UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE FÍSICA

World Tourism and Airline Networks: Structure and Correlations

Luís Manuel Domingos Rebelo

Mestrado em Física
Especialização em Física Estatística e Não-Linear

Dissertação orientada por:
Prof. Doutor Nuno Araújo

2018

Acknowledgements

I would like to thank all the people who have helped make this work possible and have been a source of support over this past year, in particular my family, colleagues, friends and also the other members of our group, at the Centro de Física Teórica e Computacional.

I specially thank my supervisor Nuno Araújo for his guidance throughout this journey and I also would like to thank Trivik Verma for his advice and collaboration.

Abstract

Tourism is an ever-growing business and its impact on the world's economy is pretty noticeable but above all it serves as a measure of the wealth transfer between different countries. This, along with the several side effects that it triggers, some good and some bad, make it a business that is so present in our everyday lives that it is nearly impossible to escape it or its effects.

Nowadays, most of tourists worldwide make use of air transportation to get to their touristic destination. This network, mainly serves people travelling for business or touristic reasons as well as to transport cargo. Recently, there has been a growth, in the literature, of the use of complex network theory in order to study transportation networks. This theory studies the correlations between the different components of systems (nodes) resulting from the interaction between them (represented by links) and maps all these elements and interactions into what is called a complex network.

The reliability and wide range of the world airline network have allowed for a boom in tourism in the last decades, with international tourism receipts increasing from US$ 495 billion in 2000 to US$ 1220 billion in 2016 while international tourist arrivals grew from 674 to 1235 million in this time interval and are expected to reach 1.8 billion by 2030, according to industry projections [1].

The recent boom in worldwide tourism and its relevance to the world economy is obvious. However, there are still questions about it that literature has not been able to answer such as how far do people really travel and if touristic travel is really global or rather limited to a specific region. In order to answer these questions we will use complex network theory which represents the parts of a system as nodes (in our case countries) and the interactions between these nodes as links (in our case, touristic flow). With this model we studied the structure and functionality of the WTN as well as the interplay between the two, whilst taking into account the influence of the world airline network (WAN) in the WTN. In order to achieve this, we use complex network theory and using methods and concepts from statistical physics allowing us to have a quantitative description of this network.

We find that touristic travelling is indeed mainly local, as it is corroborated by the strong resemblance between our obtained communities and the geopolitical divisions of the world. We also see that the structural part of the network, mainly provided by the WAN and the functional part are strongly correlated, indicating that the WAN provides a very good infrastructural support for the WTN. The difference in the clustering coefficients ($< C^{cyc} >$ and $< C^{und} >$) is significant, not only on its difference but also on its meaning since it tells us that that even though most countries are well connected, this characteristic of the network breaks down when we add directionality, meaning that most of the links of the network are not reciprocated, as it is also evidenced by the reciprocity value.

Keywords— tourism, human mobility, complex networks, world airline network, chemical distance

Resumo

Desde os primórdios do ser humano que a mobilidade foi importante para o desenvolvimento deste. Esta permitiu que os primeiros *Homo erectus* se começassem a dispersar pelos diversos continentes, algum tempo após o seu aparecimento. A mobilidade tem evoluído constantemente ao longo da história e a vida humana tem sido fortemente influenciada pela mobilidade, sendo que esta foi sempre uma boa medida do progresso humano. Para além disto, a exploração humana foi sendo tornada cada vez mais fácil à medida que a mobilidade se desenvolvia, sendo que os padrões de mobilidade estão cada vez mais globais. Estes têm o suporte de redes de transporte, que podem ser locais, regionais ou até globais, chegando a cobrir o planeta inteiro.

A rede de transportes mais global é a rede de transporte aéreo, que serve pessoas viajando por razões profissionais ou turísticas. Recentemente, tem havido um crescimento do número de artigos científicos, que fazem uso da teoria de redes complexas para estudar redes de transporte. Estas redes de transporte podem ser redes de caminhos ferroviários, redes das ruas de uma cidade ou redes de aeroportos, para dar alguns exemplos. A teoria de redes complexas é usada para representar sistemas físicos onde a distribuição do números de primeiros vizinhos não é uniforme. Para representar estes sistemas, são usados nodos que representam os constituintes de cada sistema e ligações para representar as interações entre os nodos. A esta representação, dá-se o nome de rede complexa, por oposição a redes regulares, encontradas na Física da Matéria Condensada, onde todos os elementos da rede têm o mesmos números de vizinhos. A teoria de redes complexas tem várias aplicações em diversas áreas da ciências e alguns exemplos de sistemas estudados usando a teoria de redes complexas incluem a *World Wide Web*, redes de fornecimento de energia, redes metabólicas ou até redes de contactos sociais.

A rede estudada neste trabalho designa-se por Rede Mundial de Turismo (World Tourism Network - WTN) e representa o fluxo de turistas a nível mundial. Para tal, usamos uma base de dados disponibilizada pela Organização Mundial de Turismo das Nações Unidas (United Nations World Tourism Organization - UNWTO), onde temos, para cada país, a quantidade de chegadas de turistas nesse mesmo país e qual a sua origem, para os anos entre 2004 e 2008, inclusive. Apesar de a base de dados se referir ao número de chegadas a um certo país e não ao número de turistas que usam o transporte aéreo, nós assumimos que, devido ao fator massivo que este último tem no meio de transporte utilizado pelos turistas para chegarem ao seu destino, os dois valores são equivalentes. De referir ainda que na construção da nossa rede, usamos o número total de turistas entre 2004 e 2008 e fazemos a análise toda, excepto em casos pontuais, com estes dados em vez de criarmos uma rede para cada ano.

Para representar a rede, nós consideramos os países como nodos do sistema e o fluxo de turistas de um país para o outro como as interações deste sistema, ou seja, as ligações da rede. Como ligações, consideramos o número total de turistas durante o período a que se refere a base de dados em vez de outros períodos de tempo. De referir ainda que esta rede é direta, ou seja, as ligações da rede têm uma direção, sendo que nem todas as ligações são recíprocas, ou seja, termos pares de países onde um país "envia" turistas para outro mas não recebe.

Nesta rede, a nossa métrica baseia-se na rede mundial de transporte aéreo (World Airline Network - WAN), sendo que para a WTN, nós fazemos um *coarse-graining*, onde em vez de considerarmos o número de voos que são necessários para ir do aeroporto A para chegar ao aeroporto B como a distância da rede, que designamos por distância química. Nós usamos uma base de dados pública, com cerca de 60000 rotas aéreas, para saber se existe um voo direto do aeroporto do país A para outro aeroporto do país B, e se sim, a distância química na rede do país i para o país j é igual a 1, sendo que como a grande maioria dos voos são recíprocos, consideramos também que a distância química do país B para o A seria 1.

O turismo tem crescido bastante nas últimas décadas, principalmente devido ao crescimento da WAN e da sua comodidade o que permite mais facilmente fazer viagens mais longos, sendo que as receitas internacionais provenientes do turismo aumentaram de 495 milhares de milhões de dólares em 2000 para 1220 em 2016 sendo que as chegadas internacionais de turistas cresceu de 674 para 1235 milhões neste período de tempo e prevê-se que cheguem a atingir 1800 milhões no ano 2030, de acordo com projeções da indústria.

O crescimento recente no turismo e a sua relevância tanto para a economia como a formação de relações entre diferentes culturas torna este tópico relevante em termos de estudo. Ainda existem, no entanto, diversas questões acerca do turismo que a literatura ainda não foi capaz de responder tais como o quão longe as pessoas viajam, se o turismo é realmente global ou ainda se somos enviesados nas nossas escolhas de viagens turísticas.

Depois de modelarmos a WTN, estudámos a sua estrutura e funcionalidade e a relação entre as duas, tendo em conta a influência da WAN na WTN. Estudámos também a presença de comunidades nesta rede, uma vez que estamos a falar de uma rede global com componentes diversas que interagem de maneira diferente entre elas, e este estudo pode revelar-nos algo em relação ao turismo que não sabíamos antes e que pode ser importante para os vários agentes ligados a este negócio. Por último, olhámos para a evolução da dinâmica da rede ao logo dos anos em estudo.

Como resultados, obtivemos que, em média, os turistas de um país precisam de fazer 2.32 voos entre países para chegar ao seu destino turístico e também que, em média, cada país tem um número baixo de países a um voo de distância (19.38), sendo que este resultado pode ser explicado pela presença de países/territórios isolados na base de dados usada.

Em termos da funcionalidade da rede, descobrimos que os turistas viajam para destinos perto, seja em distância química (1) ou distância geográfica ($[0, 3000]\, km$), quando comparado com a distância máxima percorrida em cada métrica (4 e 20037 km respectivamente).

A análise da relação entre a estrutura e a funcionalidade da rede permite-nos dizer que existe uma correlação forte entre estes, com um coeficiente de correlação $\rho = 0.903$, o que significa que a rede (WTN) está bem correlacionada com a estrutura oferecida pela WAN.

Para a deteção de comunidades, usámos o método de Louvain que compara a intensidade das ligações formadas entre um nodo e os seus vizinhos e compara estas com as interações que existiriam se as ligações fossem reorganizadas de forma aleatória, e conseguimos replicar as divisões geopolíticas do mundo dando força à hipótese que o turismo é na sua maioria um negócio de escalas regionais/intra-continentais. Finalmente, o estudo da evolução da rede permitiu-nos vislumbrar qual foi o crescimento da rede e vimos que a evolução da rede foi gradual e que grandes mudanças ocorreram poucas vezes.

Em resumo, podemos afirmar que as nossas escolhas para viagens turísticas são de facto enviesadas, visto que existe uma tendência para viajar para destinos perto do nosso país de origem e também que o turismo é na sua maioria um negócio regional/intra-continental, como é corroborado pela forte semelhança entre as comunidades obtidas usando o método de Louvain e as divisões geopolíticas do mundo.

Palavras-chave— turismo, mobilidade humana, redes complexas, rede mundial aérea, distância química

Contents

List of Figures

List of Tables

Introduction

Mobility has always been a presence in the life of humans beings, starting by the first *Homo erectus* specimens who began to disperse soon after their emergence [2]. Human life has evolved around mobility with exploration and growth have been made possible by it. These mobility patterns have the support of transportation networks, which can either be local, regional or even global networks covering the entirety of the globe.

The most global of all the transportation networks is the airline network, which mainly serves people traveling for business or touristic reasons as well as to transport cargo. Recently, there has been a growth, in the literature, of the use of complex network theory in order to study transportation networks. This theory studies the relationships of the different components of systems (nodes), and the interactions between them (links) and represents all these parts in what is called a complex network. Examples of systems studied using complex network theory include the World Wide Web [3], power grid networks [4], metabolic networks [5] or even social networks [6].

In the literature, airline networks have been extensively studied [7–13] and in their respective complex network representations, the nodes correspond to airports and links to connections between airports. Also, since a large part of these connections are two-way, the links are usually considered undirected.

The airline network is essential for business and tourism and this latter has grown in the last decades, with international tourism receipts increasing from US$ 495 billion in 2000 to US$ 1220 billion in 2016 while international tourist arrivals grew from 674 to 1235 million in this timespan and are expected to reach 1.8 billion by the year 2030, according to industry projections [1].

When people travel for touristic reasons, they mainly use the airline network, but unlike airline networks, in tourism the number of people from a certain territory that travel to another is not the same as the ones that do the inverse way thus making us consider the links in tourism networks as directed and also that they represent the number of tourists going from place A to place B and vice versa.

The recent boom in worldwide tourism and its relevance to the world economy makes it a very relevant topic to study. However, there are still questions about it that literature has not been able to answer such as how far do people really travel and if touristic travel is really global or even if we are travelling biased and in order to answer those questions we intend to study the structure and functionality of the WTN as well as the interplay between the two, whilst taking into account the influence of the world airline network (WAN) in the WTN. In order to achieve this, we will use complex network theory concepts coupled with a quantitative analysis using methods and concepts from statistical physics allowing us to have a quantitative description of this network and the business that supports it.

Chapter 1

Network Topology

1.1 Introduction

Complex networks are ubiquitous and complex network theory is used in numerous fields to represent different systems, such as metabolic pathways [14], food webs [15], taxonomic classifications [16], e-mail networks [17] and stock ownership [18], just to name a few. We can classify them into several different types of networks based on their structure, and the structure of each network is important to its functional and dynamical behavior [19–21]. Throughout this chapter, we will introduce the data collection and representation methods used in order to build the World Tourism Network (WTN) and characterize its structure using several network related metrics.

1.2 Data Collection and Representation

1.2.1 Data Collection

The datasets used in this work to build the World Tourism Network (WTN) were provided by the United Nations World Tourism Organization (UNWTO) and consist of data files for each territory containing the origin and number of "arrivals of non-resident tourists at national borders" from 2004 to 2008 (inclusive). These overnight visitors are visitors who travel to a country other than that in which they have their usual residence, but outside their usual environment, for a period not exceeding 12 months and whose main purpose in visiting is other than an activity remunerated from within the country visited. The sources and collection methods differ across countries. In some cases, data is from border statistics (police, immigration) and supplemented by border surveys. In other cases, data is from tourism accommodation establishments.

The data refers to number of arrivals, not to the number of people travelling, but since these should not differ by much, we consider that it represents the number of people travelling from their country of residence to another with a touristic purpose. The dataset only considers overnight visitors, so if a person from country i has country j as its touristic destination and takes a connecting flight in country k, but does not stay there overnight, it will only count as a tourist in country j, and not in country k.

These datasets also cover several years, allowing us to study the time-dependent properties throughout a number of years. We have to mention that the analysis done in this work is based on a network that aggregates the information from all years (2004-2008). The data for the World Airline Network (WAN) used in this analysis was provided by *OpenFlights* [22], and it contains thousands of airports around the world and around 60000 airplane routes.

1.2.2 Complex Network Representation

The links of a complex network can be classified by the reciprocity between their attaching nodes, i.e., if there is a two-way connection between the nodes, the link joining them is undirected but, if the direction of the connection is relevant, the link is referred to as directed. Furthermore, links can also be labeled weighted or unweighted, where the weight is usually the magnitude of a property of the links [23] which can be, for example, the number of people from country i who travel to country j to do tourism, as is the case for the WTN. Although in this work we only deal with weights in terms of the links, nodes can also have weights [24].

In order to study the network in a quantitative way, we first represent it using the adjacency matrix formalism, which can be used to represent all types of networks used in this work, be it undirected or directed networks, and also unweighted and weighted networks.

The adjacency matrix for an unweighted network, a network whose links all have the same magnitude, is defined as

$$A_{ij} = \begin{cases} 1, & \text{if node } i \text{ is connected to node } j \\ 0, & \text{otherwise} \end{cases}. \tag{1.1}$$

Directed weighted networks can be represented by either an unweighted or a weighted adjacency matrix, depending on whether pairs of nodes have different strengths of interaction or not. The adjacency matrix of a weighted network is a matrix whose elements represent the weight of the interaction between the nodes, and can be defined as

$$A_{ij} = \begin{cases} w_{ij}, & \text{if node } i \text{ is connected to node } j \\ 0, & \text{otherwise} \end{cases}, \tag{1.2}$$

where w_{ij} represents the weight of the interaction between nodes i and j.

In our case, the weights represent the number of tourists travelling from country i to country j. Also, since the number of tourists shared between countries i and j is not the same, this matrix is asymmetric.

Most works studying the WAN consider airports as nodes and the routes between them as links [8]. In this work, we consider the countries present in the WTN (World Tourism Network) as nodes of the WAN and say that there is a link between them, in the WAN, only if there is at least one connection between an airport of one country to an airport of another. In the WAN, we assume that if a flight exists from country A to B, so must exist a connection in the reverse direction, and through the same path, thus making the adjacency matrix for this network symmetric. In section 1.3, we will study the structural properties of the WTN and as such, we will deal with its unweighted version.

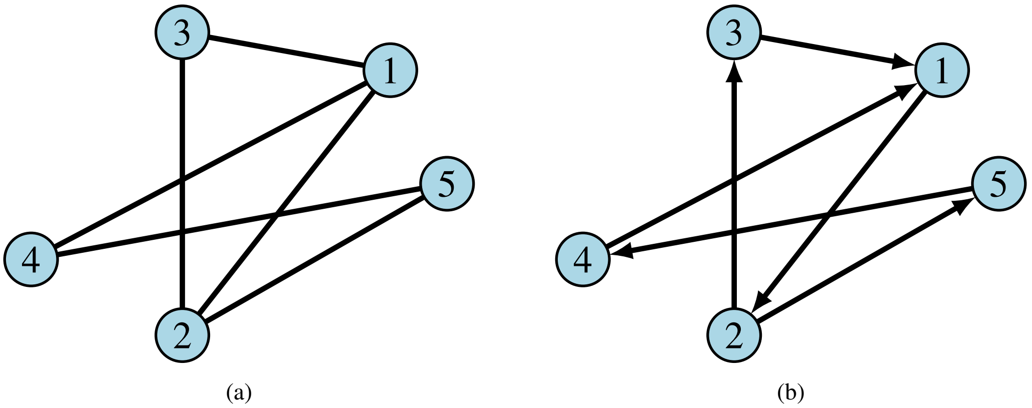


Figure 1.1: Representation of undirected (a) and directed (b) networks.

The adjacency matrix for the undirected network in Fig. 1.1 is:

$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \end{bmatrix}, \tag{1.3}$$

whilst the adjacency matrix for the directed network is:

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{1.4}$$

1.3 Network Metrics

The WTN has N = 214 nodes (countries) and L = 4148 directed links (direction of flow of tourists) with a sparse link density of

$$\frac{L}{N(N-1)} = 0.091, \tag{1.5}$$

with only 9.1% of possible links enabling tourism. In complex network theory, the degree of a node is defined as the number of nodes that share a link with the original node. For directed networks, one can specify, and measure the in- and out-degree of a node. The in-degree of node i is the number of links starting in other nodes that end in node i while out-degree means the number of links starting at node i. In this case, the out-degree of a country is the number of countries its inhabitants travel to, and the in-degree of the same node is the number of countries that send tourists to it. The average out-degree of the WTN, i.e., the average number of countries a country sends its inhabitants to, is

$$< k^{out} >= \frac{\sum_{i}^{N} k_i^{out}}{N} = 19.38\ , \tag{1.6}$$

with a standard deviation value of $\sigma_{k^{out}} = 11.67$. In Eq. 1.6, k_i^{out} represents the out-degree of country i. Since all links in the WTN have a start and end node, the total in and out-degree of the network are the same, as are their average value. The distribution of the *out-degree* metric for the WTN is shown in Fig. 1.2.

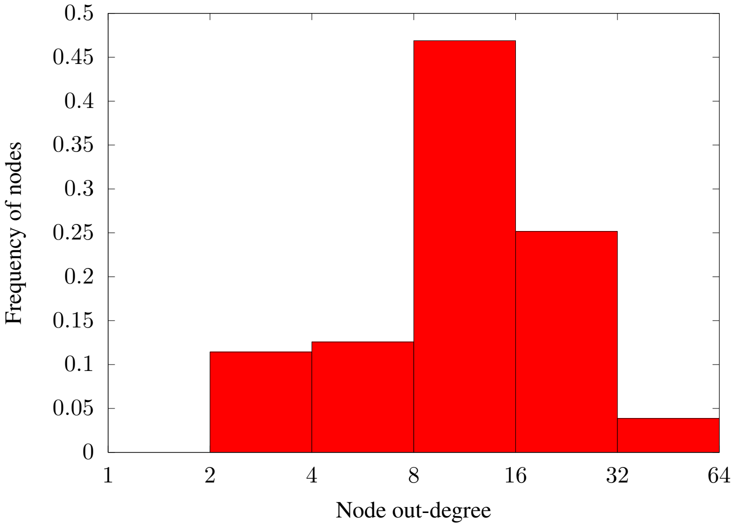


Figure 1.2: Logarithmic binning of order 2 for the distribution of the out-degree parameter in the WTN. We see the peak of the out-degree distribution for values between 8 and 16.

In order to better understand the topology of the network, one of the most used measures is the average path length, defined as the average minimal chemical distance between all connected nodes. The chemical distance between two nodes is the number of links needed to cross to go from one node to the other. In our abstracted version of the World Airline Network, for which we measure this parameter, the chemical distance between two nodes corresponds to the number of between-country flights one person has to take in order to go from their country to the destination one, where we assume that two countries are at a chemical distance of 1 if one airport of the origin country has a direct flight to an airport in the destination country. We say that this measure is minimal because, between two different nodes of a network, there is typically more than one way to connect one to the other. In here, we only care about the smallest of these paths because this is the path that is typically chosen by the users of the network, since it is the one with lowest cost.

The average path length for the WTN is given by:

$$l = \frac{\sum\limits_{ij} d_{ij}^{c}}{L},$$

where d_{ij}^{c} is the chemical distance between any two nodes of the network and L is the total number of links in the network. The value of the average path length for the whole network is $l = 2.32$ with a standard deviation value of $\sigma_l = 0.69$. The distribution of the path lengths for the pairs of nodes in the WTN is shown in Fig. 1.3. This result means that, on average, one person has to take 2.32 flights in order to go from its country to any other. It is a smaller value than what has been previously recorded [8], but this difference can be justified by the coarse-graining done in order to obtain the WAN used in this work and also by the fact that the network studied in Ref. [8] takes into account paths from any airport to any other as long as such a path is possible, while for our version of the WAN, we only count paths which are already used by people who are travelling to their touristic destination, and thus we are using smaller paths.

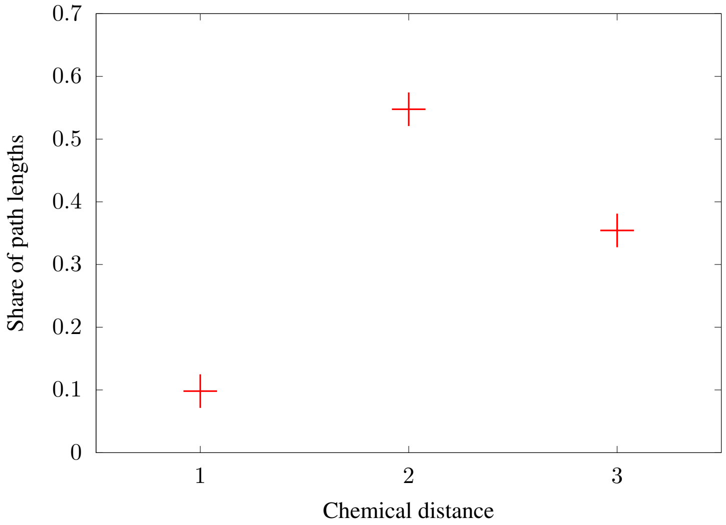


Figure 1.3: Distribution of path lengths in the WTN.

Another common measure is the reciprocity, which quantifies the relations between links belonging to a directed network which work in both directions. This can be used to assess the relationships between two different countries in terms of their economic capacity. The reciprocity coefficient is defined as [25]

$$r = \frac{1}{L}\sum_{ij} A_{ij}A_{ji}, \tag{1.7}$$

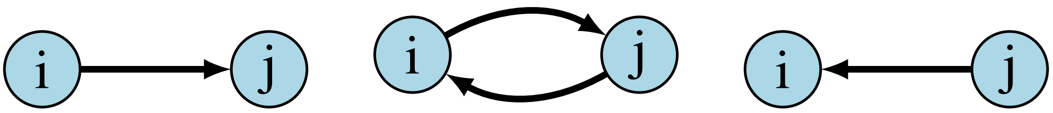
where L is the number of links in the network and $A_{ij} = 1$ if there is a link between nodes i and j or equal to 0 otherwise. The value of the reciprocity for the WTN is $r = 0.15$, meaning that for each link connecting two nodes in one direction there is a 15% chance that we have a link in the opposite direction. This value can be attributed to the heterogeneity of the economies of the countries represented in the WTN. The WAN, on the other hand, has a value of reciprocity of $r = 1$, by construction, since we assume that all flights in this network are reciprocated.

1.3.1 Clustering coefficient

We now investigate how well connected the WTN is, more specifically, how the neighbors of a node are connected among them. To quantify this, and in a more broad sense, how well connected the network is, the clustering coefficient is used. This coefficient is defined as the fraction of pairs of nodes to which a node is connected, that are connected among them. The WTN not only transfers tourism wealth from one country to another directly but also in a cyclic indirect manner. If a traveler moves from their country of residence, they contribute to the economy of the place they travel to, which in turn accumulates and contributes to the economy of another country in part if there is movement on that segment of the cycle.

For a given node in a undirected unweighted network, the number of possible links that exist between its neighbours is $\frac{n_i(n_i-1)}{2}$ and so the probability of two neighbours of i being connected is the number of links between neighbours divided by $\frac{n_i(n_i-1)}{2}$. Thus the topological clustering coefficient of a node i in an undirected network is defined as,

$$C_i = \frac{2L_i}{n_i(n_i-1)}, \tag{1.8}$$



(a) Direct link from i to j.

(b) Links in both directions for nodes i and j, also called a repeated link.

(c) Direct link from j to i.

Figure 1.4: Link possibilities in a directed network between nodes i and j.

where L_i is the number of links that exist between the first neighbors of i [26], the nodes which are at a chemical distance of one from it.

Using the adjacency matrix to express the last equation, Eq. 1.8, we can write the clustering coefficient for a node of an undirected network as

$$C_i = \frac{2}{n_i(n_i - 1)} \sum_{j,k \in \{n_i\}} a_{ij} a_{ik} a_{jk}. \tag{1.9}$$

If we want to calculate the clustering coefficient in directed networks however, we have to take into account the direction of each link [27]. The possible links we have between two nodes in a directed network are presented in Fig. 1.4 and they influence the total number of possible links between two nodes. The total number of possible links between neighbours of node i, for a directed network is defined as:

$$n_i^D = k_i^{tot} - L_i^{\leftrightarrow} , \tag{1.10}$$

where k_i^{tot} represents the sum of the in- and out-degrees of node i, the total degree, and $L_i^{\leftrightarrow}$ the number of repeated links of node i, meaning the number of links that grow out of i that have a corresponding link in the reverse direction, as is represented in Fig. 1.4b.

Another thing to take into account is that whilst for undirected networks, there is only one possible triangle formed between each triplet of nodes, for directed networks that number increases to 8, as shown in Fig. 1.5. In this work, we only consider the cycle-triangles, represented in Fig. 1.5a, for the calculation of the clustering coefficient since these triangles indicate that there is a unidirectional cyclic flow of tourism wealth among countries, which means that these cliques unintentionally transfer wealth back to themselves, thereby all countries retaining a portion of their tourism economic value spent abroad (see Fig. 1.5).

Taking into account these two factors, we introduce here the *cyclic clustering coefficient* which is defined as:

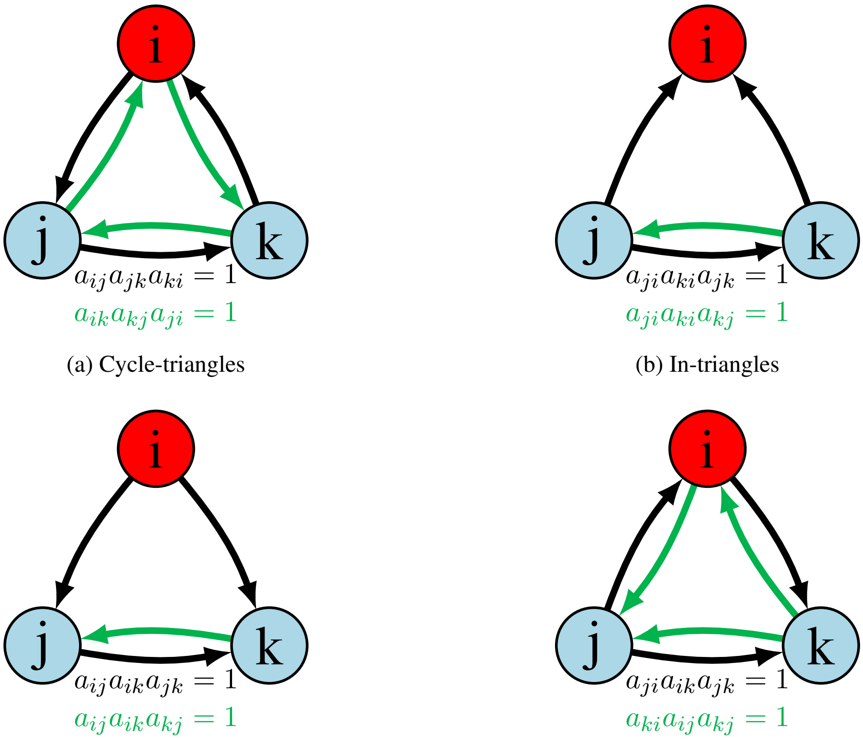
$$C_i^{cyc} = \frac{N_i^{cyc}}{n_i^D(n_i^D - 1)} , \tag{1.11}$$

where N_i^{cyc} represents the number of cycle-triangles involving node i.

We expect the probability of occurrence of cycles in a directed network to be of the order of the link density, since if a node is connected to another two, the latter only need to form a link between themselves (in the right direction), in order to form a cycle.

The average value of cyclic clustering coefficient for the nodes of the WTN is $< C^{cyc} > = 0.015$, which is of the order of the link density, as expected. It is, though, of a small order for a network built on top of the airline network, which itself has a high value (C=0.62[7]).

If we discount the directionality of the links in the WTN and calculate the clustering coefficient



(a) Cycle-triangles

(b) In-triangles

(c) Out-triangles

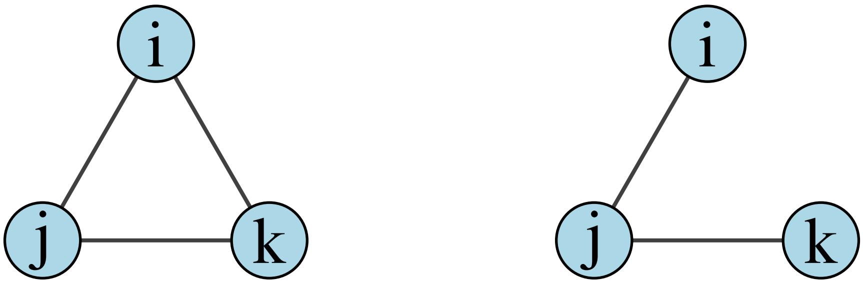
(d) Bridge-triangles

Figure 1.5: Triangles in directed networks and their designations. Each type of triangle corresponds to a different product of the adjacency matrix which is shown.

value, given by Eq. 1.9 we obtain $< C^{und} >= 0.79$, a much larger value than for the directed version of the network and even bigger than the one for the WAN. This difference sheds a light on the profile of touristic travel around the world, because it tells us that even though most countries are well connected, this characteristic of the network breaks down when we add directionality, meaning that most of the links of the network are not reciprocated, as it is evidenced by the reciprocity value.

1.3.2 Triangle decomposition

We are interested now in the study of cliques of size three, which are triplets of nodes that are all naturally connected, forming a triangle. Cliques are important in transportation networks because as one sees in Fig. 1.6, the ability to move to the desired destination inside a clique, even after an internal or external event causes a failure in one of the links, is still guaranteed although the effort to do so is now greater.



(a)

(b)

Figure 1.6: Triplets of nodes in an undirected network. On the left, we have a clique of size 3, while on the right there is one link missing. We can see that although we have one link missing, we can still get from any point in the network to any other.

For directed networks, the links have directionality which increases the number and type of cliques we can have. In order to separate from cliques in undirected and directed networks, we define a new concept, *triangles*, which are the cliques present in directed networks. We can see, by Fig. 1.5 that a node i can be part of 4 types of triangles: *cycle-triangles*, triangles where one can start at a node and travel back to it without having to reverse the direction of travel; *in-triangles*, where a node i has incoming link from each of the other nodes in the triangle; *out-triangles*, similar to in-triangles but now node i has outgoing links to the other nodes; and *bridge-triangles*, where node i has an incoming links from one node and an outgoing to another, acting as a "bridge" between the other nodes in the triangle.

The cycle/non-cycle distribution for the WTN is presented in Table 1.1. From this table, we can see that the percentage of cycles in the network is about 6%. We looked at the distribution of triangles for the network and also into the percentage of countries, for each triangle, which took part in that type of triangle, naming this measure as triangle participation percentage. In order to have a benchmark to which we can compare these results, we use the *configurational model*. The configurational model consists in randomly rewiring all the links in a network in order to remove any node-node correlations and whilst keeping the degree distribution of the original one. The results are shown in Figs. 1.7 and 1.8 and in Tables 1.2, 1.3, 1.4, and 1.5.

Table 1.1: Cycle/Non-cycle triangles distribution for the WTN.

Cycles	Non-cycles
5.89%	94.11%

Table 1.2: Triangles distribution for the WTN, according to their type, where each type is shown in Fig. 1.5.

Cycle	In	Out	Bridge
5.89%	31.37%	31.37%	31.37%

Table 1.3: Average triangles distribution for 100 configurational model networks.

Cycle	In	Out	Bridge
4.66(±0.29)%	31.78(±0.1)%	31.78(±0.1)%	31.78(±0.1)%

Table 1.4: Triangles participation coefficient for the WTN, according to the type of triangle.

Cycle	In	Out	Bridge
32.24%	31.77%	100%	32.24%

Table 1.5: Triangles participation coefficient for 100 configurational model networks, according to the type of triangle.

Cycle	In	Out	Bridge
22.3(±0.5)%	22.3(±0.5)%	99.7(±0.3)%	22.9(±0)%

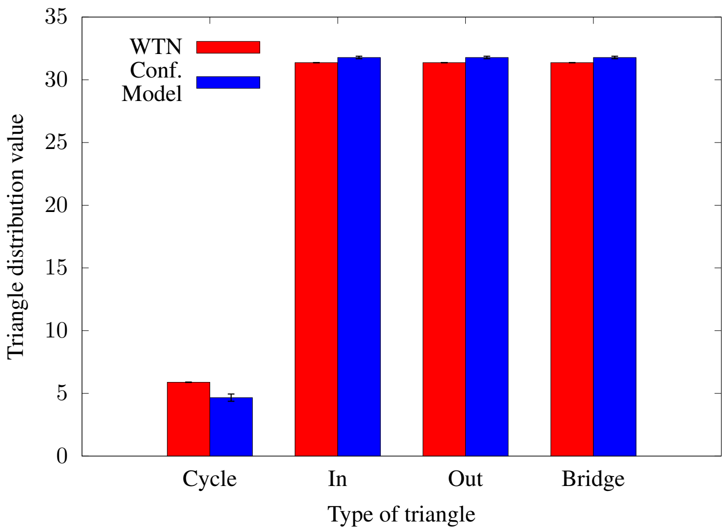


Figure 1.7: Triangle distribution values for the WTN and the configurational model networks.

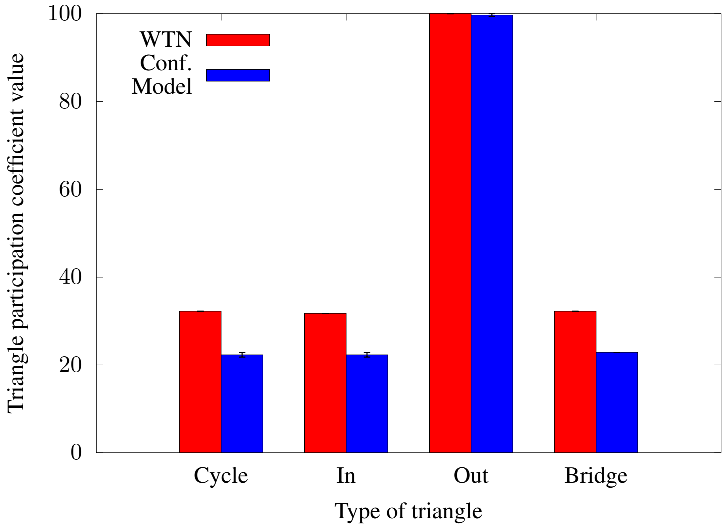


Figure 1.8: Triangle participation coefficient values for the WTN and the configurational model networks.

1.4 Conclusion

These results show that, on average, the nodes of the network connect to a significant number of other nodes in the network and that travelers have to take a couple of connections to get from their territory to the one they wish to visit, based on the average path length of the WAN. On the other hand, the average out-degree value indicates that a low (when in comparison to number of nodes in the network) number of countries are available to travel to and its standard deviation value reflects the disparity in the out-degree values between the most visiting countries and the less visiting ones. This result is surprising but could be explained by the presence of remote territories/countries in the dataset who typically only have flights to a handful of countries and this behaviour contrasts with the more connected countries. The clustering coefficient value is of the expected order, but a big difference occurs when we calculate this coefficient without taking into account the directionality of the links, making the value increase from 0.015 to 0.79. This difference can be attributed to the fact that most of the links of the network are not reciprocated, exposing the unequal behaviour of human mobility in terms of tourism, where people tend to go from richer countries to poorer ones, but not the other way around. Looking at the triangle decomposition section, we note that this network is very similar to a random network, which is a much different result than we expected since this work is based on a real network and one would expect that it did not have much in common with random networks.

Chapter 2

Network Functionality

2.1 Introduction

The dynamics of different processes occurring on networks depend not only on the dynamics of the individual nodes but also on the topology of the interactions. In the WTN, the relevant processes are related to the movement of people from their country to their destination. Such processes are responsible for several business areas and services which support them. In this chapter we will investigate the influence of both the geographical and chemical distances in the way people move through the network.

2.2 Distances in the WTN

When people look for their next touristic destination, they take several factors into consideration. Among those factors is the geographical distance to the destination and also the number of flights one needs to take to get there. The latter is typically referred as chemical distance, and for this analysis, we use the WAN and so the values of this distance can go from 1, a direct flight to 4, the highest number of between-country connections for two connected countries in the WTN. The geographical distance is also accounted for in our analysis and in order to calculate it we use $hav(x)$, given by:

$$hav\left(\frac{d^g}{R}\right) = hav\left(\phi_2 - \phi_1\right) + \cos\left(\phi_1\right)\cos\left(\phi_2\right) hav\left(\lambda_2 - \lambda_1\right), \tag{2.1}$$

where d^g corresponds to the geographical distance between two points, R to the Earth's radius, ϕ_1 and ϕ_2 to the latitude of points 1 and 2 and finally, λ_1 and λ_2 to their longitude, respectively.

The $hav(x)$ function is defined as:

$$hav\left(\theta\right) = \sin^2\left(\frac{\theta}{2}\right), \tag{2.2}$$

and using this last relation we can write equation 2.1 as:

$$\sin^2\left(\frac{d^g}{2R}\right) = \sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\ \cos(\phi_2)\ \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right). \tag{2.3}$$

Inverting this equation, we obtain:

$$d^g = 2R\ sin^{-1}\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos(\phi_1)\ cos(\phi_2)\ sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right). \tag{2.4}$$

Equation 2.4 gives the orthodromic distance between two points on Earth. We use the centroid (geographical center) of each country to make the calculations of the geographical distance between nodes of the WTN.

2.2.1 Tourists per distance analysis

We now want to look at the influence of the chemical distance (in the WAN) and the geographical distance, between two nodes of the network on the number of tourists shared by the nodes.

In order to study the influence of the *chemical distance* on the number of tourists going from one country to another, we calculate, the fraction of tourists as a function of the chemical distance (from 1, a direct flight to 4, the largest chemical distance in the WTN). The distribution of network-wide outgoing fraction of tourism flow (f^t_{ij}) for all chemical distances (d^c_{ij}) is expressed as,

$$f^t_{ij}(d^c_{ij} \in [1,4]) = \frac{F^t_{ij}}{\sum\limits_{j\in\{n_i^{out}\}} F^t_{ij}}, \tag{2.5}$$

where F^t_{ij} is the number (flow) of tourists between the nodes i and j and $\{n_i^{out}\}$ is the set of countries receiving tourists from country i.

Doing this analysis for all nodes of the network, and then averaging over all nodes in the network, disregarding the different number of outbound tourist each country has and therefore the different weights they have on the average, we obtain the average fraction of tourists as function of the chemical distance, for the entire network.

We also looked for the distribution of tourists as function of the geographical distance by dividing the distance between countries into bins of 1000 km each. We calculate the number of tourists in each bin and calculate the fraction in each, with respect to the entire outflow of tourists. To give a better view of both of these distances, we show in Fig. 2.1 a representation of a node of the WTN in our abstracted version of the WAN.

The resulting distributions of the distances are presented in Fig. 2.2 and were calculated using the following functions for the chemical distance,

$$P_f(d^c) = \frac{1}{N}\sum_i \sum_{j\in\{n_i^{out}\}} f^t_{ij}\,\delta_{d_{ij},d^c}\ , \tag{2.6}$$

where d^c represents the different values of chemical distance, and for the geographical distance,

$$P_f(d^g) = \frac{1}{N}\sum_i \sum_{j\in\{n_i^{out}\}} \frac{F^t_{ij}}{\sum\limits_{j\in\{n_i^{out}\}} F^t_{ij}} \int_{d^g}^{d^g+1} \delta(x-d_{ij})dx\ , \tag{2.7}$$

where d^g represents the index of the bins used in the geographical distance division.

In Fig. 2.2 we can see that the fraction of tourists is highest for countries one connection away and decreases monotonically until we reach the highest value for the chemical distance. The result obtained by measuring this distribution indicates that tourists tend to travel to destinations which are more easily accessible in terms of flight connections.

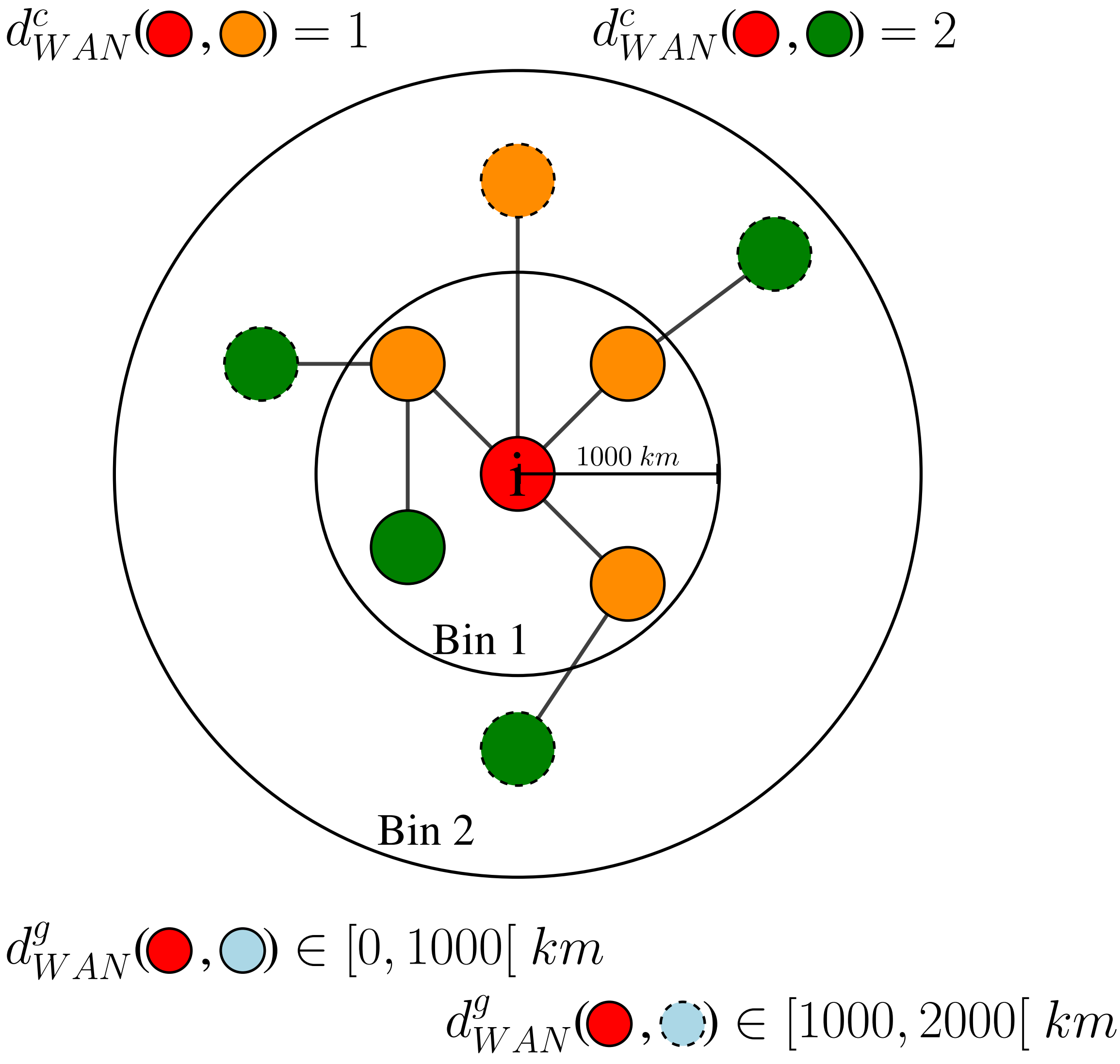

d^c_{WAN}(●,●) $= 1$
d^c_{WAN}(●,●) $= 2$
i
1000 km
Bin 1
Bin 2
d^g_{WAN}(●,○) $\in [0, 1000[\ km$
d^g_{WAN}(●,○) $\in [1000, 2000[\ km$

Figure 2.1: Representation of node i and some of its neighbors, in the WAN network. All the nodes in this figure are neighbors of node i in the WTN, but they are at different chemical distances in the WAN. Orange colored nodes are at a chemical distance of 1 from node i and green nodes are at a distance 2, meaning that people from node i have to take two between-country flights to reach those nodes. The figure also shows two circles, of 1000 and 2000 km in radius, respectively, which represent the binning used in the distributions related to the geographical distance. Solid circle nodes belong in the first bin while nodes represented with dashed circles belong in the second one.

For the geographical distance, we can see that the peak of fraction of tourists is in the [0,1000[km interval, and after that, it follows an expected distribution with the fraction of tourists decreasing at each bin of the histogram, with some exceptions. To note that the first two bins have very similar values, which means that people are almost as likely to travel to a destination that is at most 1000 km away as they are to a destination between 1000 and 2000 km away.

2.2.2 Links per distance analysis

We now look into the distribution of the number of flights (links) that connect countries, in the WAN, between nodes that share tourists. We proceed the same way as above, and calculate, for the whole network, the fraction of touristically-connected countries as a function of the chemical and geographical distances. The result is shown in Fig. 2.3.

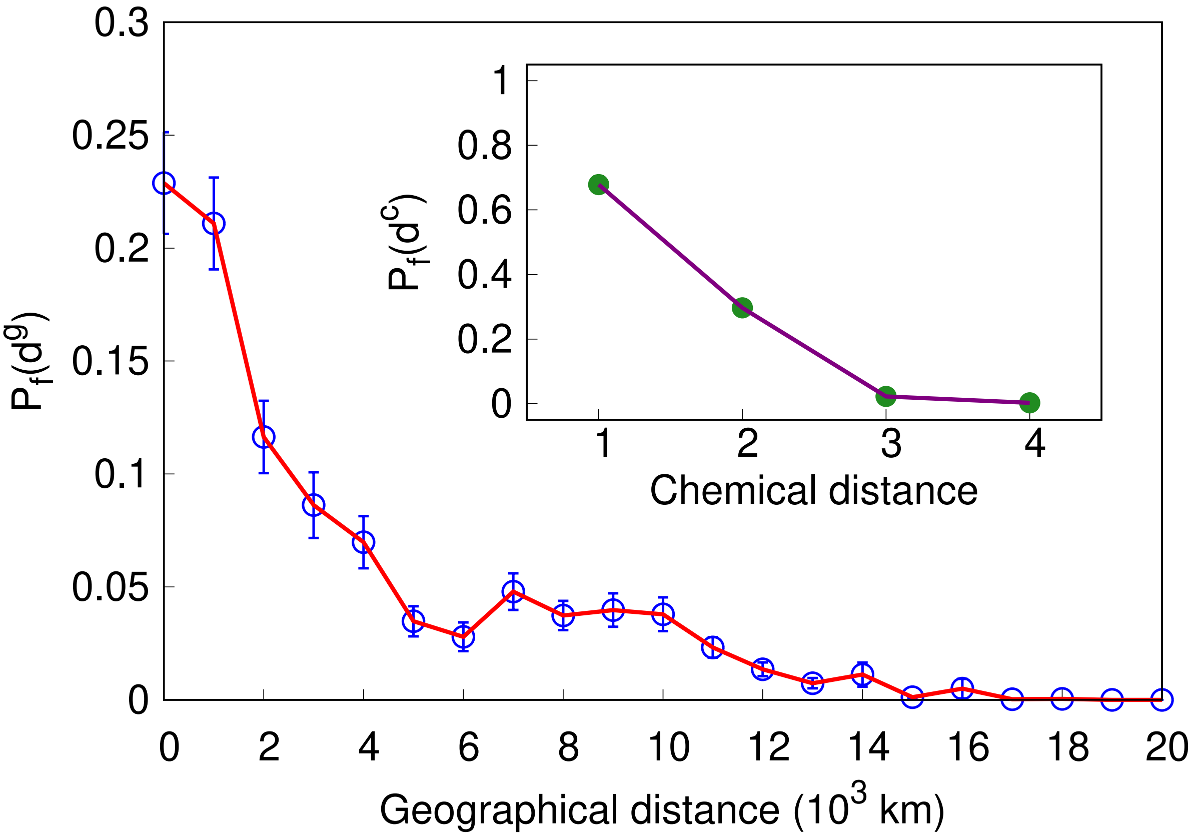


Figure 2.2: Geographical and chemical (inset) distance distributions of outgoing tourists in the WTN.

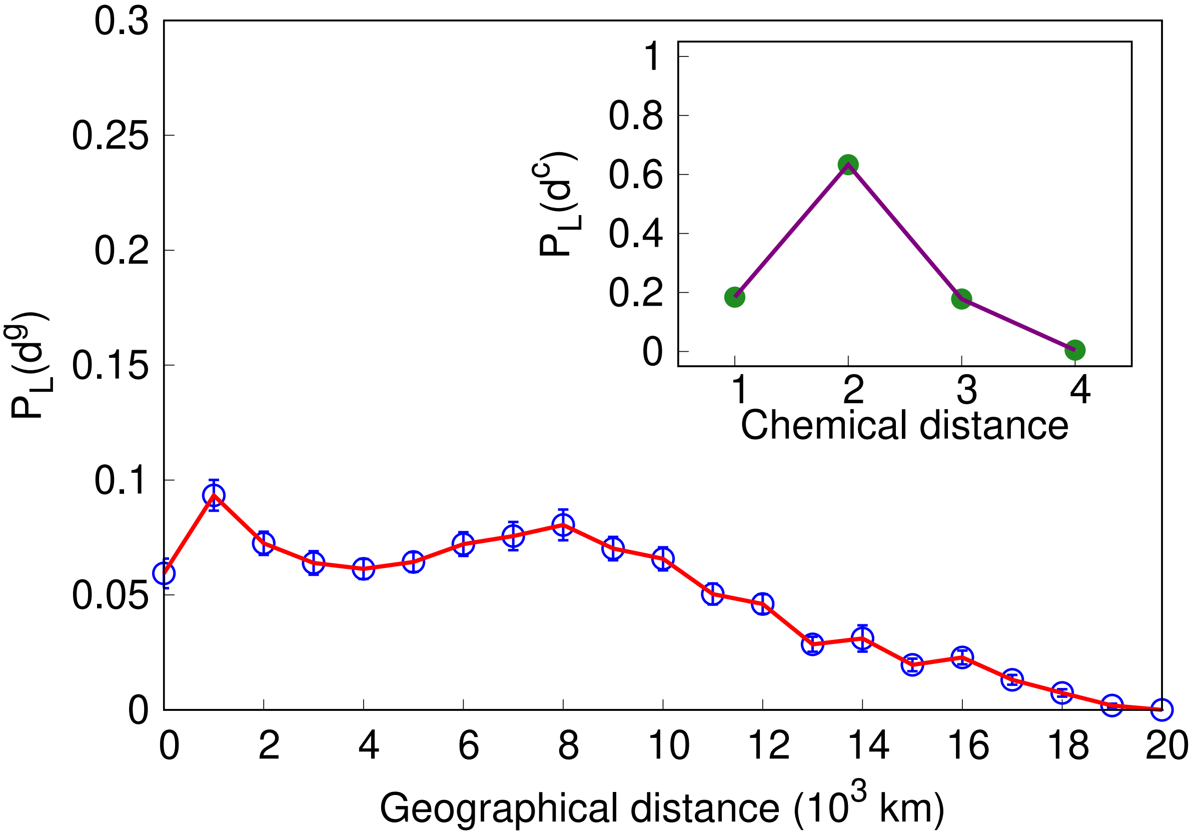


Figure 2.3: Geographical and chemical (inset) distance distributions of the touristic destinations of the WTN.

The behaviour shown in this figure suggests that most of the links in the WTN are formed between countries two or more flight connections apart from each other, and more specifically, the more common number of connections is 2, indicating that even though the number of links with connections to all countries on average is 2, typically a lot more travellers fly to countries that are directly connected for tourism. Geographically speaking, the link distribution does not exhibit a large difference in the values, at least for the $[0, 11000[$ km range, as touristic flow distribution, indicating that there is still some variety in choice of travel to farther countries.

2.2.3 Passengers and flights correlation

Figures 2.2 and 2.3 show the average distributions of touristic flow and links between countries for various distances and we can see that the peaks of both distributions for the chemical distance do not occur at the same value. The peak for the passenger-distance distribution is at a chemical distance of 1 and the peak for the distribution of flights' chemical distance is at 2. This means that most amount of traffic moves on direct links and lesser amount of traffic has a rather large subset of links available due to a variety of travelling behavior. However, the relationship between each individual country's travel preference in terms of the two perceptions of distance (chemical and geographical) is unclear. Figure 2.4 compares the chemical distance at which the distributions of touristic flow and links maximizes for each country. From the figure, we can see that there seems to be no correlation between the peaks of each country. We can also see that there is a large percentage of network nodes for which the touristic connections peak has a value of 2 and a tourist flow peak of 1, indicating that on average, the most visited destination per country is typically one flight away although most flights available in a given country to destinations which receive tourists are two connections away. In this sense, the network, and some countries specifically, those whose value of chemical distance of the maximal touristic flow is 1 and of touristic connections is 2, could be optimised by taking some flights that take two connections to complete and which carry a significant fraction of tourists, and transforming them into one connection flights.

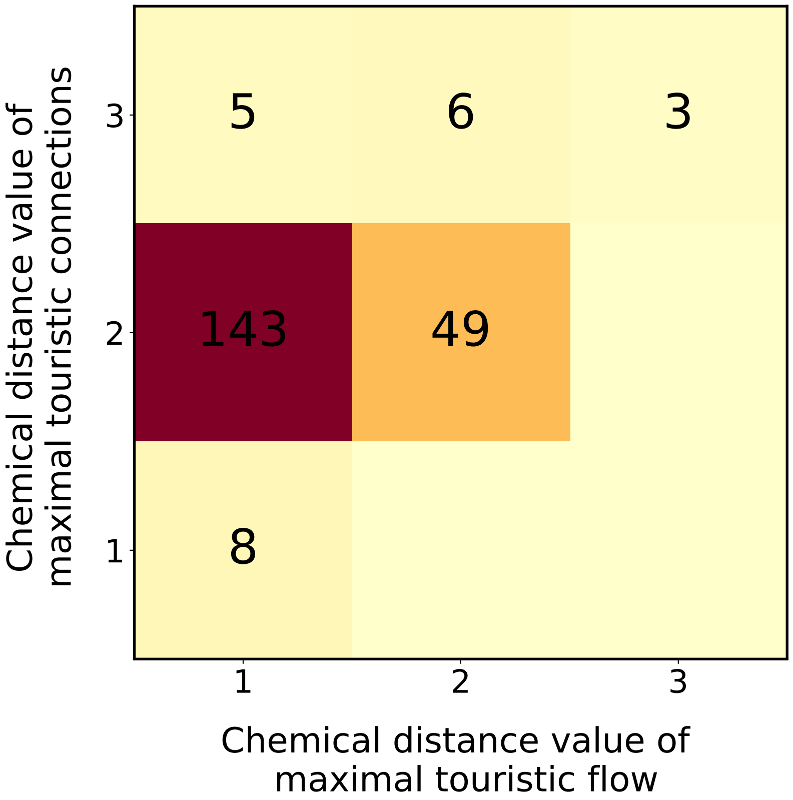


Figure 2.4: Heat map indicating the relationship between chemical distances for which the touristic flow and number of touristic connections have their maximum for the WTN.

We also see that the peaks of the geographical distributions do not match and that they are very different from each other. We wonder if there is a relationship between the peaks in each distribution and in order to verify that, we plot a heat map of both distribution and the result is shown in Fig. 2.5. Since most tourists per country prefer a distance in the range $[0, 2000]$ km on a direct link, there are not enough data points beyond that distance range to report anything statistically significant.

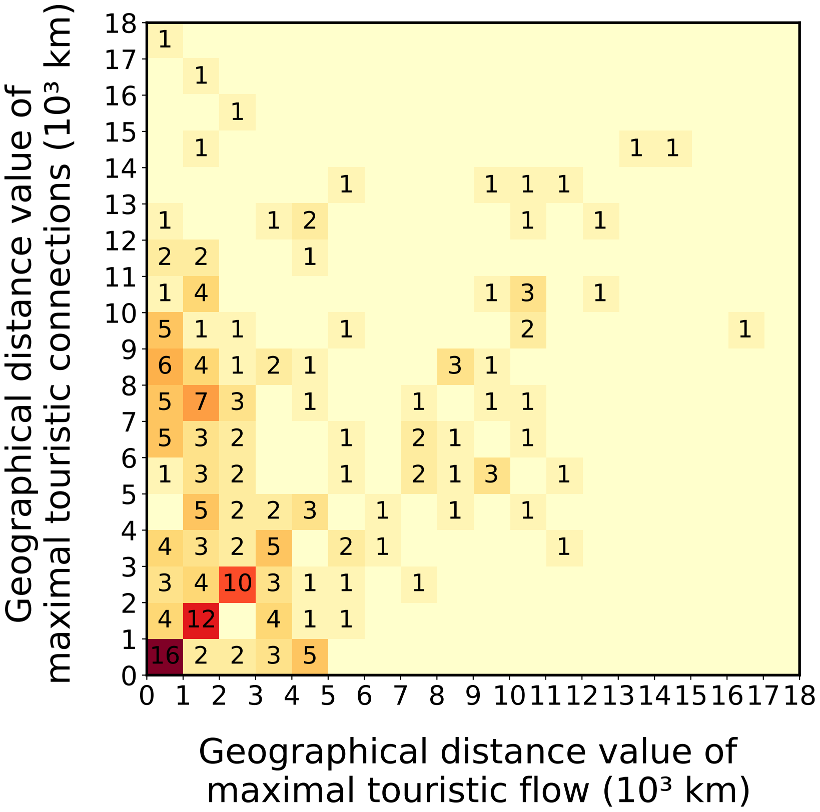


Figure 2.5: Heat map indicating the relationship between geographical distances for which the touristic flow and number of touristic connections have their maximum for the WTN. The squares without a number correspond to non-registered occurrences.

2.3 Connection strength evolution

Having studied the functionality of the WTN, we also wanted to take a look at its dynamical evolution. Since our data covers several years (2004-2008), one of the first things we can study is the evolution of the strength of connections. A simple measure of strength of evolution of tourism flow α_{ij} between a pair of countries i and j,

$$\alpha_{ij} = \frac{F^t_{ij}(t_f) - F^t_{ij}(t_0)}{F^t_{ij}(t_0)}, \tag{2.8}$$

where t_f and t_0 correspond to the end and start of discrete-events in the network dataset (2008 and 2004, respectively), shows that large changes in flow have occurred for a very few connections over the years (trailing tail) and $\approx 30\%$ of the links have shown no change in traffic. The results are presented in fig. 2.6.

The higher values for positive α in fig. 2.6 tells us that large changes have occurred and that the growth of links has been gradual, since most of the connections have an α value between $\alpha = 0$ and $\alpha = 1$. Another representation of the alpha parameter values can be shown in Fig. 2.7 where we have a donut chart showing the evolution of tourism in the years present in this work (2004-2008). Finally, Table 2.1 shows the top and bottom five connections in terms of their α parameter value. These α values could be considered outliers due to their difference to the mean.

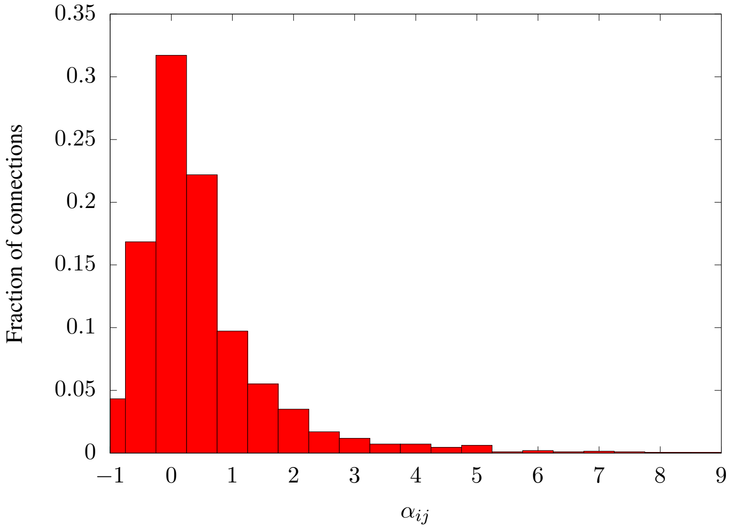


Figure 2.6: Fraction of connections in the WTN in terms of their α parameter value with $t_0 = 2004$ and $t_f = 2008$.

Table 2.1: Top and bottom five values of the α parameter.

Origin country	Destination country	α_{ij}
Macedonia	Israel	881
Laos	Cambodia	35.75
Tajikistan	Kyrgyzstan	35.23
Rwanda	Ukraine	26.11
Uzbekistan	Kyrgyzstan	17.66
Kuwait	Israel	-0.93
Georgia	Trinidad and Tobago	-0.96
United Arab Emirates	Israel	-0.96
Saudi Arabia	Israel	-0.99
Macau	Malaysia	-0.99

2.4 Conclusion

This chapter provides an analysis of how both the chemical and geographical distance influence the choice of country to visit. In the chemical distance analysis, we found that people mainly want to go destinations close, in terms of connections, to their country. Another significant result was the difference in the peaks of touristic flow and touristic connections for the entire network where we see several countries whose airline network could be optimized in order to better respond to the needs of touristic passengers.

In terms of geographical distance, our analysis reveals that people prefer closer destinations and also that the flight share has very similar values for flight in the $[0 : 11000]$ km range. Also, the analysis of the maximums for this distance tells us that there is some correlation in the $[0 : 3000]$ km interval but not much after that.

From the study of the correlation between the maximal touristic flow and connections values, we see that the WTN can be optimized with regards to tourism, if not during all the year, during high demand seasons.

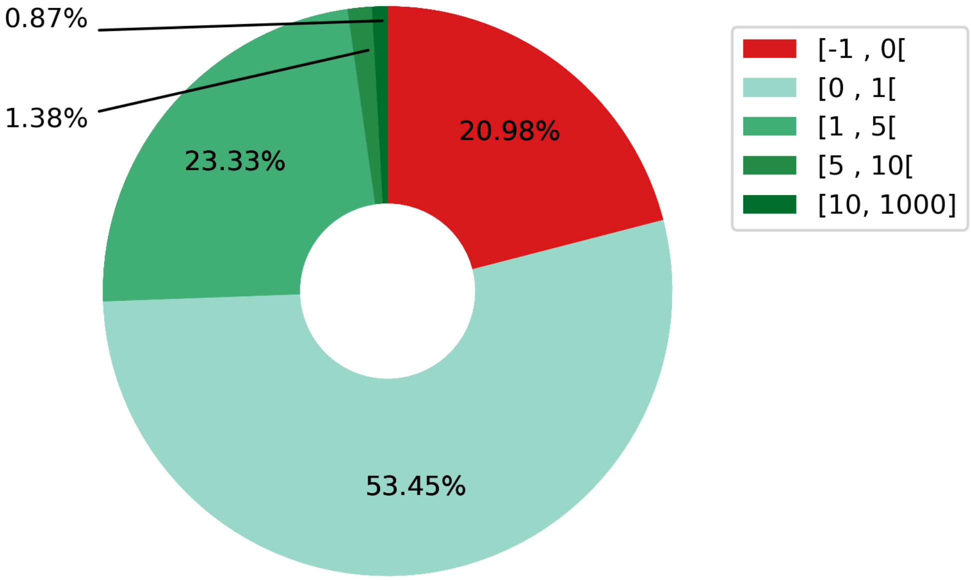


Figure 2.7: Alpha parameter values distribution.

The analysis of the evolution of the connection strength has allowed a glimpse into the growth of the network. We can see that the network's evolution has been gradual and that large changes have occurred few times, a sign of the constancy in people's touristic choices in the years this work covers (2004-2008).

Chapter 3

Structure & Functionality Interplay in the WTN

3.1 Introduction

After getting a better idea about the structure and functionality of the WTN, we now wish to focus on the interplay between both of these aspects of the network, for example, if there is any relationship between them and also how the structure of the network enables its functionality.

In order to do so, we will dwelve deeper into the triangles of the network, a structural aspect of it, and the number of passengers in each link, a sign of the functionality of the network. We also introduce and discuss a parameter related with the balance of the links, in order to better characterize the movement of people throughout the network.

3.2 Node Structural Balance

The network is not in balance, i.e. the reciprocity of tourism is not bidirectional and equal. The disparity in the flow distribution however, is not a product of the link distribution because they are formed over the WAN which is reciprocal. In order to study the structural balance of the network, we propose a new parameter, called the *T* parameter, defined as,

$$T_i = \frac{k_i^{in} - k_i^{out}}{k_i^{in} + k_i^{out}}, \tag{3.1}$$

where $k_i^{in/out}$ corresponds to the in/out-degree of country i. By definition, the *T* parameter values are located in the [-1,1] interval in which a value of -1 for this parameter indicates a country which only has outbound links while a value of 1 indicates a country which is only connected through inbound links, without any outbound ones.

A map with the distribution of the values of the *T* parameter is shown in Fig. 3.1 and its histogram is showed in Fig. 3.2. The map gives us a better view of the structure of the network, because it tells us most countries are in the region between -1 and -0.75 in terms of their *T* parameter value, which indicates that most countries are sending tourists to more countries than from where they receive.

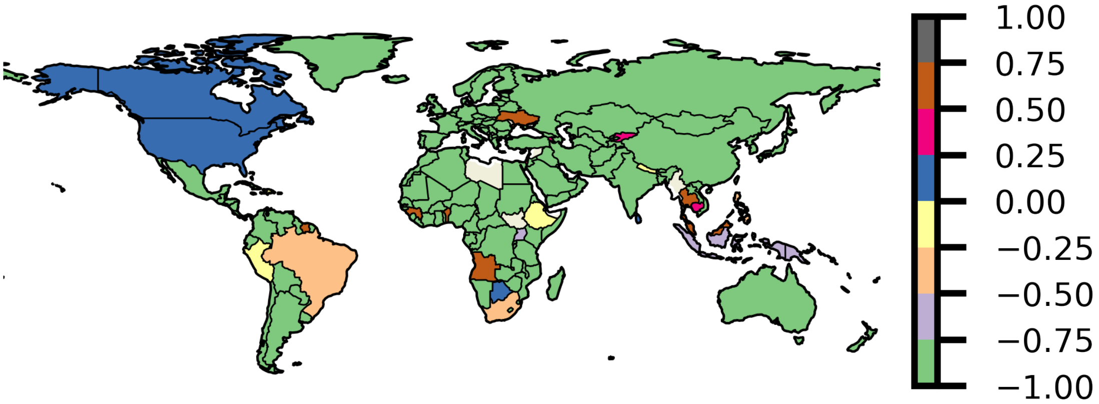


Figure 3.1: Color map for the T parameter values of the WTN. We can see that most countries have a low T parameter value, which means that they are mostly tourist-sending countries. The countries for which we have no data are represented in grey.

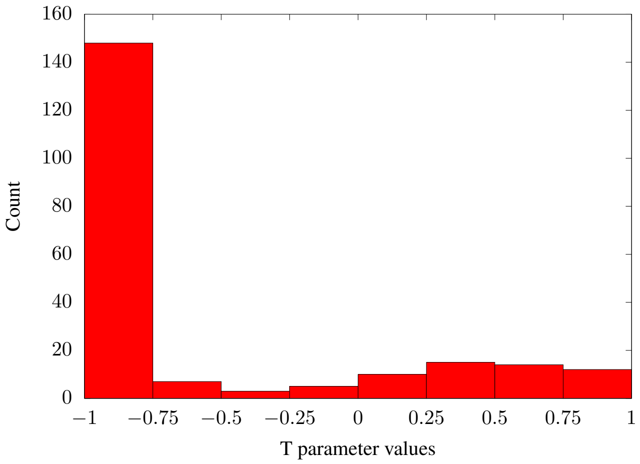


Figure 3.2: Distribution of the T parameter values in the WTN. We can see a clear domination of the values between -1 and -0.75 indicating that most countries don't possess any inbound links.

3.3 Node Functional Balance

After studying a property associated with the structure of the network, we now focus on its functionality and look into the tourist balance of each node in terms of both outgoing and incoming tourists, and for that we introduce a new parameter, called *P* parameter defined as

$$P_i = \frac{F_i^{in} - F_i^{out}}{F_i^{in} + F_i^{out}}, \tag{3.2}$$

where $F_i^{in(out)}$ is the total number of tourists coming to (going out of) node i, also called its in(out)-strength. This parameter can take values in the interval [-1,1]. P = -1 for countries that only have outgoing tourists. The other extreme case, P = 1, is for countries that only receive tourists. The distribution of this parameter throughout the network can be seen in Fig. 3.3 with the corresponding histogram presented in Fig. 3.4. These figures show that most of the countries have a low value of *P* which means they mainly have outbound tourists.

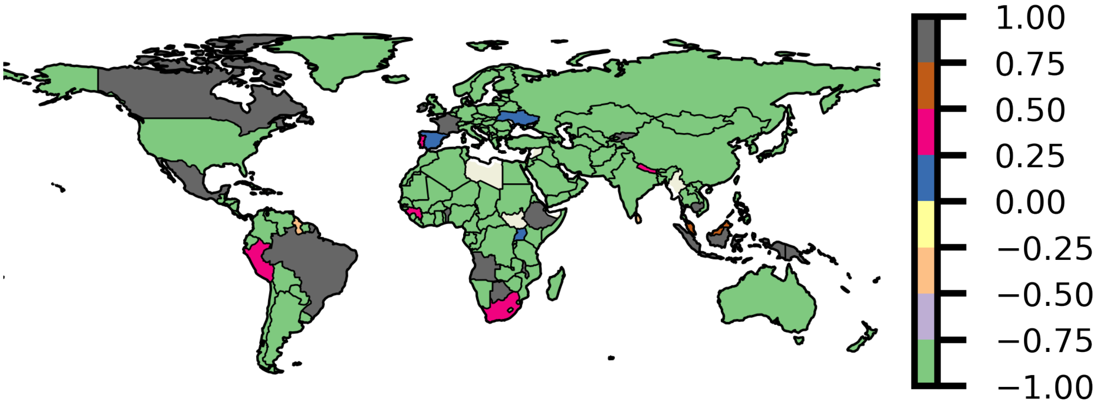


Figure 3.3: Color coded map of the countries in the WTN for the P parameter. The majority of countries are in the [-1,-0.75] region indicating that most of the nodes in the countries don't receive any tourists, although we now can clearly see countries which belong to the topmost subinterval which correspond to high values of the P parameter. The countries for which we have no data are represented in grey.

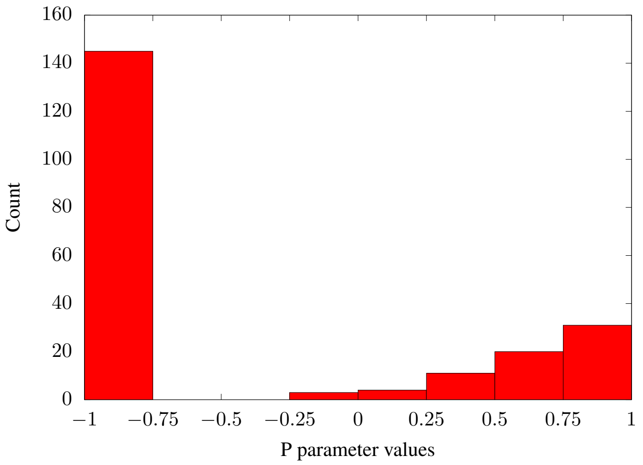


Figure 3.4: Distribution of the P parameter values in the WTN.

After having looked into the T and P parameter separately, we now want to see if there is any correlation between these two parameters, which if true, could be indicative of a relation between the structure and the functionality of the WTN and also about how the airline transportation network, WAN, enables the mobility of tourists between countries.

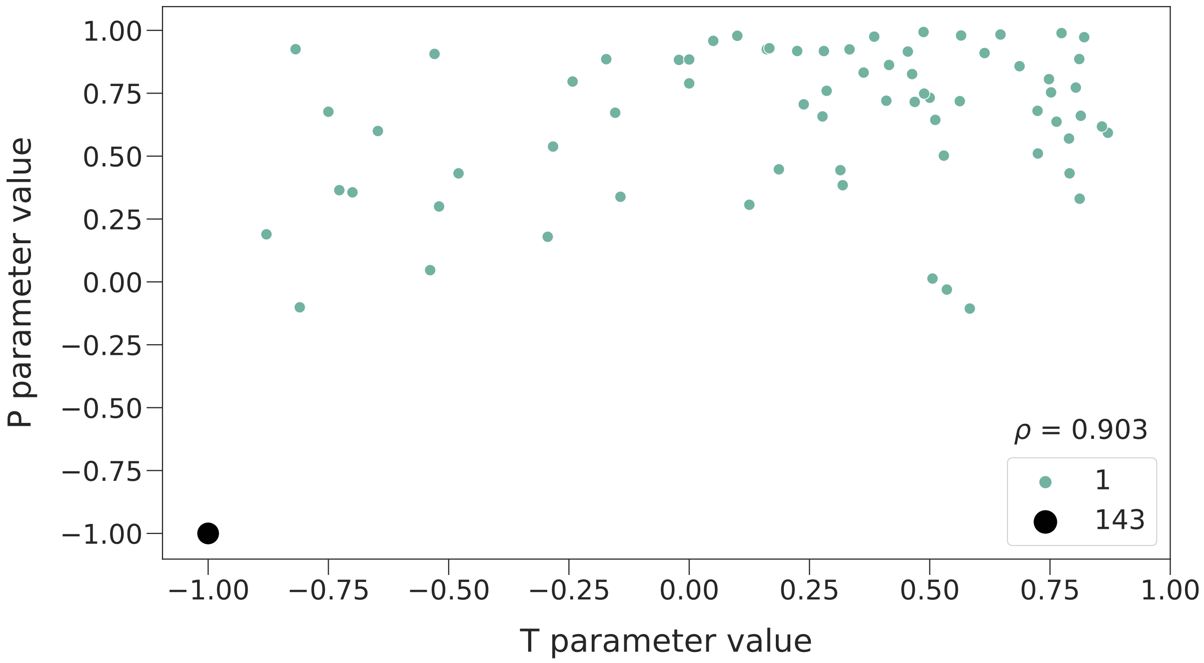


Figure 3.5: Scatter plot of the T and P parameter values for the WTN. The correlation coefficient for this parameters is 0.903. The size of the points in the plot is proportional to the number of pairs of points of T and P parameters values, as is shown in the key.

Figure 3.5 shows the correlation between the T and P parameters, showing that, except for the trivial case (-1,1), there is no strong correlation between the two parameters.

3.4 Network Balance

We now want to know the touristic balance of the links in the network. For each country, we want to evaluate if it is more of tourist-receiving or tourist-sending country. To do so, we define the *network balance* parameter which is given by

$$B_i = \frac{1}{N_i^{pairs}} \sum_{\substack{j \\ j\in\{n_i^{out}\} \\ j\in\{n_i^{in}\}}} \left| \frac{F_{ij}^t - F_{ji}^t}{F_{ij}^t + F_{ji}^t} \right| \tag{3.3}$$

where N_i^{pairs} represents the number of pairs of connections that node i belongs to. Equation 3.4 calculates, for each pair of links which are bidirectional, the absolute value of the fraction between the difference in incoming and outgoing tourists for each node in that link. A value of 0 for the *network balance* parameter value corresponds to a country which sends as many tourists as it receives in its bidirectional links, while a value of 1 is a sign of a country which either mainly only sends or receives tourists in its bidirectional links. In Figs. 3.6 and 3.7 we can see that most countries either do not have another country with which to exchange tourists and also we see that most of the countries are balanced in terms of their links and extreme values of the *network balance* parameter are rare although they do exist.

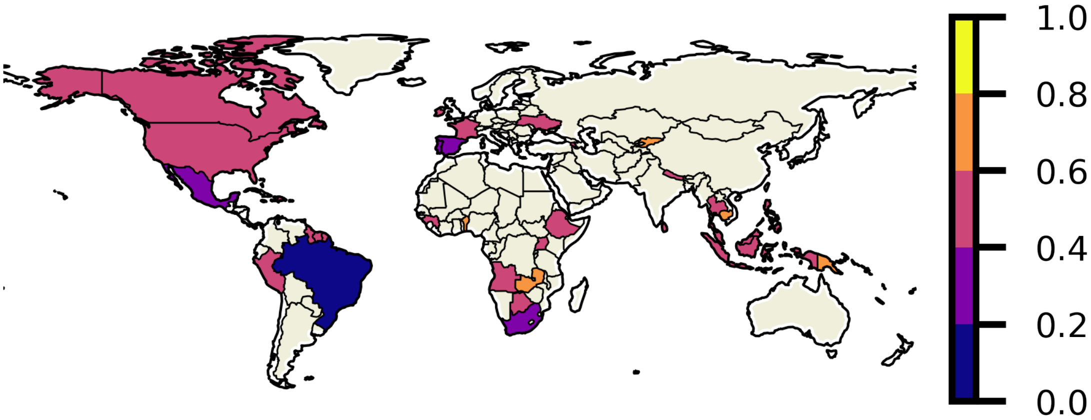


Figure 3.6: Color coded map of the countries in the WTN for the network balance parameter. We see that most countries either don't have another country with which to exchange tourists or the dataset does not have enough information about them.

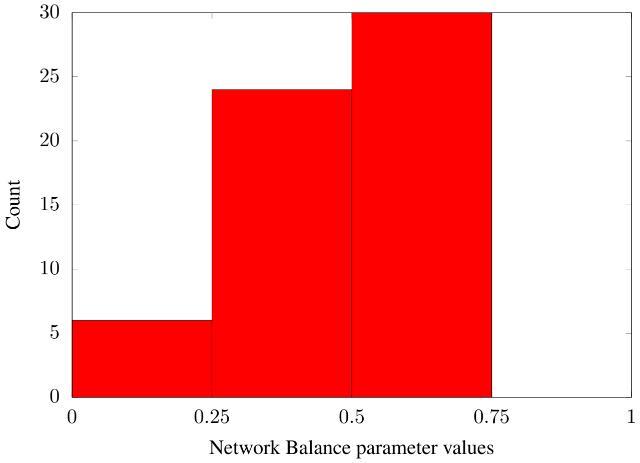


Figure 3.7: Distribution of the Network Balance parameter values in the WTN.

The *network balance* parameter represents a functionality parameter of the network, and we want to see if there is a correlation between the *T parameter* and the network balance parameter values. In order to calculate the correlation between these two parameters, we will use the Pearson correlation coefficient. Fig. 3.8 shows the correlation between the network balance parameter and the T parameter. The correlation coefficient value for these two parameters is $\rho = 0.073$ which means that there is almost no correlation between the values of the *network balance* and T parameters. With this result, we can not say anything for certain about how touristically balanced a country is by only looking at its T parameter (structure), and vice-versa.

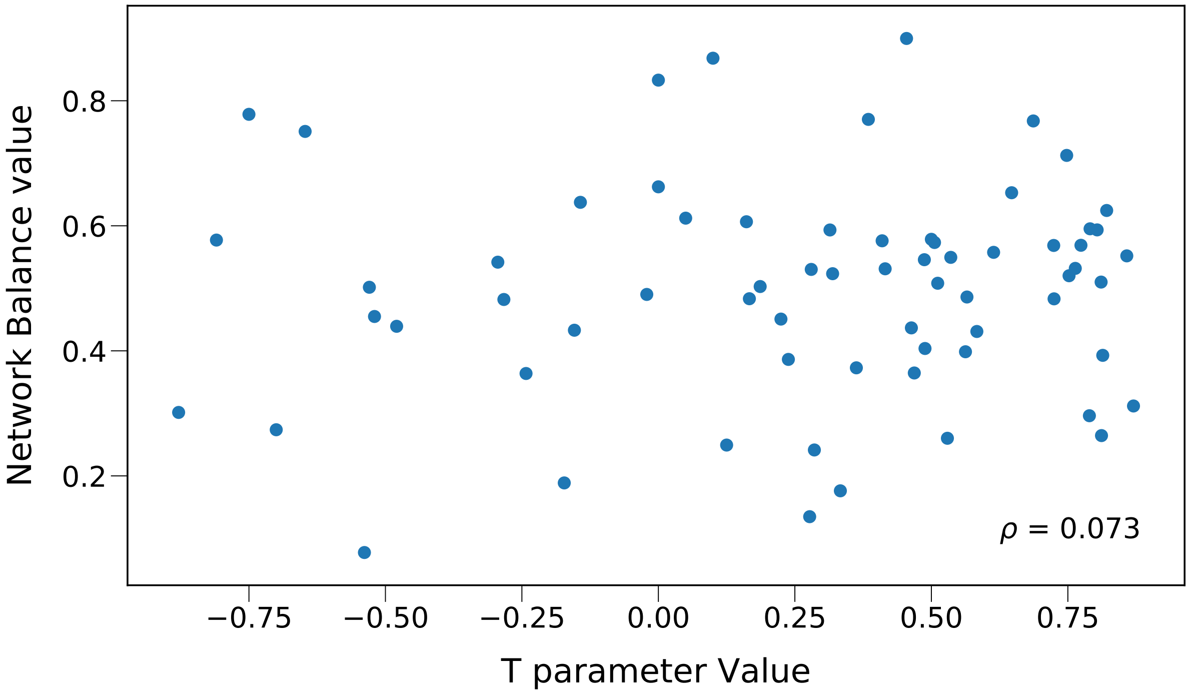


Figure 3.8: Network Balance and T parameter correlation plot. The Pearson correlation value is 0.073 .

3.5 Conclusion

Having looked at measures related to the structure and functionality of the network and the relationship between them, we see that most of the countries are typically tourist-sending countries and also that there is a strong correlation between the T and P parameters. The network balance analysis tells us that most of the countries analyzed are balanced in terms of their link values and also that high values of this parameter are rare, meaning that for reciprocated pairs of links, a touristic flow balance exists. Finally, we looked at the correlation between the T and *network balance* parameters which is very low ($\rho = 0.073$).

Chapter 4

Community Detection

4.1 Introduction

Networks consist of nodes that (in general) interact differently with each other. This behaviour, present in networks of all sizes gives rise to the presence of communities. One of the classical examples of this is Zackary's karate club network [28], in which the author studied the interactions of 34 members of a karate club from 1970 to 1972. One peculiarity of this study was that during it, a conflict between an administrator and an instructor of the club arose which led to the split of the club into two groups and the author correctly assigned all but one member of the club, to the community they actually joined after the split. This was one of the first works to use community detection. The usual definition of a community in a complex network is a group of nodes that are more densely connected amongst themselves than with the rest of the network.

The identification of communities could provide insights into the organization of the network and the analysis of each community allows us to classify the nodes based on the role the play inside the community, for example if their connections are strictly intra-community or if they also interact with nodes outside of their own community.

There are several methods used in order to find communities in a network, which can be based on statistical inference, optimisation or on dynamics. The method used to find communities in our work is based on optimisation and it is called the *Louvain* method [29].

In order to detect communities in a directed network like ours we use a software produced by Dugué and Perez [30], which is a modified version of Louvain's algorithm, in order to handle directed networks based on the notion of directed modularity defined by Leicht and Newman [31].

4.2 Louvain method

Methods based on optimisation, like the Louvain algorithm, measure the quality of the partitions obtained by measuring a quantity called the modularity of the partition. A partition here corresponds to the division of the network into one or more communities. The modularity calculated from a partition can have values between -1 and 1 and its most popular function is defined as [32]

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - P_{ij})\, \delta(C_i, C_j), \tag{4.1}$$

where m is the number of links in the network, A_{ij} the value of the adjacency matrix for the link connecting nodes i and j, C_i and C_j refer to the communities of countries i and j, the δ corresponds to the Kronecker delta function $\delta(i,j)$ which by definition is 1 if $i = j$ and 0 otherwise, and P_{ij} is called the $null\ model\ term$.

If we redistribute the links of our network while keeping the degree of the vertices unchanged, the probability of an link existing between vertices i and j is given by $k_i k_j/2m$ where k_i is the degree of vertex i. The modularity function can then be written as

$$Q = \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j). \tag{4.2}$$

This definition allows us to measure how different the network is from a randomization of it. This approach was inspired by the idea that when we randomize the network structure, communities are destroyed and so the comparison between the network and the $null\ model\ term$ reveals how non-random the structure of the network is.

The definition for the modularity function present in Eq. 4.2 is valid for undirected and unweighted networks. However, as was shown previously, the WTN is a weighted and directed matrix, and as so the modularity function used in this work is

$$Q = \frac{1}{s}\sum_{ij}\left(A_{ij} - \frac{s_i^{out} s_j^{in}}{s}\right)\delta(C_i, C_j), \tag{4.3}$$

where A_{ij} is the number of tourists from country i in country j, and s_i^{out} and s_j^{in} are the out and in strengths of nodes i and j, respectively

$$s_i^{out} = \sum_j A_{ij}, \tag{4.4}$$

and

$$s_j^{in} = \sum_i A_{ij}, \tag{4.5}$$

and s is known as the total strength of the network, given by,

$$s = \sum_i s_i^{out} = \sum_j s_j^{in} = \sum_{ij} A_{ij}. \tag{4.6}$$

The goal of Louvain's algorithm is to maximize this function in order to obtain the largest possible value of modularity for a network.

Louvain Method Algorithm

The community detection algorithm used in this work follows the steps:

1. In the first step, each node is assigned to its own community, with this division being called a partition.

2. The algorithm tries to iteratively join communities together, accepting these newly formed communities if there is an improvement in the value of modularity of the new partition.

3. If this newly formed community is accepted, it is created and its is part of a new network where the nodes are the communities built from the previous step. Any links between nodes of the same community are now represented by self-loops on the new node and links from multiple nodes in the same community to a node in a different community are represented by weighted links between communities.

4. The second and third part are repeated until there is no gain in modularity by joining communities.

Network A - Unitary Weights

The first implementation of this algorithm in the WTN was used using the directionality of the links but changing all its weights to a unitary value making all interactions equal magnitude wise. The result is present in Fig. 4.1 and we see that four communities are found. We can see that, apart from a few nodes which are misplaced, we have a community covering America and its three subcontinents, a community linking Western and Central Africa to Eastern Europe, another linking Western Europe to the Far East, Eastern Africa and Oceania and finally a community constituted mainly of Pacific Islands.

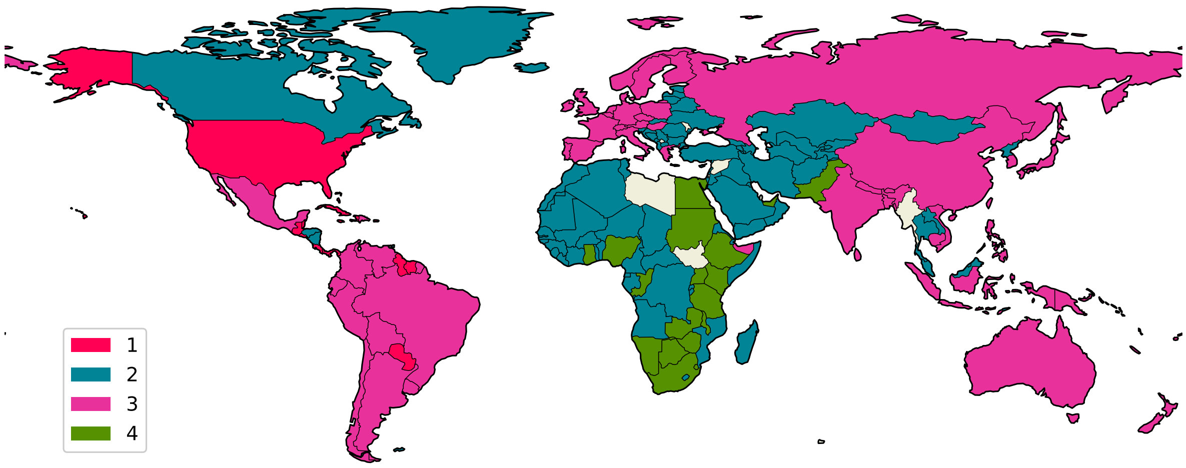


Figure 4.1: Community structure of the WTN for the case where all interactions are unitary. The number of communities is 4 and the modularity value is Q = 0.14.The countries for which we have no data are represented in grey.

Network B - Original Weights

In this network, the weights are the ones from the original matrix (WTN). The result of the implementation of Louvain's algorithm for this network is presented in Fig. 4.2.

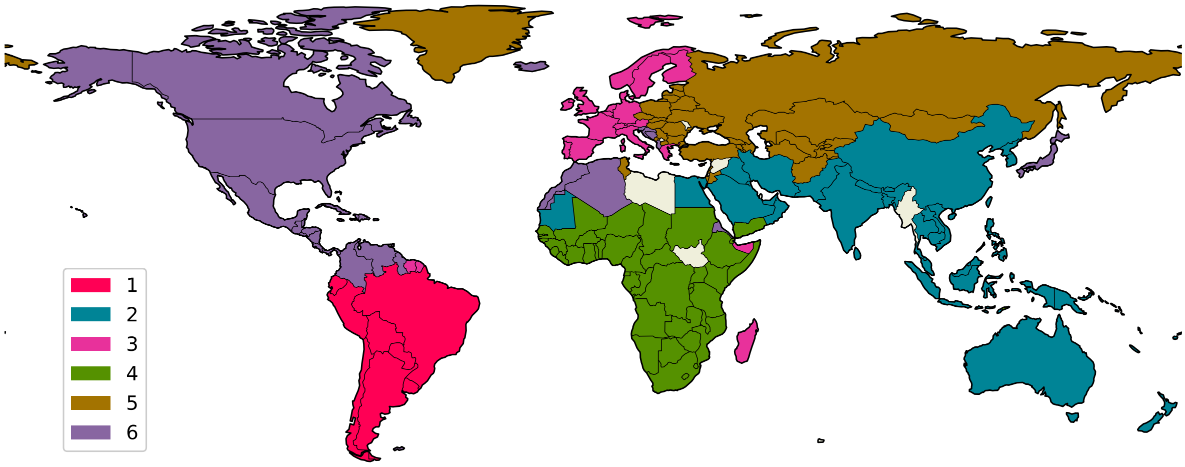


Figure 4.2: Community structure of the WTN for the case where interactions are the ones from the World Tourism Network. The number of communities is 6 and the modularity value is Q = 0.54. The countries for which we have no data are represented in grey.

If in the previous partition, we could map communities to continents, that mapping is clearer and with a higher resolution, albeit with a few nodes geographically out of place), than the previous one. We can clearly map the six communities to different regions of Earth: Western Europe, Eastern Europe, Middle and Far East and Oceania, South America, North America and Africa.

Network C - Weights rescaled by the population of the sending country

Even though the communities presented in Fig. 4.2 are a close approximation to the geopolitical division of our planet, every country has a different scale of originating tourism. Considering this, we analyze the communities by using rescaled weights of tourism flow by population of the originating country. The result is shown in Fig. 4.3 and displays a rather clear pattern of global tourism worldwide. Similar results were observed by the authors in Ref. [8] and we can conjecture that people prefer to travel to countries within their continent or sub-continent, even after the effects of globalization have created an intricate network of airlines around the world.

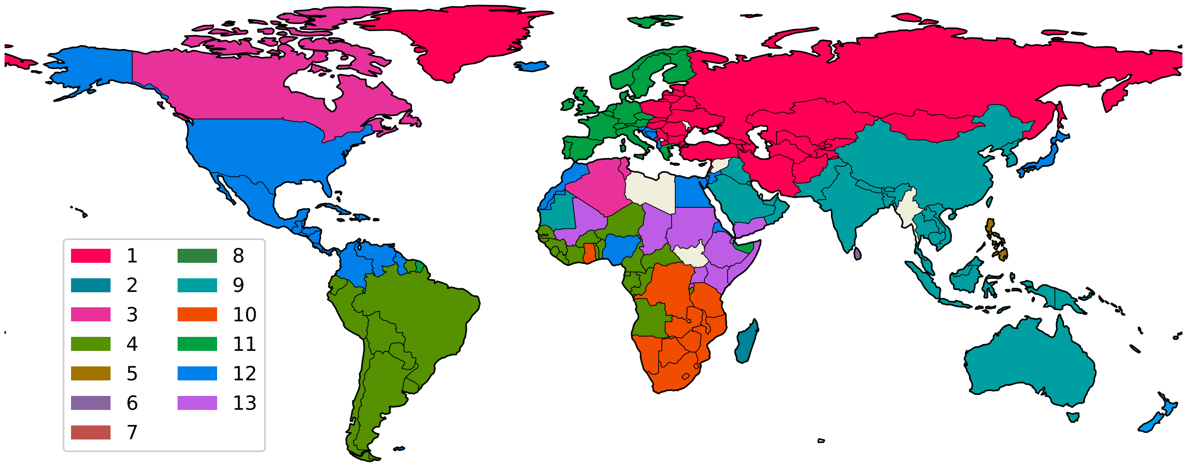


Figure 4.3: Community structure of the WTN for the case where interactions are the ones from the original matrix divided by the population of the sending country. The number of communities is 13 and the modularity value is Q = 0.7. The countries for which we have no data are represented in grey.

Network D - Weights rescaled by the out-strength of the sending country

We consider now the fraction of passengers in each link divided by the out-strength of the sending country as the values for the links. These weights can be expressed as:

$$f_{ij} = \frac{A_{ij}}{s_i^{out}}, \tag{4.7}$$

where A_{ij} represents the number of passengers going from country i to j. The results of this analysis is presented in Fig. 4.4. In this map, we can see a better geopolitical and administrative division than for the previous figure, since we can see a South and North America divide, though not as precise, and also a Western and Eastern where we see a division provided by tourism. To note also the connection between Asia and Oceania, which belong to the same community, according to this map.

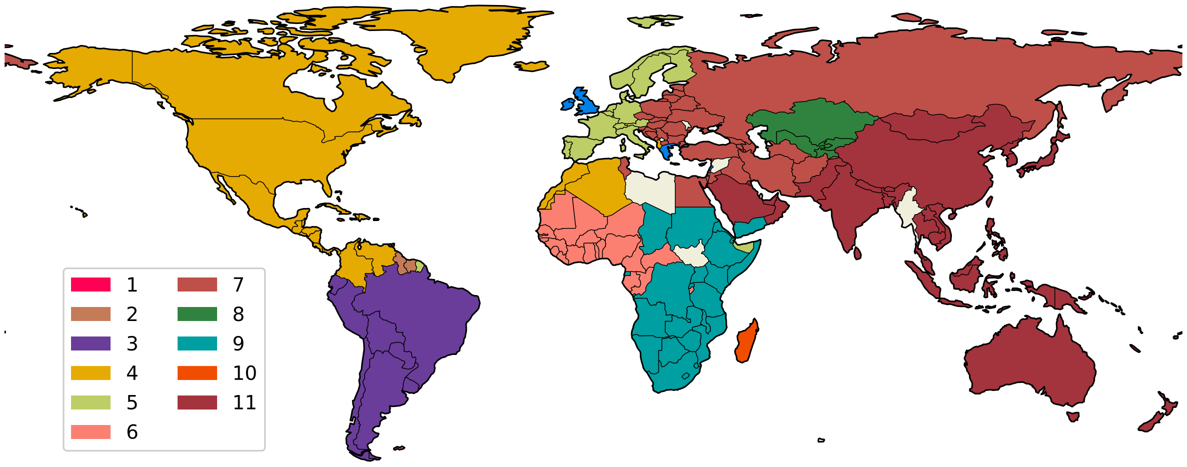


Figure 4.4: Community structure of the WTN for the case where interactions are the ones from the original matrix divided by the out-strength of the sending country. The number of communities is 11 and the modularity value is Q = 0.58. The countries for which we have no data are represented in grey.

4.3 Conclusion

Using the Louvain method in order to detect communities in the WTN, we were able to replicate, in some cases, the geopolitical division of the planet giving strength to the hypothesis that tourism is mostly a local/intra-continental business even after the process of globalisation we have been experiencing in the last decades. This method has also allowed us to get to know better which different parts of the world interact, if they even interact at all.

Conclusion

This study has allowed us to have a better view of the big business that is tourism but through a new, scientific perspective. We found out that people have, on average, to go through another country in order to get to their destination ($l^{WTN} = 2.32$, the number of countries passed by is equal to $l - 1$). Also, we observed that a low number of countries are readily available to travel, although this result could be explained by the presence of remote countries/territories in the dataset. The difference in the clustering coefficients (weighted directed and undirected versions) is also remarkable and can be attributed to the fact that most links are not reciprocated, exposing the unequal behaviour of tourism.

We found out that tourists prefer to travel to countries close by (one between-country flight away), in terms of chemical distance, even though there is a wider offer for destination two between-country flights away. It specifically highlights that most amount of traffic moves on lesser direct links and lesser amount of traffic has a rather large subset of links available due to a variety of travelling behaviour. Geographically, we observe some variety in the choice of travel to farther away countries, although the larger bulk of tourists prefers to go to closer destinations $[0 : 3000]km$. However, the relationship between each individual country's travel preference in terms of the two perceptions of distance (chemical and geographical) is unclear.

We analysed the evolution of the connection strength of the network and it allowed a glimpse into the growth of the network. We saw that the network's evolution has been gradual and that big changes have occurred few times, a sign of the constancy in people's touristic choices in the years this report covers (2004-2008).

After having looked at measures related to the structure and functionality of the network, we then set out to find if there was any the relationship between them, and we observed that most of the countries are typically tourist receiving countries and also that there is a strong correlation between the a parameter pertaining to the network structural balance (*T* parameter) and another describing the functional balance of the network (*P* parameter). The network balance analysis performed tells us that most of the countries analysed are balanced in terms of their link values and also that extreme values of this parameter are rare. Finally, we looked at the correlation between the *T* and *network balance* parameters which is very low ($\rho = 0.073$) indicating a lack of correlation between these two last parameters and indicating that information about the structural properties of a node won't tell us much about its touristic balance, and vice-versa.

We also performed a community detection study regarding the WTN, using the Louvain method and were able to replicate, in some cases, the geopolitical divisions of the planet giving strength to the hypothesis that tourism is mostly a local/intra-continental business even after the process of globalisation we have been experiencing in the last decades. This method has also allowed us to get to know better which different parts of the world interact, if they even interact at all, with each other, though the exchange of tourists.

In short, we found that touristic travelling is indeed biased, as people tend to travel to close-by destinations and also that touristic travelling does not seem to be global, as is indicated by the strong resemblance between our obtained communities and the geopolitical division of the world.

To conclude, we would like to say that there is still many work to be done about this subject and a few topics to be explored are for example, possible optimization strategies for the WAN in order to meet the demands of touristic travel and also the influence of the recent rise of market share by the low-cost carriers in the tourism industry.

Bibliography

[1] World Tourism Organization. *UNWTO Tourism Highlights 2017 Edition*, 2017.

[2] Y.N. Harari. *Sapiens: A Brief History of Humankind.* HarperCollins, 2015.

[3] A. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281:69, 2000.

[4] B. A. Carreras, V. E. Lynch, I. Dobson, and D. E. Newman. Complex dynamics of blackouts in power transmission systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 14:643, 2004.

[5] A. Barabasi and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101, 2004.

[6] G. F. Davis, M. Yoo, and W. E. Baker. The small world of the american corporate elite, 1982-2001. *Strategic Organization*, 1:301, 2003.

[7] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences*, 102:7794, 2005.

[8] T. Verma, N. A. M. Araújo, and H. J. Herrmann. Revealing the structure of the world airline network. *Scientific Reports*, 4:5638, 2014.

[9] Massimiliano Zanin and Fabrizio Lillo. Modelling the air transport with complex networks: A short review. *The European Physical Journal Special Topics*, 215(1):5–21, 2013.

[10] Ernesto Estrada and Jesús Gómez-Gardeñes. Network bipartivity and the transportation efficiency of european passenger airlines. *Physica D: Nonlinear Phenomena*, 323-324:57 – 63, 2016. Nonlinear Dynamics on Interconnected Networks.

[11] O. Lordan, J. M. Sallan, P. Simo, and D. Gonzalez-Prieto. Robustness of the air transport network. *Transportation Research Part E: Logistics and Transportation Review*, 68:155, 2014.

[12] M. Azzam, U. Klingauf, and A. Zock. The accelerated growth of the worldwide air transportation network. *The European Physical Journal Special Topics*, 215:35, 2013.

[13] Guillaume Burghouwt and Jacco Hakfoort. The evolution of the european aviation network, 1990–1998. *Journal of Air Transport Management*, 7(5):311 – 318, 2001. Developments in the Deregulated Air Transport Market.

[14] Hawoong Jeong, B Tombor, R Albert, Z.N. Oltvai, and Albert-Laszlo Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–4, 11 2000.

[15] Jennifer A. Dunne, Richard J. Williams, and Neo D. Martinez. Food-web structure and network theory: The role of connectance and size. *Proceedings of the National Academy of Sciences*, 99(20):12917–12922, 2002.

[16] Cecile Caretta Cartozo, Diego Garlaschelli, Carlo Ricotta, Marc Barthélemy, and Guido Caldarelli. Quantifying the taxonomic diversity in real species communities. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224012, 2008.

[17] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68:065103, Dec 2003.

[18] Diego Garlaschelli, Stefano Battiston, Maurizio Castri, Vito D.P. Servedio, and Guido Caldarelli. The scale-free topology of market investments. *Physica A: Statistical Mechanics and its Applications*, 350(2):491 – 499, 2005.

[19] Hermann Riecke, Alex Roxin, Santiago Madruga, and Sara A. Solla. Multiple attractors, long chaotic transients, and failure in small-world networks of excitable neurons. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 17(2):026110, 2007.

[20] J Poncela, J Gómez-Gardeñes, L M Floría, and Y Moreno. Robustness of cooperation in the evolutionary prisoner's dilemma on complex networks. *New Journal of Physics*, 9(6):184, 2007.

[21] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4):175 – 308, 2006.

[22] OpenFlights. `https://web.archive.org/web/*/https://openflights.org/data.html`. Accessed: 2017-09-30.

[23] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101:3747, 2004.

[24] J. Heitzig, J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes. *The European Physical Journal B*, 85:38, 2012.

[25] Mark Newman. *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA, 2010.

[26] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440, 1998.

[27] Giorgio Fagiolo. Clustering in complex directed networks. *Physical Review E*, 76:026107, 2007.

[28] W.W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452, 1977.

[29] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[30] Nicolas Dugué and Anthony Perez. Directed Louvain : maximizing modularity in directed networks. 2015.

[31] E. A. Leicht and M. E. J. Newman. Community structure in directed networks. *Phys. Rev. Lett.*, 100:118703, Mar 2008.

[32] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69:066133, 2004.