the mean distance for two people in the world in order to their known people was six. This concept is not so difficult to understand because the people you can reach increases exponentially with every degree, however, this other people may be underhanded creating little clusters of people.

1.4.2.3. Correlations

Another feature of the real networks is the relation between nodes with the same degree. That is, the probability of reaching a node by following a link is independent of the node from which the link emanated. This property can be studied analyzing the average degree of neighboring nodes as a function of degree. We can also calculate the correlation coefficient, $r$, between the degrees of neighboring sites:

$$r = \frac{\langle k_i k_j \rangle - \langle k \rangle^2}{\langle k^2 \rangle - \langle k \rangle^2}$$  \hspace{1cm} (1.24)

where averages are taken over all pairs of neighbors, $i$ and $j$.

It is also proved that uncorrelated power-law graph having $2 \leq \gamma \leq 3$ will also have a small diameter $d \sim \ln(\ln(N))$ where $N$ is the number of nodes in the network. So from a practical point of view, the diameter of a growing scale-free network might be considered almost constant.
1.5. Optimization algorithms

There are some types of optimization problems, some of them can be resolved easily by using simple methods because the problem shows a linear behavior; but in other problems there are more complications to find a solution and we will need to use systems to find a close approximation value due to the necessity of a high amount of time to solve them. In this chapter we will present two of the methods which we will use in this project.

In our case we will try to optimize an initial graph with the shape of a grid by using a cost function to turn it into a graph that shares some properties with the real airport network.

1.5.1. Threshold method

This method is based on the acceptability of a random generated solution between a threshold and a cost function that gives a value to the quality of the solution.

This will be the scheme of this method:

1. Define the number of iterations.
2. Define the function that will measure the quality of the approximation, the cost of the model.
3. Define the parameter that will reduce the error accepted on each iteration.
4. Compute the cost of the first model.
5. Define the error that the threshold will accept (taken from a percentage of the first cost of the model).
6. Copy the approximation to another variable to make changes.
7. Make a random change in the approximation.
8. Measure the cost of the new approximation.
9. If the new cost is lower than the older cost plus the error that we accept, we take it, if not, we don’t take the change and try again
10. We multiply the error with the reduction parameter to make the solution more sensible as the iterations go by.

1.5.2. Simulated annealing (SA)

This method is based on the acceptability of a random generated solution by a probability that will show if we accept a model that gives a worse cost than the previous one. The name of the method comes from the first algorithm that used this technique, which studied the cooling of some materials.

This will be the scheme of this method:
1. Define the number of iterations.
2. Define the function that will measure the quality of the approximation, the cost of the model.
3. Compute the cost of the first model.
4. Copy the approximation to another variable to make changes.
5. Make a random change in the approximation.
6. Measure the cost of the new approximation.
7. If the new cost is lower than the older one, we accept it.
8. If it is higher we will compute the probability of accepting a change that worsen the model.
9. We generate a random number to compare with the probability.
10. If the number generated is in the probability, we also accept the change.
11. If not, we don’t take the change and try again.

The first probability used in this method was:

\[ P[\delta E] = e^{-\frac{\delta E}{kT}} \]  

(1.25)

where \( \delta E \) is the difference between the cost of the model and the cost of the changed model, \( k \) is the Boltzmann constant and \( T \) is the cooling temperature given from an initial temperature and a reduction parameter.
CHAPTER 2. GLOBAL AIRLINE NETWORK ANALYSIS

2.1. Basic properties

Initially, we will study and determine the characteristics of the airline network, first for the two biggest parts of it, the North America network and the Asia and Middle East network, and finally for the entire global airline network.

For this study we will use the Python programming environment due to its package NetworkX, a very strong tool with such strong functions as the possibility of obtaining the features of the networks, generating new networks with given models, changing the networks, drawing the networks and plotting the degree distribution of the given network. For obtaining all the basic parameters of the networks we will use that package, and for treating the data, Microsoft Excel will be the main program to use.

2.1.1. Asia and Middle East airline network

The Asia and Middle East airline network is the second larger community of the global airline network, with 706 nodes and a total number of connections of 2574, due to this, the average degree for this network is 7.2918. Now we will see in a table the ten most connected nodes of this network.

Table 2.1. The ten most connected nodes of the Asia and Middle East airline network

<table>
<thead>
<tr>
<th>Node</th>
<th>BC</th>
<th>Degree</th>
<th>CC</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>69</td>
<td>0.096416928</td>
<td>87</td>
<td>0.416666667</td>
<td>Beijing</td>
</tr>
<tr>
<td>71</td>
<td>0.134054031</td>
<td>78</td>
<td>0.458984375</td>
<td>Bangkok</td>
</tr>
<tr>
<td>97</td>
<td>0.047763961</td>
<td>77</td>
<td>0.392538976</td>
<td>Guangzhou</td>
</tr>
<tr>
<td>210</td>
<td>0.07074402</td>
<td>73</td>
<td>0.435185185</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>544</td>
<td>0.033638053</td>
<td>71</td>
<td>0.390581717</td>
<td>Shanghai</td>
</tr>
<tr>
<td>539</td>
<td>0.104162774</td>
<td>70</td>
<td>0.4341133</td>
<td>Seoul</td>
</tr>
<tr>
<td>644</td>
<td>0.080971254</td>
<td>69</td>
<td>0.414462081</td>
<td>Tokyo</td>
</tr>
<tr>
<td>554</td>
<td>0.114321697</td>
<td>67</td>
<td>0.452212957</td>
<td>Singapore</td>
</tr>
<tr>
<td>468</td>
<td>0.061025838</td>
<td>65</td>
<td>0.410599884</td>
<td>Osaka</td>
</tr>
<tr>
<td>160</td>
<td>0.061874206</td>
<td>60</td>
<td>0.417159763</td>
<td>Dubai</td>
</tr>
</tbody>
</table>

If we talk about the distance terms in this network, we find that the diameter of the network is 9, but the average distance, in comparison, is very small; its value is 3.5489. Here we already find the small-world phenomenon because
the mean number of connections that we will have to do from one airport to another in this region of the world will be nearly four.

Now, it’s time to talk about the clustering coefficient of this network, it has a value of 0.4661. We can see that this value shows that almost the half of the nodes share a connection to the other ones, this also collaborates to create the small-world phenomenon.

The degree distribution of this network, showed in the figure (Fig. 2.2, Fig. 2.3), does not follow a clean power law, in the case of number of nodes it has a exponent of $\gamma = 1.314$. The shape of the accumulated probability distribution is not a clear power law in the total distribution, but if we divide the plot in separated parts it’s clearly visible.

**Fig. 2.2** Degree distribution of the Asian and Middle East airline network in terms of number of nodes.

**Fig. 2.3** Degree distribution of the Asian and Middle East airline network in terms of accumulated probability.
The next experiment that we will do is to remove the 10% of the hubs of the network, the seventy nodes that have the higher degree. We will try this change in this network because it is the smaller that we will study and the removal will be easier. We will expect a huge variation of the basic features of the network and a reduction of its well behaviour.

After an erase that seventy nodes, the first thing that we will check is if the new network is connected, and the answer is no. The first subgraph that we can find has 330 nodes and 483 edges with an average degree of 2.9273; the second biggest subgraph that we will find has 32 nodes and 63 edges with an average degree of 3.9375. Now if we focus in the giant component the degree distribution will be as shown in the figure (Fig. 2.4), we can see that the giant component also follows a power law due to his similar shape to the entire network. The highest degree now is 16 and the average degree is approximately 3, they have been reduced drastically. The clustering coefficient is now 0.2883 and this shows us that the most nodes don’t share connections due to the loss of the hubs.

![Degree distribution](image)

**Fig. 2.4** Degree distribution for the giant component of the Asian and Middle East airline network if we remove the 10% most connected nodes, expressed in number of nodes.

The degree distribution of the entire network is similar to the original, but in this case we will find that 167 nodes are not connected to anything due to the disappearance of the hubs.

### 2.1.2. North America airline network

Now we will talk about the North America airline network, the largest community of the global airline network in the world. It has a total amount of nodes of 940 and they are connected by 3446 edges. The average degree of this network is 7.3319. Now we will see in a table the ten most connected nodes in this network.
Table 2.2. The ten most connected nodes of the North America airline network

<table>
<thead>
<tr>
<th>Node</th>
<th>Degree</th>
<th>BC</th>
<th>CC</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>113</td>
<td>145</td>
<td>0.105955604</td>
<td>0.412203687</td>
<td>Chicago</td>
</tr>
<tr>
<td>43</td>
<td>131</td>
<td>0.036174816</td>
<td>3.64E-01</td>
<td>Atlanta</td>
</tr>
<tr>
<td>160</td>
<td>121</td>
<td>0.05092776</td>
<td>0.366510539</td>
<td>Dallas/Fort Worth</td>
</tr>
<tr>
<td>159</td>
<td>114</td>
<td>0.088403923</td>
<td>0.377715205</td>
<td>Denver</td>
</tr>
<tr>
<td>470</td>
<td>111</td>
<td>0.116056279</td>
<td>0.403350515</td>
<td>Minneapolis/St Paul</td>
</tr>
<tr>
<td>539</td>
<td>109</td>
<td>0.050573538</td>
<td>0.361710324</td>
<td>Pittsburgh</td>
</tr>
<tr>
<td>146</td>
<td>103</td>
<td>0.008817179</td>
<td>0.358396947</td>
<td>Cincinnati</td>
</tr>
<tr>
<td>174</td>
<td>100</td>
<td>0.024647408</td>
<td>0.362128808</td>
<td>Detroit</td>
</tr>
<tr>
<td>280</td>
<td>100</td>
<td>0.023391466</td>
<td>0.358807795</td>
<td>Houston</td>
</tr>
</tbody>
</table>

Now, talking in distance terms, the diameter of the network is 13 but in comparison with the Asian and Middle East network the ratio between the diameter and the average distance, in this case the average distance is 4.1333, is much larger. It means that the nodes are better connected in spite of the larger size of the network. It may be result of the higher average degree, although such a large change in the ratio may not be reflected in the change in the average degree between the networks, which is so small.

This network also shows clearly us a feature that is also shared with the global airline network but it’s not evident at a glance. This property is that the most connected nodes, the hubs are not the best connected nodes. This means that the hubs sometimes don’t have the highest betweenness centrality. In the next table we will see which nodes have the higher betweenness centrality and try to explain why.

Table 2.3. The ten best connected nodes of the North America airline network

<table>
<thead>
<tr>
<th>Node</th>
<th>Degree</th>
<th>BC</th>
<th>CC</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>35</td>
<td>0.281186939</td>
<td>0.361153846</td>
<td>Anchorage</td>
</tr>
<tr>
<td>617</td>
<td>78</td>
<td>0.170726305</td>
<td>0.392394484</td>
<td>Seattle</td>
</tr>
<tr>
<td>979</td>
<td>85</td>
<td>0.12662443</td>
<td>0.406669554</td>
<td>Toronto</td>
</tr>
<tr>
<td>470</td>
<td>111</td>
<td>0.116056279</td>
<td>0.403350515</td>
<td>Minneapolis/St Paul</td>
</tr>
<tr>
<td>113</td>
<td>145</td>
<td>0.105955604</td>
<td>0.412203687</td>
<td>Chicago</td>
</tr>
<tr>
<td>998</td>
<td>35</td>
<td>0.1028399</td>
<td>0.339847991</td>
<td>Winnipeg</td>
</tr>
<tr>
<td>159</td>
<td>114</td>
<td>0.088403923</td>
<td>0.377715205</td>
<td>Denver</td>
</tr>
<tr>
<td>905</td>
<td>49</td>
<td>0.083167567</td>
<td>0.357170027</td>
<td>Montreal</td>
</tr>
<tr>
<td>208</td>
<td>34</td>
<td>0.07750508</td>
<td>0.303785183</td>
<td>Fairbanks</td>
</tr>
</tbody>
</table>
We can see that the Anchorage International Airport is the best connected. This airport is placed in Alaska and it is easily visible that the most of the out flights to Asia from North America will pass through it.

This is the shape what the network will have:

![North American airline network](image)

**Fig. 2.6** North American airline network

In this network, the average clustering coefficient is 0.5175. It is higher than in the previous network, so this means that in this one, there are more nodes which share a connection to, this may be another explanation to the phenomenon of the higher ratio between the diameter and the average distance.

Finally, we will talk about the degree distribution of the network, the next figures will show us which are for this network. We can clearly see in the second figure (**Fig. 2.8**) that the degree distribution for the most of the values follows a power law if we divide the plot in two sections.
2.1.3. Global airline network

Finally it’s time to talk about the entire airline network. It is a huge network with 3618 nodes and 14142 edges, so it seems that the amount of edges for the total nodes may be excessive. This two data combined show an average degree of a node of 7.8176. Now we will see the top 25 airports in terms of connections.

Tab 2.4. Top 25 most connected airports in the world

<table>
<thead>
<tr>
<th>Node</th>
<th>BC</th>
<th>Degree</th>
<th>CC</th>
<th>Mean distance</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>2343</td>
<td>0.093420381</td>
<td>250</td>
<td>0.371966269</td>
<td>2.688415814</td>
<td>Paris</td>
</tr>
<tr>
<td>1775</td>
<td>0.084988984</td>
<td>242</td>
<td>0.379379064</td>
<td>2.635886093</td>
<td>London (GB)</td>
</tr>
<tr>
<td>962</td>
<td>0.065577711</td>
<td>237</td>
<td>0.372732894</td>
<td>2.68288637</td>
<td>Frankfurt</td>
</tr>
<tr>
<td>119</td>
<td>0.040492131</td>
<td>192</td>
<td>0.366019024</td>
<td>2.732098424</td>
<td>Amsterdam</td>
</tr>
<tr>
<td>2020</td>
<td>0.052210812</td>
<td>186</td>
<td>0.347420997</td>
<td>2.878352226</td>
<td>Moscow</td>
</tr>
<tr>
<td>548</td>
<td>0.044443497</td>
<td>184</td>
<td>0.35603898</td>
<td>2.808681228</td>
<td>Chicago</td>
</tr>
<tr>
<td>2229</td>
<td>0.069283491</td>
<td>179</td>
<td>0.369119298</td>
<td>2.70915123</td>
<td>New York</td>
</tr>
<tr>
<td>179</td>
<td>0.024896178</td>
<td>172</td>
<td>0.342908608</td>
<td>2.916228919</td>
<td>Atlanta</td>
</tr>
<tr>
<td>732</td>
<td>0.022851668</td>
<td>147</td>
<td>0.337753292</td>
<td>2.960740946</td>
<td>Dallas/Forth Worth</td>
</tr>
<tr>
<td>1205</td>
<td>0.017457043</td>
<td>144</td>
<td>0.335124618</td>
<td>2.983964612</td>
<td>Houston</td>
</tr>
<tr>
<td>2073</td>
<td>0.008585744</td>
<td>143</td>
<td>0.335279941</td>
<td>2.98258225</td>
<td>Munich</td>
</tr>
</tbody>
</table>
In terms distance, the diameter of the network is 17 and the average distance is 4.4396. It is curious to see how we can reach any point of the world in an average number of almost 4 connections, what is in fact done by the globalization phenomenon. It is also curious that from the most important airports, with only 3 jumps to a plane we can reach any other airport of the world. Another curious feature of this network is that the most connected airports are not necessarily the most transited airports in terms of millions of passengers (if we compare the data with the data of the year 2009, easily findable in the Wikipedia [35]).

The clustering coefficient for this network is 0.4957, this means that almost the half of the nodes share a connection with the others, and normally these connections will be the hubs.

The degree distribution for this network is shown in the next figures (Fig. 2.10, Fig. 2.11). In the first one, we can see that the representation of the number of nodes follow a power law distribution with an exponent of 1.457. In the second one with the accumulated probability, it has a power law distribution but divided in three different parts with three different exponent values shown in the figure by the slope of the three lines that are over the curve.
2.2. Simulation of the global airline network

In this final part we will try to simulate a model that bears similarities with the global airline network. For this simulation we will use Networkx and different optimization algorithms, whose main aim is to obtain a graph with a degree distribution equal to the real and similar parameters.

For the simulations we will use in first term two computers which have these processor features:

- Intel Core 2 Duo 2.00 GHz.
- Intel Core 2 Quad 2.40 GHz.
The main program used will be Python in the 2.7 version, and the NetworkX package in the 1.4 version.

### 2.2.1. Generation of the model

First of all, we will need to find a model which fits in the desired real network. Once explained the classic models of graph generation we could see that the perfect model for this type of network will be the BA model due to its power law degree distribution, but due to the situation of the network, it is placed in a sphere which is the Earth, it is more accurate to use a grid graph.

Our grid will not be a normal grid, it will be a periodical grid with a toroidal shape in order to represent that all nodes at the end of the grid can be connected to the other in the other end due to the Earth’s shape. We can also represent a starting approach to the real airline network with its average degree; with the data obtained from the previous analysis of the network, we will create a graph with average degree 8. The number of nodes is also given by the previous analysis.

Once we have these assumptions, Networkx will give us an easy function to create this graph and we will only need to add the diagonal edges to the network.

At the beginning of the simulations we tried a grid of 27x27 nodes, which had a number of nodes similar to the Asia and Middle East airline network, and with a 30x30 grid, similar to de North America airline network. Due to the high simulation times, we reduced this to a 10x10 grid after some simulations as we shall see later.

The initial aspect of both graphs was as shown in the next figure (Fig. 2.12); both graphs shared similar aspect so in the figure we will show one different view for each one.
Fig 2.12 Initial models for the global airline networks, Asian and Middle East (left) and North America (right).

2.2.2. **Cost function**

Once we have the model, we will need to define something to optimize in order to make the model an approach to the real network. For this case, we have developed a cost function which determines the global cost of the model in a given moment so we can compute the impact of the changes in our graph.

The cost function is focused on edges; the function will obtain some features of the edges and will make an accumulative value for the entire graph. These three aspects that we considered important in the airline network will be explained now.

2.2.2.1. *Distance between the nodes*

This is an important factor because in the global network, when a plane flies further, it costs more to the operator of that flight; so this was the first term of the cost function that we will take in account.

For the programming part, the computation of this value was easy due to the grid shape of the network, so we could treat them as points and the distance between them was the sum of the squared values of their coordinates.

2.2.2.2. *Betweenness centrality of the edge*

This term will increase the cost if in an edge has a high number of short paths that cross through it. This term tries to avoid the excessive load of some edges and the fair distribution of the flights through the connections.

This value was easy to obtain because Networkx has a function to give all the edge betweenness centrality of every node and we only had to know which edge it is.

2.2.2.3. *Combustible factor*

The combustible factor tries to catch the idea which if you fly further; the combustible wasted per nautical mile will be less than if you go to a place near your departure airport. For computing this factor we have considered a 10% of change in the total cost from the nearest airport to the furthest. We have done
this by using the maximum distance in the graph and a simple equation system to find it. The system will have an aspect as:

\[
\begin{align*}
1.1 &= (1 + K) \cdot G \\
1 &= (distance_{max} + K) \cdot G
\end{align*}
\] (1.3)

2.2.3. Optimization using the threshold method

The first attempt to the creation of the graph will be using the threshold method explained before applied to this case using some slight variations of the cost function.

The basic algorithm to make this optimization follows this structure:

- Definition of the cost function.
- Definition of the function to compute the average distance.
- Creation of the grid
- Add the new edges to make an average degree of 8.
- Compute the initial values of:
  - Diameter
  - Clustering coefficient
  - Average distance
  - Degree distribution
  - Initial cost
- Define the initial value of the threshold, a 10% of the initial cost.
- Start iterating NK times for a total value of 30 times.
- Make another iteration of 30 changes in every NK step.
- Compute the cost of the actual graph.
- Save the graph in another variable.
- Pick a random node with a degree higher or equal to 1.
- Pick another random node of all that the previous one had connections.
- Remove that edge.
- Add an edge between two other random nodes.
- Check if the graph is connected.
- If not, undo the process and try again.
- Compute the new cost.
- If the cost is better than the initial, we accept it.
- If the cost is better than the initial plus the threshold, we accept it.
- If not, we discard the changes and start again.
- Repeat for all the total number of changes.
- Reduce the threshold a 10% of its actual value
- Repeat for till we reach the NK value.
- Obtain the final values of:
  - Diameter
- Clustering coefficient
- Average distance
- Degree distribution
- Initial cost

This scheme was used in four different variations of the cost function to obtain the desired graph.

2.2.3.1. \textit{Distance x Edge BC x Factor}

The first try was assuming that the three different values had the same importance in the cost equation. After some simulations the final values of the basic characteristics were:

Table 2.5. Comparison between the original North American network and the model after the simulation.

<table>
<thead>
<tr>
<th></th>
<th>North America</th>
<th>Initial values</th>
<th>Final values</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>940</td>
<td>900</td>
<td>900</td>
</tr>
<tr>
<td># edges</td>
<td>3446</td>
<td>3600</td>
<td>3600</td>
</tr>
<tr>
<td>Average degree</td>
<td>7,3319</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Diameter</td>
<td>13</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>Average Clustering Coefficient</td>
<td>0,5175</td>
<td>0,4286</td>
<td>0,1707</td>
</tr>
<tr>
<td>Average Distance</td>
<td>4,1333</td>
<td>10,0167</td>
<td>3,8196</td>
</tr>
</tbody>
</table>

The reduction of the average distance and the diameter seems a good start for this simulation but once we see the degree distribution of the simulated graph, it is clear that things haven’t gone so well.

The degree distribution in the figure (Fig. 2.10) shows a clear Poisson distribution typical from the random graphs of the ER model.

![Fig. 2.10 Degree distribution of the simulated model.](image)
2.2.3.2. \( \text{Distance}^2 \times \text{Edge BC} \times \text{Factor} \)

The second assumption was that distance could affect in a squared way to the cost. After some simulations the values were:

**Table 2.6.** Comparison between the original Asia and Middle East network and the model after the simulation.

<table>
<thead>
<tr>
<th></th>
<th>Asia and Middle East</th>
<th>Initial Values</th>
<th>Final Values</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>706</td>
<td>729</td>
<td>729</td>
</tr>
<tr>
<td># edges</td>
<td>2574</td>
<td>2916</td>
<td>2916</td>
</tr>
<tr>
<td>Average degree</td>
<td>7,2918</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Diameter</td>
<td>9</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>Average Clustering Coefficient</td>
<td>0,4661</td>
<td>0,4286</td>
<td>0,1371</td>
</tr>
<tr>
<td>Average Distance</td>
<td>3,5489</td>
<td>9</td>
<td>3,6026</td>
</tr>
</tbody>
</table>

With the data of the table we could think that we were on a right situation, but the low final clustering coefficient doesn’t say so. Again the degree distribution will be clearly Poissonian.

![Fig 2.11 Degree distribution for the second simulations.](image)

2.2.3.3. \( e^{\text{Distance}} \times \text{Edge BC} \times \text{Factor} \)

Another approach was considering the distance of between the nodes in an exponential way. The results were:
Table 2.7. Comparison between the original North America network and the model after the simulation.

<table>
<thead>
<tr>
<th></th>
<th>North America</th>
<th>Initial Values</th>
<th>Final Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># nodes</strong></td>
<td>940</td>
<td>900</td>
<td>900</td>
</tr>
<tr>
<td><strong># edges</strong></td>
<td>3446</td>
<td>3600</td>
<td>3600</td>
</tr>
<tr>
<td><strong>Average degree</strong></td>
<td>7,3319</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td><strong>Diameter</strong></td>
<td>13</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td><strong>Average Clustering Coefficient</strong></td>
<td>0,5175</td>
<td>0,4286</td>
<td>0,1706</td>
</tr>
<tr>
<td><strong>Average Distance</strong></td>
<td>4,1333</td>
<td>10,0167</td>
<td>3,8167</td>
</tr>
</tbody>
</table>

In this case we are in the same situation than the two cases before. The degree distribution has also a Poisson shape.

![Degree distribution of the third series of simulations.](image)

2.2.3.4.  *Log*(Distance) x Edge BC x Factor

The last approach for this method was the thought that distance had a logarithmic relationship with the cost. The results in this case were:

Table 2.8. Comparison between the original Asia and Middle East network and the model after the simulation.

<table>
<thead>
<tr>
<th></th>
<th>Asia and Middle East</th>
<th>Initial values</th>
<th>Final Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># nodes</strong></td>
<td>706</td>
<td>729</td>
<td>729</td>
</tr>
<tr>
<td><strong># edges</strong></td>
<td>2574</td>
<td>2916</td>
<td>2916</td>
</tr>
<tr>
<td><strong>Average degree</strong></td>
<td>7,2918</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Diameter</td>
<td>9</td>
<td>13</td>
<td>6</td>
</tr>
<tr>
<td>---------</td>
<td>---</td>
<td>----</td>
<td>---</td>
</tr>
<tr>
<td>Average Clustering Coefficient</td>
<td>0.4661</td>
<td>0.4286</td>
<td>0.1385</td>
</tr>
<tr>
<td>Average Distance</td>
<td>3.5489</td>
<td>9</td>
<td>3.6227</td>
</tr>
</tbody>
</table>

The results this time were also inconclusive, so we thought that the variation of the distance in the cost was not the cause of the creation of a scale-free network, so we tried other variables.

![Fig. 2.13 Degree distribution of the simulations.](image)

2.2.3.5. Variations of the cost in the edge betweenness centrality

Due to the bad results of this first series of simulations we considered that the distance variable was not the most influential in the optimal network, so we tried the same type of operations (squared, exponential, logarithmic) in the betweenness centrality of the edge. The results were almost the same than before and we discarded that possible solution.

2.2.4. Optimization algorithm using simulated annealing

Due to the bad results of our first method, we tried again but using the simulated annealing method instead the threshold method. In this case we started using the 10x10 grid in order to improve our simulation times.

The basic algorithm to make this optimization follows this structure; it’s mostly the same than the previous method:

- Definition of the cost function.
- Definition of the function to compute the average distance.
- Creation of the grid
- Add the new edges to make an average degree of 8.
- Compute the initial values of:
  - Diameter
  - Clustering coefficient
  - Average distance
  - Degree distribution
  - Initial cost
- Define the initial value of the temperature, a 10% of the initial cost.
- Start iterating NK times for a total value of 30 times.
- Make another iteration of 30 changes in every NK step.
- Compute the cost of the actual graph.
- Save the graph in another variable.
- Pick a random node with a degree higher or equal to 1.
- Pick another random node of all that the previous one had connections.
- Remove that edge.
- Add an edge between two other random nodes.
- Check if the graph is connected.
- If not, undo the process and try again.
- Compute the new cost.
- Subtract the initial cost from the new cost.
- If the result of the subtraction is positive, we accept the change.
- If not, we generate a random number between 0 and 1.
  - If the random number is smaller than $e^{\frac{\Delta \text{cost}}{\text{Temperature}}}$, we accept the changes. The temperature is a factor that changes in every model, in our case it will be the 10% of the initial cost.
- If not, we discard the changes and try again.
- Repeat for all the total number of changes.
- Repeat for till we reach the NK value.
- Obtain the final values of:
  - Diameter
  - Clustering coefficient
  - Average distance
  - Degree distribution
  - Initial cost

Table 2.9. Results obtained using the simulated annealing

<table>
<thead>
<tr>
<th></th>
<th>Initial values</th>
<th>Final Values</th>
</tr>
</thead>
<tbody>
<tr>
<td># nodes</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td># edges</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>Average degree</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Diameter</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Average Clustering Coefficient</td>
<td>0.4286</td>
<td>0.0856</td>
</tr>
<tr>
<td>Average Distance</td>
<td>3.3838</td>
<td>2.4747</td>
</tr>
</tbody>
</table>
The new results are as inconclusive as the previous ones, and very similar so we can conclude that simulated annealing and the threshold method have the same results in a long term, so from now on we will use the threshold method due to his simplicity.

### 2.2.5. Add, remove and change with a given probability

The previous simulating methods have not given good results, so maybe the approach was wrong. The only change of some edges may be too restrictive at some points and it may do that the most nodes tend to have the average degree as their degree. Now we will try to simulate the network with the possibility of adding and removing edges to a node too. This process will be done with certain probability in order to maintain the graph bounded and in the desired size.

The core of the simulation will be the same; the change will be done in the iteration process. We will generate a random number; if the number is between the first probability margin, we will change an edge randomly, if it’s in the second, we will remove an edge in a random way and if it’s in the third we will add an edge randomly. In every iteration we will evaluate the cost of the function, if it’s better we will accept it, if not (out of the threshold), we won’t. The cost function that we will use will be the first one due to its simplicity.

One of the first things that we will need to do is to assure that the Python random number generator is reliable. In a simulation of a hundred series of a hundred random values, the mean value of the overall probability is 0.50043, and the mean values of each series seem random too; so we can consider that the approach is fine.
2.2.5.1. **Probabilities of 80%, 10% and 10%**

The first attempt will be with an 80% of probability of changing an edge at random, and 10% for both removing and adding an edge at random. The graph studied will have 15x15 nodes.

After the first simulations with short iteration times, we can see that the distribution now tends to move to the state where we can find higher number of low degree nodes. As we can see in the next figure (Fig. 2.7) the mean degree has gone under from the initial 8 value.

![Degree distribution of the three first simulations.](image)

It seems a better degree distribution compared to the previous ones in terms of number of nodes with lower degree, but if we see the average clustering coefficient of 10 simulations it will be so low again, with a mean value of 0.0454 which is still so far away from the 0.5 values that it has in the real network; and also the distribution follows a clear Poisson shape.

2.2.5.2. **Probabilities of 70%, 15% and 15%**

The second attempt will be with a 70% of probability of changing an edge at random, and 15% for both removing and adding an edge at random and we will see how this variation affects to the distribution.
2.2.5.3. Focusing on the nodes instead of the edges

Finally, as we have seen that focusing on the edges is not giving us good results we have focused on the nodes. The cost function in this case has been the same, having the betweenness centrality of every node and, instead the distance between the nodes, the average distance of that node to the others. The combustible factor has been ignored in these simulations.

The results were also as bad as the other, so we returned to the initial assumption that edges carry the weight of the network with a much more simple function that seemed to have a better behaviour, but the results seemed as bad as previously.

2.2.6. Conclusions

Networks are a very important part of our world. The current theories don’t explain exactly the mechanisms which have lead to the structure and properties of real networks. Thus, the study of new models matching real data should provide new clues to understand the process followed in the formation of real networks characterized by short paths among nodes and a good reliability.

The global airline network, as we have seen, is a network which has two important characteristics of complex networks: it is small-world (with a short maximum distance between any pair of nodes) and it is scale-free (with a power
law degree distribution). In some sense, this network makes the world so small that in a few jumps we can reach every airport of the world from another one.

Using NetworkX, we have analyzed the global airline network and we have verified its main properties: diameter, average distance, clustering, degree distribution. We have also studied the network resulting of deleting part of the nodes, and we see that, deleting only about 10% of the highest degree nodes, the network will be hugely disconnected into smaller networks; this is due to the relevance of a low number of hubs, confirming known results. If we focus on the most connected airports of the world (degree centrality), we see, by comparing this data with the busiest airports in the world, from Wikipedia [35], that they don't match. This is because the most connected airports are placed in particular places and from there the traffic is distributed to the rest of the continent, but this don't necessarily means that the routes they have are the most transited.

In our models we have considered two evolutionary algorithms for the optimization process: threshold optimization and simulated annealing. We have seen that both algorithms produce approximately the same results. Another result is that the second series of simulations with an extra probability of adding and removing edges, shows a better behaviour (e.g. a higher clustering) than the first set of experiments based only on reconnecting links, but both methods produce almost random graphs. The impossibility of producing a graph with the same characteristics than the real network could be for different reasons: the initial graph, the cost function, the optimization algorithm and some of the many parameters involved. As the evolutionary algorithms have been fully tested in other similar problems, we think that the most relevant factor is the cost function. There is need to test other modifications of this function, such that include other parameters accounting for real properties of airline connections. These comments and our analysis should be taken into account in future research on this topic.

2.3. Bibliography and references

Global airline network analysis


A model for the global airport network based on the intermediation degree of the edges


