On the multi-dimensionality and sampling of air transport networks

Belkoura, S., Cook, A.J., Pena, J.M. and Zanin, M.

NOTICE: this is the authors' version of a work that was accepted for publication in Transportation Research Part E: Logistics and Transportation Review. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Transportation Research Part E: Logistics and Transportation Review, Volume 94, October 2016, Pages 95–109.

Transportation Research Part E: Logistics and Transportation Review is available online at:

https://dx.doi.org/10.1016/j.tre.2016.07.013

On the multi-dimensionality and sampling of air transport networks

Abstract

Complex network theory is a framework increasingly used in the study of air transport networks, thanks to its ability to describe the structures created by networks of flights, and their influence in dynamical processes such as delay propagation. While many works consider only a fraction of the network, created by major airports or airlines, for example, it is not clear if and how such sampling process bias the observed structures and processes. In this contribution, we tackle this problem by studying how some observed topological metrics depend on the way the network is reconstructed, *i.e.* on the rules used to sample nodes and connections. Both structural and simple dynamical properties are considered, for eight major air networks and different source datasets. Results indicate that using a subset of airports strongly distorts our perception of the network, even when just small ones are discarded; at the same time, considering a subset of airlines yields a better and more stable representation. This allows us to provide some general guidelines on the way airports and connections should be sampled.

Keywords: Air transport, complex networks, network topology.

1. Introduction

During the last decade, the application of complex network theory [1, 2, 3, 4] to the study of air transport has experienced a tremendous growth. Such theory has demonstrated its usefulness in the analysis of many real-world complex systems, from social networks, to the Internet or the human brain [5]; in all these cases, it has been possible to obtain a better understanding of the system structure, and of the corresponding dynamics. Air transport has been no exception, with examples including simple topological analyses, its structural evolution through time, the resilience of the system to perturbations, the dynamics of passengers, or air transport's contribution to epidemic spreading - see [6, 7] and references within for further details.

Creating a network representation of a given system entails two steps: map the elements composing the system into nodes, and establish links between pairs of nodes when a relationship is detected among them. Such processes may be far from trivial, as for instance for spatially extended systems lacking a characteristic resolution (*i.e.* a characteristic spatial scale); or for systems without explicit relationships between elements, in which case such relationships have to be derived from the dynamics of the composing elements (the case of functional networks) [8]. Such difficulty is not *prima facie* present when creating network representations of an air transport system, as nodes and links can directly be mapped - the former from airports, the latter from direct flights or passenger itineraries. The definition of nodes and links is nevertheless complicated by two problems.

First, air transport network representations may be subject to a sampling process, *i.e.* when a subset of airports is considered, for instance the most connected ones, or when connections correspond to a subset of airlines. Several reasons may be hidden behind such sampling: the need to reduce the computational cost; the interest in the analysis of one airline, or of a region of the airspace; the reduced availability and reliability of data for small airports and airlines; or biases in the source datasets, which may have been collected according to some incomplete processes. The literature provides numerous examples of such sampling processes. [6] shows that different analyses of the same air transport system reported a substantially different number of nodes (and links).

For instance, China included 128 airports and 1165 connections in [9], 144 and 1018 in [10], and 203 airports and 1877 connections in [11] - additional intermediate values can be found in [12, 13]. Also of interest is the case of the USA air transport network, which was represented respectively by 215, 272, 305 and 732 airports in [14, 15, 16, 17]. In both cases, the variation in the number of reported airports is significant: the largest network is 59% larger than the smallest one for China, and 240% for the USA - see Table 1 for a full review. Also, the number of airports is well below credible estimates of the likely true numbers, respectively 442 for China and 5194 for the USA [18].

Notably, in some cases the number of airports is not even reported [19, 20]. It is well known that considering a sampled version of a network (*i.e.* a subnetwork) has important consequences for the observed topological features, as some properties may be lost and others may emerge in a spurious way [21, 22].

US

Year	# airports	# routes	Reference
-	215	*116,725	[14]
2005	272	6,566	[15]
2010	305	2,318	[16]
2010	732	6,086	[17]

China

Year	# airports	# routes	Reference
-	128	1,165	[9]
2007-2008	144	1,018	[10]
2008-2009	140	1,044	[25]
2012	170	1,129	[26]
-	186	-	[12]

Table 1: Number of airports and routes reported in the literature for the US and Chinese networks. The number marked with an asterisk (*) refers to the total number of flights in the network. A dash symbol (−) indicates that the corresponding value was not reported.

In the previously reported examples of China, for instance, the clustering coefficient varied between 0.69 [10] and 0.73 [9]; more startling is the case of the Italian network, which was reported to have clustering coefficients between 0.07 in [23] and 0.42 in [24]. While the impact of a sampling process has been studied in different theoretical and applied contexts [21, 22], it has largely been neglected in air transport.

The second problem emerges when one considers the multi-dimensional nature of the air transport system [27]. In general terms, a complex system can be represented as an object composed of networking elements, which lie in a multi-dimensional space (see Fig. 1). When creating a (single-layer) network, this multi-dimensional nature is discarded. On the other hand, some information can be retained when creating a *multi-layer* representation, which is tantamount to projecting the original object in one dimension, such that that dimension is represented by different layers [3, 28]. Nevertheless, as represented in Fig. 1, even when just three dimensions are considered, two multi-layer projections can be obtained; furthermore, for highly dimensional systems, some dimensions have necessarily to be discarded in order to create the representation. When considering the air transport, several

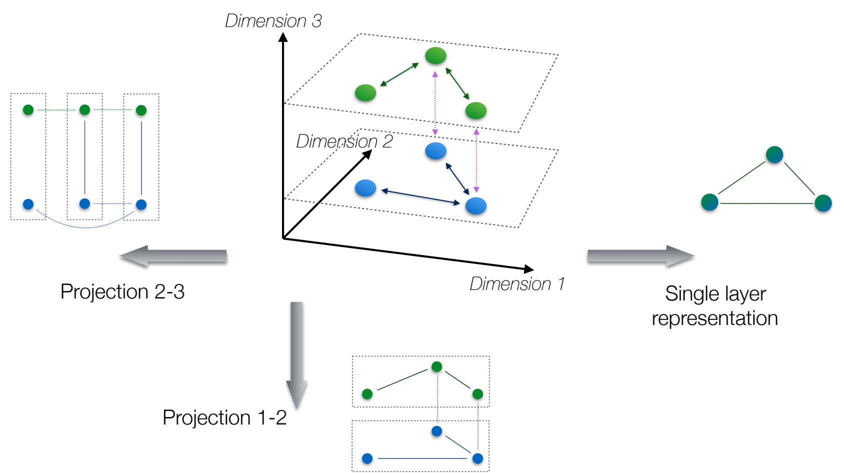

Figure 1: Projecting a multi-dimensional system. The original system (in the centre) can be represented as a single (right arrow) or as multi-layer networks (left and bottom arrows). Even in the latter case, some dimensions are discarded, leading to a loss of information.

dimensions may be included. Some of them, *e.g.* airlines, are trivial and have already been considered in past research [29]; others, like aircraft types or time windows, have mostly been neglected. One problem thus emerges: is it safe to discard some dimensions to create multi-layer networks? Is the multi-layer representation obtained still representative of the system under analysis? Notice how this second problem is strictly connected to the former: discarding some dimensions is equivalent to discarding some information in the projection, *i.e.* to sample links according to some *hidden* variables.

Merging both ideas, in this work we aim to assess if and to what extent a sampling process biases the topological and dynamical properties one observes, with respect to what would be obtained by studying the complete topology. Additionally, we also study whether a better sampling process exists, *i.e.* one that ensures a minimisation of the observed error. We here tackle these problems by studying the evolution of some commonly used metrics, as a function of different criteria (number of airports, number of connections, types of aircraft and time windows), and by comparing different sampling strategies. Eight of the most important air transport networks (Australia, Brazil, Canada, China, Europe, India, Russia and the USA) and three different datasets are considered.

The remainder of this contribution is organised as follows. Section 2 introduces the datasets and the complex network metrics analysed. Section 3 presents the observed results, considering airport (Section 3.1), aircraft type and time window (Section 3.2), and airline (Section 3.3) sampling processes, the dynamical analysis of the networks (Section 3.4), and the creation of an optimal sampling process (Section 3.5). Section 4 finally draws some conclusions, and introduces some recommendations about the best sampling methods.

2. Materials and methods

2.1. Data sets

In this work, we consider ten air transport networks, corresponding to eight countries / regions of interest, chosen for having the largest number of airports: Australia, Brazil, Canada, China, Europe, India, Russia and the USA. The networks have been reconstructed from three datasets:

- *ALL-FT+*, a Flight Trajectory (ALL-FT+) data set provided by the EUROCONTROL PRISME group. It includes information about planned and executed trajectories for all flights operating within, or crossing, the European airspace under Instrument Flight Rules (thus excluding general aviation and leisure aircraft). Intercontinental flights are excluded. The data set covers the period from 1st March to 31st July 2011.

- *On-Time Performance data set*, provided by the Research and Innovative Technology Administration (RITA), of the United States Department of Transportation. It includes detailed information about executed commercial passenger(non freight) US flights, as reported by the 16 US non-freight certified air carriers that account for at least one percent of domestic scheduled passenger revenues. Intercontinental flights are excluded. Data used covered the same time window as the ALL-FT+ data set.

- *OpenFlights*. Open source repository of flights and airport data, with worldwide coverage. It does not include freight flights, nor information about the frequency of individual routes. Available at `http://openflights.org`. Downloaded on August 17^{th} 2015.

Country	Data set	# airports	# airlines	A/c avail.	Time avail.
Australia	OpenFlights	112	12	No	No
Brazil	OpenFlights	119	12	No	No
Canada	OpenFlights	204	24	No	No
China	OpenFlights	185	17	No	No
Europe	ALL-FT+	1854	100	Yes	Yes
Europe	OpenFlights	497	153	No	No
India	OpenFlights	71	8	No	No
Russia	OpenFlights	104	36	No	No
USA	RITA	286	16	Yes	Yes
USA	OpenFlights	595	81	No	No

Table 2: Summary of network characteristics.

Table 2 reports some basic statistics about the ten reconstructed networks, including the availability of the aircraft type (fifth column) and of the time stamp of the flight (sixth column). Fig. 2 depicts the standard and cumulative distributions of airports as a function of the number of operations. All these are long-tailed distributions, in some cases (Australia, Canada and USA OpenFlights) presenting a scale-free behaviour. This suggests that air transport networks have no characteristic scales, and that therefore no threshold can be fixed beforehand for selectively filtering the network.

The European and USA air transport systems are two special cases, in that two data sets were available for each. It is worth noting the differences, both in number of airports and airlines. Specifically, the RITA data set includes information for a reduced number of carriers, which in turn limits the number of nodes mapped in the network; on the other hand, the ALL-FT+ representation of the European network includes more airports, but fewer airlines, than OpenFlights. This highlights the complexity of the problem: even for a single country, different datasets obtained through various sources may present distinctly different characteristics. As each of them samples reality in a different way, one should be aware that none is a perfect representation of the actual air transport network. While we use the most commonly-accepted official data sources for Europe and the USA (RITA and ALL-FT+), the reader should note that they are not exempt from this problem.

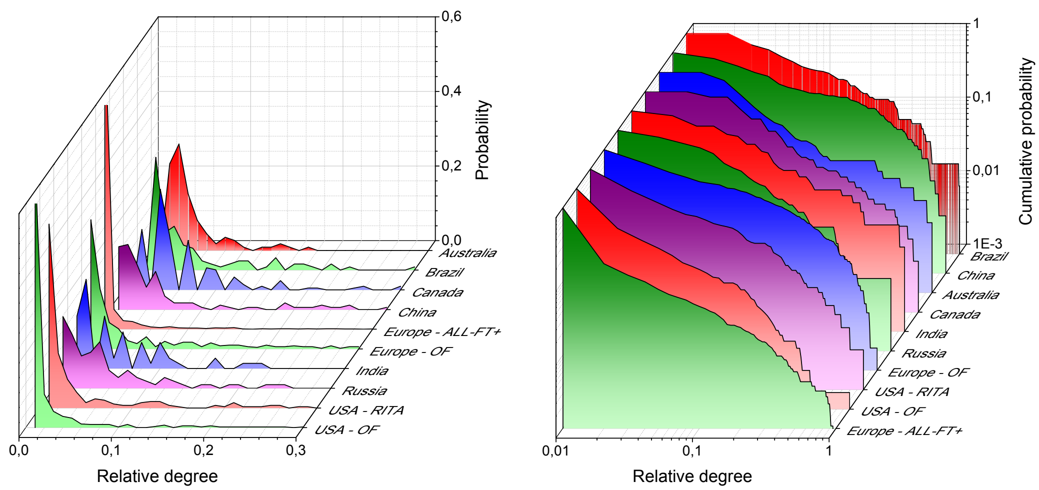

Figure 2: Normal (left panel) and cumulative (right panel) degree distribution of the 10 air transport network studied.

2.2. Topological metrics

Following the standard notation [1, 2, 3], a network composed of n nodes is fully represented by its *adjacency matrix* $\mathcal{A}$ of size $n \times n$, whose element $a_{i,j} = 1$ if a link exists connecting nodes i and j, and $a_{i,j} = 0$ otherwise.

Note that, with the exception of the analysis performed in Sections 2.3 and 3.4, we here consider unweighted and directed networks; in other words, the number of flights connecting pairs of airports is discarded, and it may be that a pair i, j is connected by a direct flight $i \rightarrow j$ but not by the reciprocal $j \rightarrow i$.

Different *topological* metrics can then be extracted from $\mathcal{A}$, each of them describing some specific structural characteristic of the network [30]. In this work, we consider the following:

Link density Is defined as the proportion of links that are active, with respect to the total number of potential links, *i.e.*

$$l_d = \frac{\sum_{i,j} a_{i,j}}{n^2} = \frac{l}{n^2}. \tag{1}$$

It is therefore defined in the interval $[0, 1]$, 0 and 1 respectively indicating a void and a fully connected network.

Maximum degree The maximum degree of a network is defined as the degree of the most connected node: $k_{max} = \max_i k_i$. Here k_i represents the degree of node i, *i.e.* $k_i = \sum_j a_{i,j}$.

Clustering coefficient The *clustering coefficient*, also known as *transitivity*, measures the presence of triangles in the network [31]. It is mathematically defined as the relationship between the number of triangles in the network N_Δ (three vertices with edges between each pair of them), and the number of connected triples N_3 (set of three vertices where each vertex can be reached from all other, directly or indirectly):

$$\begin{aligned} &C = \tfrac{3N_\Delta}{N_3}, \text{where} \\ &N_\Delta = \sum_{k>j>i} a_{ij}a_{ik}a_{jk}, \\ &N_3 = \sum_{k>j>i} (a_{ij}a_{ik} + a_{ji}a_{jk} + a_{ki}a_{kj}). \end{aligned} \tag{2}$$

C close to 1 indicates that all triangles are *closed*; or, as usually depicted in social network analysis, that 'the friend of my friend is also my friend'.

Degree correlation This metric, also called 'assortativity', represents the conditional probability $P(k'|k)$, *i.e.* the probability of a link from a node of degree k to point to a node of degree k'. When expressed by means of a Pearson correlation coefficient, it is defined as [32]:

$$\frac{1}{M} \sum_{j>i} \frac{1}{2} (k_i + k_j) a_{ij}, \tag{3}$$

M being the total number of links in the network.

Entropy of the degree distribution The entropy of the degree distribution provides a measure of the heterogeneity of the network [33], and is defined as:

$$E_{dd} = -\sum_k p(k) \log_2 p(k). \tag{4}$$

Values close to 1 indicate a uniform degree distribution, while the minimum $E_{dd} = 0$ is achieved when all vertices have the same degree.

Efficiency The efficiency of a network represents how easily one can move between two nodes, *i.e.* how many intermediate nodes one has to visit in order to reach the 'destination'. While the length of the shortest path between two nodes might be used to estimate this property, it has an important drawback: when the network is disconnected, no path may exist connecting two nodes i and j, and therefore $d_{ij} = \infty$ and $l = \infty$. The efficiency solves this by considering the inverse of the distance [34]:

$$E = \frac{1}{N\left(N-1\right)} \sum_{i \neq j} \frac{1}{d_{ij}}. \tag{5}$$

Information content (IC) This metric assesses the presence of regularities in the adjacency matrix, and thus the presence of mesoscale structures [35]. It is calculated by first identifying pairs of nodes that share a similar connectivity pattern, which are thus redundant and whose merging would suppose the smallest information loss (from a Shannon information theory perspective). Once the best pair has been detected, both nodes are merged (thus yielding a network one node smaller), and the quantity of information lost in the process is calculated. The final information content is then the sum of all information loss when shrinking the original network to a single node, normalized with respect to an ensemble of random networks. The metric thus represents the quantity of information needed to reconstruct the network, or the quantity of information encoded in its structure. The lower such information, the more regular the underlying topology, as for instance certain patterns of connectivity are repeated; thus, a relatively low IC indicates the presence of mesoscale structures, *e.g.* communities.

2.3. Dynamical analysis

In order to understand how the observed differences in topology may affect a dynamical process taking place in the network, here we also consider a simple dynamical model mimicking delay propagation. Such a model is based on a random propagation process, similar to other models used to simulate rumour spreading [36]. We suppose delays are generated at airports in a random fashion, and then propagated through the network according to its flight structure. It should be noted that this model does not aim at understanding the real propagation process, as it does not include elements such as aircraft, crew or passenger dependencies (connectivities), which may contribute to the generation of (additional) reactionary delays, or indeed the sequestering thereof. For the sake of simplicity, all complex details about the real process have been disregarded, in order to better highlight the biases introduced by the different sampling strategies studied here - see Section 3.4 for further details.

Let's assume that the accumulated delay at an airport i is proportional to the sum of the delays of neighbouring (connected) airports [1]:

$$d_i = \frac{1}{\lambda} \sum_{j \in M(i)} d_j, \tag{6}$$

$M(i)$ being the set of neighbouring nodes of i, and λ a proportionality constant. The set of neighbours can be obtained through the adjacency matrix, such that:

$$d_i = \frac{1}{\lambda} \sum_{j} A_{i,j} d_j. \tag{7}$$

Finally, the previous equation can be rewritten as:

$$A\mathbf{d} = \lambda \mathbf{d}, \tag{8}$$

A being the full adjacency matrix, and $\mathbf{d}$ the vector of centralities $(d_1, d_2, \ldots, d_n)$. The previous equation is an eigenvector equation, and the vector $\mathbf{d}$ corresponds to the stationary distribution of the Markov chain represented by the row-normalised adjacency matrix. In other words, $\mathbf{d}$ represents the expected equilibrium distribution of delays when they propagate according to a random process in the flight network. This measure, assessing the importance of each node, is called 'eigenvector centrality', and has been extensively used to understand the importance of individuals in social networks [37]. This approach has the advantage of being analytically solvable, as $\mathbf{d}$ corresponds to the eigenvector associated with the largest eigenvalue of the A matrix [2].

[1]Note that this model makes no assumptions on the way delays are generated, but only considers how they would propagate once they have appeared. As such, the delay observed at one airport only depends on external contributions, and not on internal generation.

[2]The existence of an unique $\mathbf{d}$ is ensured by the Perron-Frobenius theorem for positive and non-negative matrices. Additionally, this theorem guarantees that the largest eigenvalue is associated with the only eigenvector with all non-negative and real elements, and

2.4. Optimisation procedure

In order to better understand which nodes are responsible for the observed network structure, and should thus be retained in a sampling procedure, we have here implemented a greedy optimisation algorithm.

The process starts with the full network, which is characterised by a set of topological metrics $\mathbf{t}$. The node i is then temporarily deleted from the network, to assess how its deletion affects the topological structure (represented by a new set of metric values $\mathbf{t}'_i$); this process is repeated for all nodes. Finally, the error introduced by the deletion of each node i is quantified as:

$$e_i = \frac{1}{n_t} \sum_{j=1}^{n_t} (t_j - t'_j)^2, \tag{9}$$

n_t being the number of elements of $\mathbf{t}$. The node associated with the smallest error is permanently deleted from the network, and the whole process is repeated until just one node remains.

Note that, in order to obtain meaningful results, all metrics included in $\mathbf{t}$ should have similar values and be defined in similar domains; this condition will be fulfilled in Section 3.5, where C and E (both defined in $[0, 1]$) are used. Additionally, it is important to note the difference between the optimisation procedure implemented here, and the concept of 'backbone' in network theory [38]. While both concepts aim at extracting a subset of nodes representing the basic connectivity structure, the latter explicitly considers only highly connected nodes (the 'rich club'), disregarding other network properties.

3. Results

3.1. Sampling airports

Air transport networks are known to have a scale-free like structure with a degree probability distribution showing a long tail [6], that is, they are composed of a few highly connected nodes (the hubs of the system), and a large number of secondary nodes (*e.g.* regional airports). This is confirmed by Fig. 2, depicting the standard and cumulative distribution of degrees - *i.e.*, the probability of finding an airport with a given number of flights (in the case of the cumulative distribution, this probability corresponds to

that thus these elements can be interpreted as probabilities.

finding at least a given number of flights). While the sampling properties of scale-free networks are well known [21], previous studies have focused on random sampling, *i.e.* the selection of nodes is performed on a random basis. This is generally not the case with the air transport network: both when the sampling is purposive, and when it reflects the limitations of the original dataset, larger airports are over-represented to the detriment of smaller ones.

In order to simulate such bias, we here introduce a sequential sampling process, in which nodes (airports) are sequentially added to the considered network following their number of connections, from highly to sparsely connected ones. Fig. 3 represents the evolution of the seven topological metrics defined in Section 2.2, as a function of the fraction of nodes sampled from the original data set, for the ten studied networks.

Some interesting facts can be observed from Fig. 3. First of all, all topological metrics strongly vary when the number of considered airports is changed. The most extreme example is represented by the entropy of the degree distribution E_{dd}, which starts from zero when the core of the network is considered (*i.e.* all airports are connected to all others and thus all have the same degree), to stabilise around one, representing a large heterogeneity in airport sizes. This confirms the presence of a *rich club*, *i.e.* the fact that the most important nodes of the network are more interconnected compared with what would be expected from a random structure, and thus that their importance is not only the result of connections with secondary nodes [38, 39]. Additionally, while most of them have a monotonic behaviour, some (*e.g.* E_{dd} and IC) change the direction of their evolution. Thirdly, metrics like the clustering coefficient, efficiency, and E_{dd} do not saturate, *i.e.* they do not reach a stable value, even for high node fractions. These three facts imply that the network topology is continuously changing when nodes, even small ones, are added, and therefore there is no natural threshold according to which nodes may be safely discarded. While this is partly due to the scale-free nature of the air transport network, the phenomenon described here is quite remarkable and specific to air transport. Past studies have shown that sampling 60% of the nodes is sufficient to achieve a good approximation of the full network (*e.g.* the Internet and the online pre-print repository arXiv) [22]. Nevertheless, these results were based on random sampling procedures, while the usual approach in air transport is guided by external information, such as airport sizes, or data availability constraints. Our sampling approach is probably the reason behind the results observed in Fig. 3.

As a fourth point, it is interesting to compare the topologies of the US and European networks for the two datasets available for each of them, *i.e.* RITA and OpenFlights for the former, and ALL-FT+ and OpenFlights for the latter. Both pairs of data sets present different characteristics, the most important one here being the different number of airports: OpenFlights is more complete in the case of the USA, but includes fewer airports for Europe. Such a difference is represented in Fig. 3 by the two vertical dashed lines, each one representing the proportion of airports in the large data set that corresponds to the size of the small one. Remarkably, all metric evolutions follow the same shape for the same region. The most clear example is the local maxima and minima for the clustering coefficient of the two USA networks, respectively at 0.25 and 0.6. If both pairs of datasets had been constructed using similar sampling criteria, topological metrics should be expected to converge to similar values for an equal number of airports. In other words, the topological metrics for the full OpenFlight European network should be equal to those obtained in the ALL-FT+ case when including the 497 most important airports; a similar result should be obtained in the case of the US network. On the contrary, this does not happen: when the same number of nodes are considered, topological metrics are different (see vertical dashed line in Fig. 3), while both final values appear to be quite similar. As the RITA data set is obtained by just considering a subset of all available airlines, such an effect may be the result of introducing an airline sampling process - this topic will be further discussed in Section 3.3.

Lastly, it is worth noting that, while the evolution of topological metrics presents a qualitatively similar behaviour for most regions, an important exception is represented by China, India and partly Russia. Specifically, the link density l_d displays a flat slope, while the clustering coefficient CC increases slowly with the fraction of airports added. While Fig. 3 alone does not provide enough information to understand the causes of such different behaviour, in the case of China, we suspect that it is related to the fact that it is not a fully competitive market - this issue will be further discussed in Section 3.3.

3.2. Sampling aircraft types and time windows

A different, although less common, way of sampling air transport networks is to filter specific aircraft types, or consider only specific time windows.

Aircraft types are related to the kind of operation performed, and selecting only some types allows the creation of specific networks for different

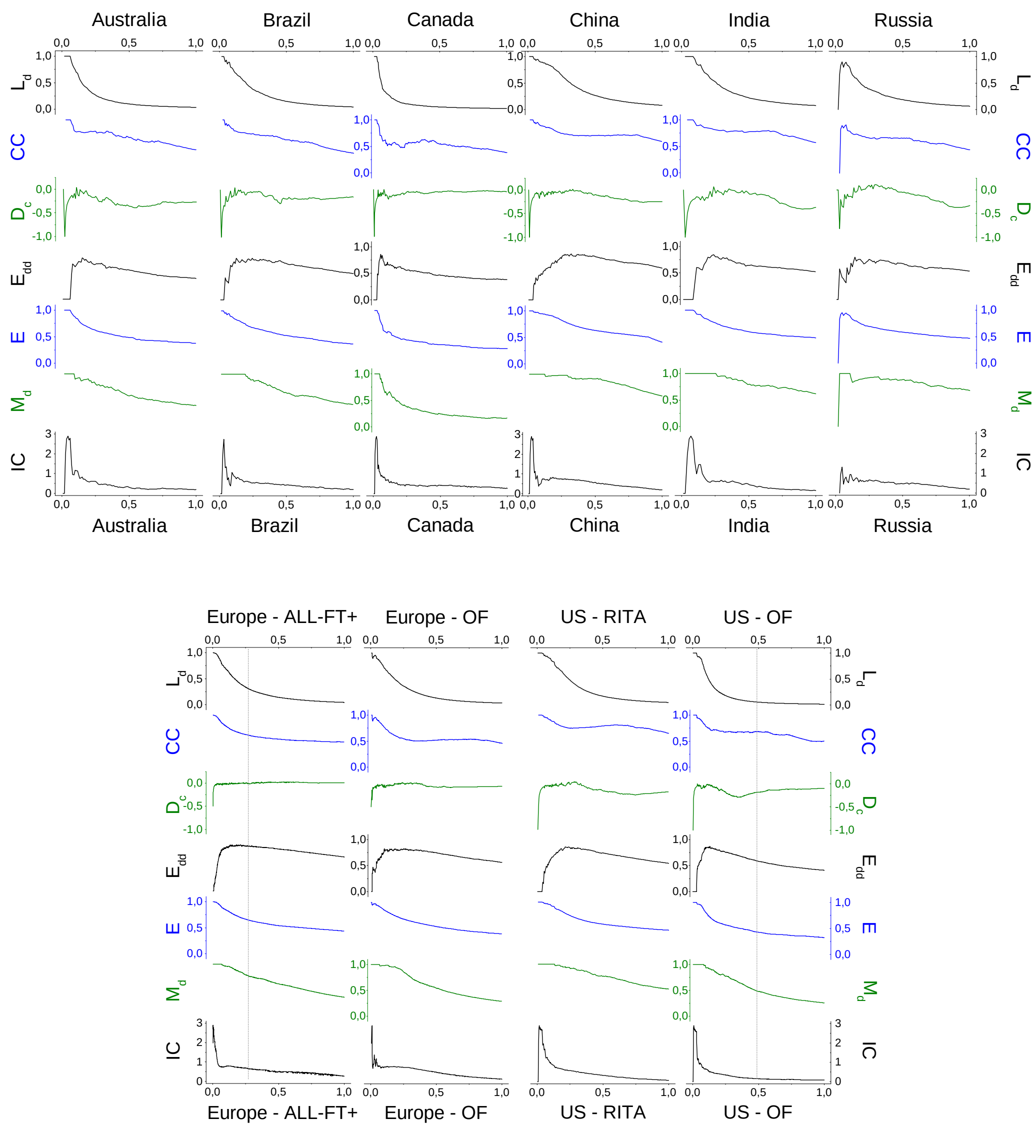

Figure 3: Evolution of network topological metrics, as a function of the fraction of airports included in the sampling process (*i.e.* the horizontal axis represents the normalised number of airports). Airports are included in decreasing order of degree.

markets: for instance, widebodies for higher density and long-haul routes; narrowbodies and turboprops for shorter hauls. Cargo-only and business aviation also need to be considered. Such specificity may be relevant to understand social and economic processes [27]. Fig. 4 depicts the evolution

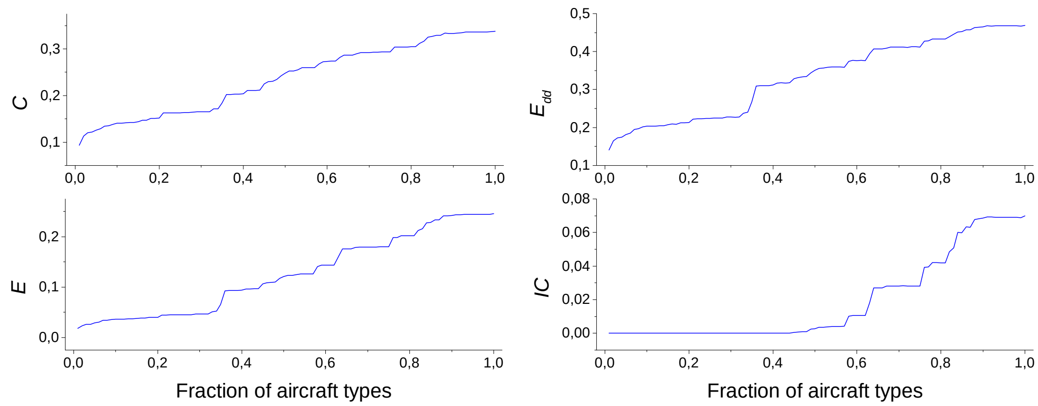

Figure 4: Evolution of topological metrics of the European (ALL-FT+) network, as a function of the fraction of aircraft types included. Aircraft are included in decreasing order of number of operations.

of four main topological metrics for Europe as a function of the fraction of aircraft types included, sorted in decreasing order of frequency. The complex nature of the air transport network appears only when most of the aircraft types are included - see, for instance, the decrease in the mesoscale complexity above the sampling fraction of 0.6, as highlighted by the increasing IC. Additionally, considering only a few aircraft types yields a mostly disconnected network, as indicated by the low initial efficiency and clustering coefficient.

It is further known that the topology of the air transport network strongly changes when considering different time windows: for instance, between summer and winter, but also when considering different days of the week. In spite of this, some research works focus on small time windows, for instance to understand the system behaviour under specific abnormal conditions [16]. Fig. 5 depicts how four major topological metrics are affected by this sampling, by considering random time windows of a given size. Similarly to what obtained in Fig. 4, realistic and complex structures are obtained only when large time intervals are analysed, thus averaging weekly and seasonal effects.

3.3. Sampling airlines

As already introduced, air transport networks can also be sampled across the airline dimension. This may be done intentionally, for instance when the topology of one airline (or of one alliance) is the focus of the study [9, 7, 40], or, indeed, represents the only data available. However, it may also be the

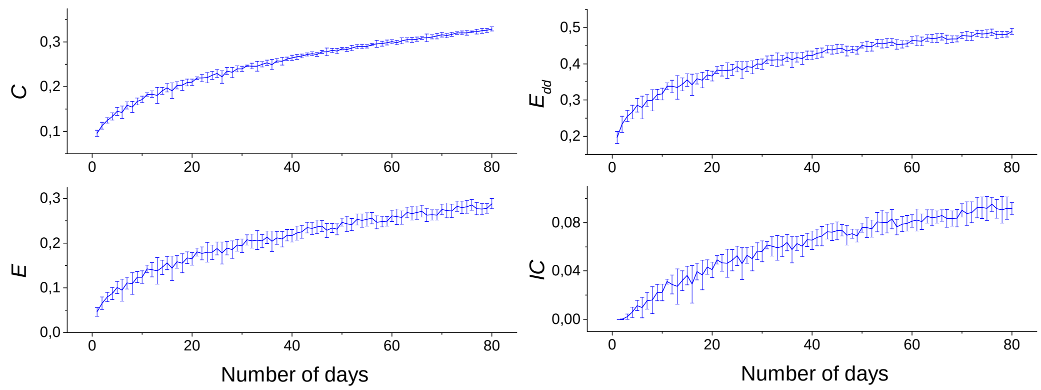

Figure 5: Evolution of topological metrics of the European (ALL-FT+) network, as a function of the number of days analysed. Each point is the average of 200 time-window calculations, and vertical whiskers indicate the standard deviation.

result of the way a data set is collected, as in the RITA case. It is therefore of interest to understand the biases introduced by such sampling.

Analogously to Fig. 3, Fig. 6 reports the evolution of the seven topological metrics considered, as a function of the fraction of airlines included in the sampled networks. This sampling method yields results quite different from those previously observed. Specifically, we note that a low number of airlines is usually sufficient to recover a good topological approximation of the complete topology.

The Chinese network presents an abnormal behaviour in Fig. 6, as it did in Fig. 3. Specifically, few airlines are not enough to recover the complete structure, and topological metrics (except IC) do not stabilise before at least one third of the airlines are included. As previously speculated, this may be due to the nature of the Chinese market: with the exception of Cathay Pacific/Dragonair in Hong Kong, which operates like a typical western airline, airlines are tightly controlled by the Civil Aviation Administration of China, which allocates them routes or regions they can operate, and determines the aircraft they are allowed to order. Due to this, including only a few airlines, even if major ones, may leave some regions of the network under-represented.

An alternative strategy may involve sampling airlines using some criteria not depending on the number of flights operated. This may be the case, for instance, of a study involving a set of airlines of some characteristics, *e.g.* region of operation, alliance membership, *etc.*; this may also occur due to some limitations in the data set used, such that not all major airlines were included. For the sake of simplicity, we here emulate this process by considering two possibilities:

a. A random process, in which a set of airlines is randomly drawn from the complete pool.

b. A substitution process. Airlines are selected in decreasing order of number of operations, but some of them are discarded and randomly replaced by other airlines. For instance, we start with the set of the five highest-operation airlines, *e.g.* DLH, RYR, EZY, AFR and BER[3]; one of them is then discarded (*e.g.* RYR) and substituted by an airline not included in the top 5 (*e.g.* SAS).

Results of both processes are reported in Fig. 7, for the European network and as a function of the number of airlines sampled. Specifically, the red solid lines report the evolution of the resulting topology for the random draw process (*a*), while the black and blue squares correspond to the substitution process (*b*). In this latter case, black symbols represent a 5:1 substitution ratio (every 5 airlines, one is substituted), blue representing 10:1; additionally, the symbols represent the average value, while the whiskers the corresponding standard deviation. It can be seen that the second strategy provides a much more stable representation of the complete topology: five airlines (four of the biggest ones, and one drawn at random) are enough for obtaining, on average, the topology of the true network. These results are easy to explain in terms of sampling probabilities: when sampling at random, there is a high probability of picking only small airlines, with a subsequent distortion of the topology; on the other hand, the substitution process ensures that the core of the network, as created by the biggest airlines, is represented.

3.4. Dynamical analysis

While complex networks have initially been studied from a structural perspective, *i.e.* encompassing the analysis of their topology as a fixed object, it was soon realised that an important problem was gaining an understanding of the dynamics of networks. According to this framework, nodes of the network represent dynamical systems, which are coupled according to the network topology. The structure of the network is thus the background on

[3]Respectively: Deutsche Lufthansa, Ryanair, EasyJet, Air France and Air Berlin.

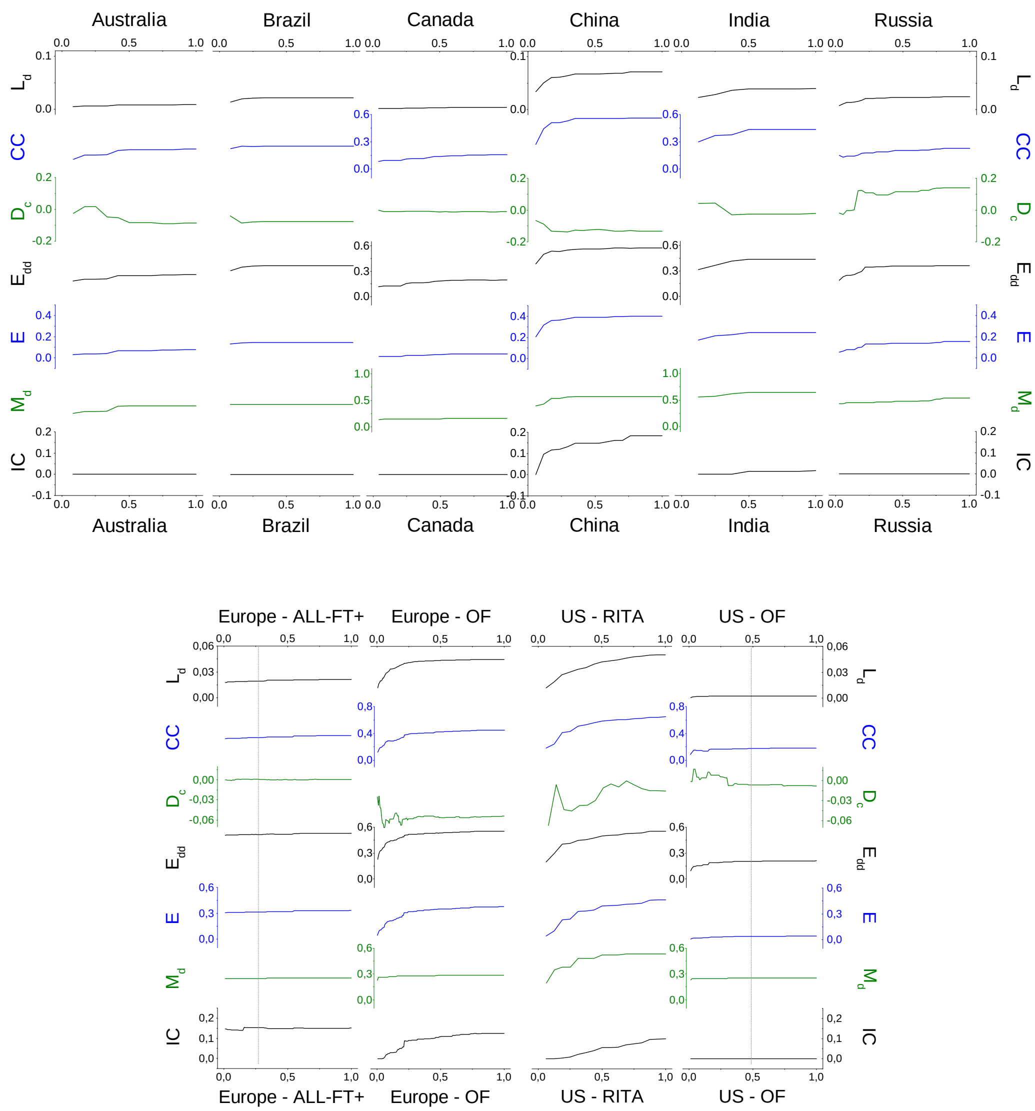

Figure 6: Evolution of network topological metrics, as a function of the fraction of airlines included in the sampling process. Airlines are included in decreasing order of number of operations.

top of which some dynamical processes take place [41, 42]. Some notable examples include the synchronisation of groups of chaotic oscillators interacting between them [43], and contact processes such as opinion formation and epidemic spreading [44].

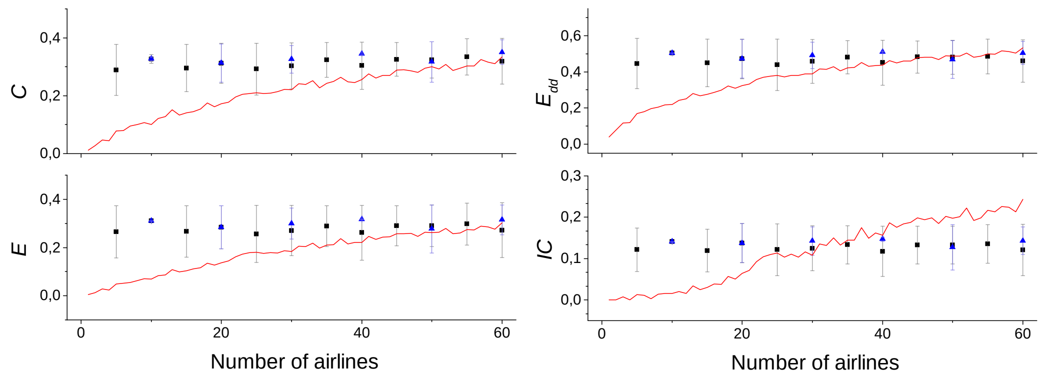

Figure 7: Topological metrics *vs.* a mixed sequential-random airline sampling process for the European (ALL-FT+) network see main text for details. Red lines indicate the evolution of metrics for a completely random sampling process.

The relationships between the topologies and dynamics of networks are far from trivial. In some cases, the topology may have little influence on the way the dynamics take place, such that different networks may be equivalent from a dynamical point of view. Nevertheless, in many other instances certain topological properties have a strong impact on the dynamics - as a relevant and practical example, it was shown that vaccination, even of a large fraction of people, cannot stop epidemics in a scale-free network [45].

Shedding light on such topological-dynamical relationships is clearly of interest also in the case of air transport. While a characterisation of the air transport network structure may be relevant from a theoretical point of view, one is usually interested in understanding how such topology influences some dynamical processes, *e.g.* delay propagation [16] or system resilience [6]. It is in principle possible that a change in the observed topological metrics has no impact on the dynamics of interest, especially if the network retains some global characteristic (*e.g.* scale-freeness); nevertheless, the opposite may also occur. It is therefore necessary to ascertain to what extent the representation biases previously described may affect our comprehension of the dynamics of the system.

In this section we are to study whether the biases observed in network topology caused by the sampling processes may affect dynamical processes. For the sake of simplicity, we have chosen a minimal model of propagation, which mimics a random delay diffusion process and is similar to other models used to simulate rumour spreading [36] - see Section 2.3 for details. While such a simple model cannot reproduce all the complexities lying behind the actual propagation of delay, it presents the advantage that results are easy to relate to topological differences, not being potentially masked by the complexity of the model. In other words, if the sampling process modifies the output of this simple dynamical model, one can conclude that any other model would (very likely) be equally affected, especially if additional non-linearities were included.

The left panel of Fig. 8 depicts the evolution of the eigenvector centrality ($\mathbf{d}$) for the 5 most important (*i.e.* highly connected) nodes of the USA air transport network (according to the RITA data set), as a function of the number of airports included (airports are added in descending order of number of connections). Note that values are normalised such that the most important node in the network has a centrality of 1, *i.e.* $d_i = d_i / max(\mathbf{d})$. Also, horizontal dashed lines correspond to the centrality as calculated by only using the number of flights operated at each airport (*i.e.*, its raw connectivity, or its strength centrality in network terms). Centralities slowly converge to a final value; nevertheless, even considering 200 airports distorts the centrality of Denver International Airport DEN (pink line) by 50%, from 0.6 to 0.9.

In Fig. 8 (left panel), the asymptotic centrality values are clearly different from the expected strength ones; this may be due to the simplification introduced by considering unweighted links, *i.e.* by discarding the actual number of flights connecting two airports. This has been addressed in Fig. 8 (right panel), representing the same information for the corresponding weighted network. In order to achieve this objective, the adjacency matrix A of Eq. 8 has been substituted by the weight matrix W, whose element w_{ij} represents the number of flights between airports i and j. This modification solves the bias in the asymptotic values, and slightly improves the sensitivity of results to the number of nodes (although more than 150 are still needed to get a good approximation of the dynamics).

In summary, the results presented in Fig. 8 highlight the fact that the biases observed in Section 3.1 are not restricted to the topological domain, but also have important repercussions in dynamical models. Even when taking into account the weight of nodes, disregarding part of the network can result in an invalid estimation of the importance of nodes, and therefore, for instance, of the most important airports from the point of view of delay propagation.

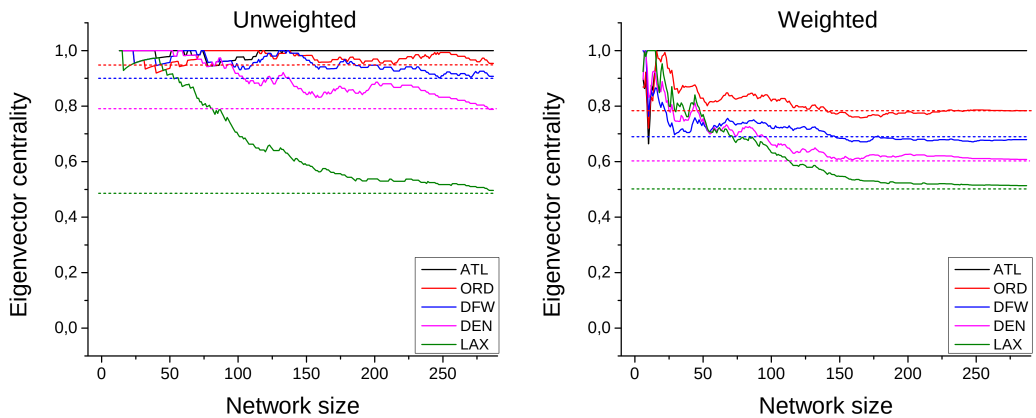

Figure 8: Eigenvector centrality (see Section 2.3 for definition) of the top five USA RITA airports, as a function of the number of airports sampled in the network, disregarding (left panel) and including (right panel) link weights. Horizontal dashed lines correspond to the node centrality calculated as the number of destinations (left) and as the number of flights (right) operated at each airport.

3.5. Structure optimisation

Once it has been proved that sampling airports according to their degree introduces a strong error in the topology of the network, one may ask whether a better criterion exists to guide such sampling; and what is the distance between any given criterion and the optimal one, *i.e.* the one ensuring a minimum topological bias in a static analysis.

In order to shed light on this issue, Fig. 9 (left panel) depicts the evolution of two topological metrics (namely, the clustering coefficient and efficiency, black and blue lines, respectively) for the US RITA network, when sampled using the greedy algorithm presented in Section 2.4. Note that, with both C and E defined in $[0, 1]$, no adverse scaling effects emerge. Even though it only performs a local optimisation, the greedy procedure ensures a good solution, while significantly reducing the computational cost of the analysis. In creating Fig. 9 (left panel), the error minimised corresponds to the deviation of both C and E - see Eq. 9. Additionally, the grey circles indicate the degree (right ordinate) of the node that is deleted at each step. In general, it can be appreciated that the efficiency is quite stable, maintaining the same value even with just 100 nodes; on the other hand, the clustering coefficient is stable up to 200 nodes (a reduction of 30% in the network size), after which it significantly decreases. At the beginning of the sampling process

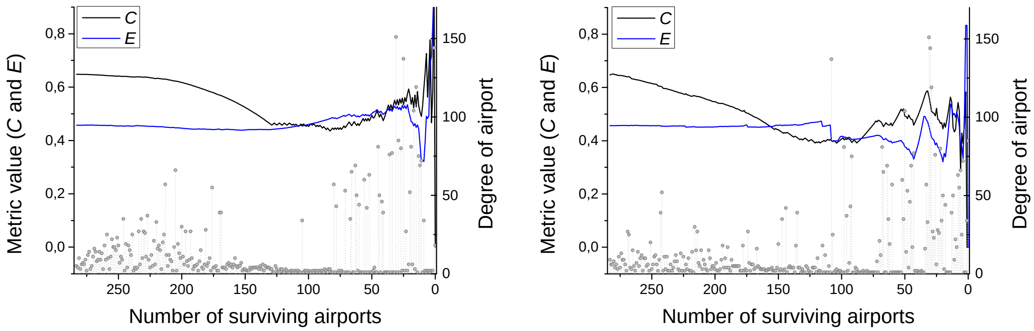

Figure 9: Evolution of C and E when the sampling process is guided by a greedy optimisation algorithm - see main text for details. The left panel corresponds to an optimal process, the right one to an optimisation guided by E_{dd}. Graphs should be read from left to right, *i.e.* starting from the complete network and ending with a single node. Grey circles (right ordinate) indicate the degree of the airport deleted at each step.

(*i.e.* left part of the graph), the deleted nodes present a high heterogeneity in degree (from 1 to 70 connections), confirming that the elimination of just the smallest nodes introduces an important topological bias.

If deleting small nodes is not a good option, one may suspect that a stratification strategy may yield better results: that is, maintain the proportion of small and large nodes, or, in other words, keep constant the distribution of degrees. Fig. 9 (right panel) depicts the results of the greedy optimisation minimising the error associated with E_{dd}. Notably, the evolution of the efficiency is very similar to that of Fig. 9 (left panel), maintaining constant down to 110 nodes. As for the clustering coefficient, its evolution is similar to the optimal case, although the drop is faster here. All in all, Fig. 9 shows that minimising the error associated with the entropy of the degree distribution is a good criterion, if a network sampling procedure has to be executed.

4. Conclusions

In this contribution, we present a study of the topological stability of complex network representations of air transport systems, aimed at understanding how observed properties are affected by representation choices, and specifically, by different sampling strategies. While one ought to use a representation of the system as true to reality as possible, this may not be possible or desirable: for instance, due to the limited coverage of available data sets;

due to the deliberate choice of discarding part of the airports/airlines, to focus the study on specific regions; or simply to reduce the computational cost. Numerous examples can be found in the literature of different sampling processes [9, 10, 14, 15, 16, 19, 20]. The aim is thus to understand how the topological and dynamical metrics observed are biased by representation choices and limitations.

Theoretical and applied studies [21, 22] suggest that sampling 60% of the nodes is usually enough to recover the statistical properties of the whole system. However, the nature of the air transport system appears to make such simplifications ineffective. Results indicate that selecting a subset of airports from the full system is not a good strategy. Due to the sensitivity of the network topology to small airports, almost all airports should be included to maintain the representation error below 10%, suggesting that a sampling process based on including highly connected airports should be avoided whenever possible - see Tab. 3 for details. In the case of the USA RITA network, the only exception found, an acceptable error can be obtained with 84% of the airports, probably because of the way the data set was created; nevertheless, one should take into account that this data set is not representative of the whole network, but just the part created by 16 airlines. Even considering this most favourable situation, it is clear that the heterogeneity of network sizes reported in the literature (see Table 1) is likely to imply an unreliable estimation of the topological properties of air transport systems.

Sampling aircraft types or different time windows does not yield better results. As for the former, complex mesoscale structures start to appear only when more than half of the aircraft types have been included, suggesting that narrowbody and turboprop aircraft are essential to maintain the topology of the network. Similarly, structural properties are correctly estimated only when more than 40 days are taken into account, thus merging weekly and seasonal variations.

A better strategy is to select a subset of the most important airlines, especially when some of them are randomly substituted with smaller ones. In this case, even considering just 5 airlines in the European network allows, on average, the recovery of all major topological metrics - see Fig. 7. Note that, due to the random nature of the substitution process, some specific combinations may still yield biased results. The reason behind such effectiveness seems attributable to the fact that sampling airlines is equivalent to sampling both big and small airports, effectively retaining the structure of the system. Sampling airlines is thus like sampling structures that are coherent with the global structure of the system (as the latter is composed of many airlines). This suggests the existence of an optimal and coherent sampling strategy to sample networks reproducing the structure of the whole system and stabilising the metrics using the lowest possible proportion of nodes. This is confirmed in Section 3.5, in which we present a sampling process guided by the entropy of the degree distribution. By including both major and small airports, it is possible to reduce by half the size of the network, while still preserving metrics like the clustering coefficient and the efficiency.

Finally, Section 3.4 confirms that the topological bias introduced by the sampling process produces impacts on simple dynamics taking place in the network. While considering a minimal model for delay propagation, obtaining a good estimation of airport importance requires the sampling of a great fraction of nodes - 200 nodes out of 286 for achieving a 12% error.

It is important to note that the results presented here do not mean that one should not consider a network created by a subset of nodes, by a single aircraft type, or by a single day of the week. This may make perfect sense, to understand the role of a given type of connection inside the global air transport picture. What should certainly be avoided is taking such a partial representation as a proxy of the complete network. This is especially important when studying processes like delay propagation: the researcher must ascertain whether the network representation used is a complete representation of the system, and that some dimensions (*e.g.* aircraft types, airport sizes, *etc.*) are not distorted.

The sampling analysis performed here can be further extended if additional elements are considered. First of all, regional air transport networks are not independent entities, but comprise a larger scale, worldwide system; therefore, the sampling problem can affect the topological properties observed at the global scale. For instance, as the clustering coefficient decreases when more airports are included (see Fig. 3), a sparser sampling of the global network may be associated with an increase of the observed modularity - as each module, or regional network, would have a denser topology. Second, the topology of airlines is partly defined by their business model, with many legacy operators having hub-and-spoke structures, and low cost carriers usually preferring point-to-point ones. Sampling only airlines of one type may introduce further biases in the structure of the resulting network - see [29] for an extensive discussion on this topic. Third, one should be aware that the flight network is only one of the many elements contributing to the dynamics of the system. For instance, the role of passengers has extensively

5% error threshold						
Country	No. of nodes	l_d	CC	E	E_{dd}	IC
Europe - ALL-FT+	1375 (98.2%)	3.72%	0.78%	0.82%	0.90%	4.37%
USA - RITA	579 (97.3%)	4.81%	1.18%	1.81%	1.34%	0.57%
USA - OpenFlight	281 (98.3%)	3.32%	1.79%	0.66%	0.91%	1.70%
China	182 (98.4%)	3.38%	1.64%	3.33%	1.50%	4.91%
Australia	109 (97.3%)	4.38%	2.81%	0.62%	1.49%	0.29%
10% error threshold (excluding ld)						
Country	No. of nodes	l_d	CC	E	E_{dd}	IC
Europe - ALL-FT+	1350 (96.4%)	7.45%	1.51%	1.73%	1.76%	8.63%
USA - RITA	499 (83.9%)	5.71%	2.77%	5.64%	2.52%	8.31%
USA - OpenFlight	281 (98.3%)	13.80%	9.02%	3.12%	4.60%	8.86%
China	180 (97.3%)	36.54%	7.85%	9.22%	9.89%	7.89%
Australia	103 (92.0%)	3.32%	1.79%	0.66%	0.91%	1.70%

Table 3: Minimum number of airports *vs.* acceptable metric errors.

been studied within the context of delay propagation [46, 47], as passengers represents another network of connections, which does not correspond well with the simple flight network. Therefore, the problem tackled in this contribution may be encountered when managing (sampled) passenger data, or indeed, not having such data at all.

Beyond what is presented here, this analysis of statistical and dynamical properties of air transport networks opens new doors towards the understanding and measurement of the topology of complex systems in general. Similar problems can be found in other multi-dimensional real-world systems, as for instance the human brain [8, 42], for which data available are already the result of a sampling process - in this case, due to the limitations of available technology.

References

[1] M. E. Newman, The structure and function of complex networks, SIAM Review 45 (2003) 167–256.

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

[3] S. Boccaletti, G. Bianconi, R. Criado, C. I. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin, The structure and dynamics of multilayer networks, Physics Reports 544 (2014) 1–122.

[4] M. Zanin, D. Papo, P. A. Sousa, E. Menasalvas, A. Nicchi, E. Kubik, S. Boccaletti, Combining complex networks and data mining: why and how, Physics Reports 635 (2016) 1–44.

[5] L. d. F. Costa, O. N. Oliveira Jr, G. Travieso, F. A. Rodrigues, P. R. Villas Boas, L. Antiqueira, M. P. Viana, L. E. Correa Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Advances in Physics 60 (2011) 329–412.

[6] M. Zanin, F. Lillo, Modelling the air transport with complex networks: A short review, The European Physical Journal Special Topics 215 (2013) 5–21.

[7] O. Lordan, J. M. Sallan, P. Simo, Study of the topology and robustness of airline route networks from the complex network approach: a survey and research agenda, Journal of Transport Geography 37 (2014) 112–120.

[8] D. Papo, M. Zanin, J. M. Buldú, Reconstructing functional brain networks: have we got the basics right?, Frontiers in Human Neuroscience 8 (2014).

[9] W. Li, X. Cai, Statistical analysis of airport network of china, Physical Review E 69 (2004) 046106.

[10] J. Wang, H. Mo, F. Wang, F. Jin, Exploring the network structure and nodal centrality of chinas air transport network: A complex network approach, Journal of Transport Geography 19 (2011) 712–721.

[11] W.-B. Du, X.-L. Zhou, O. Lordan, Z. Wang, C. Zhao, Y.-B. Zhu, Analysis of the chinese airline network as multi-layer networks, Transportation Research Part E: Logistics and Transportation Review 89 (2016) 108–116.

[12] H. Wang, R. Wen, Analysis of air traffic network of china, in: Control and Decision Conference (CCDC), 2012 24th Chinese, IEEE, 2012, pp. 2400–2403.

[13] J. Zhang, X.-B. Cao, W.-B. Du, K.-Q. Cai, Evolution of chinese airport network, Physica A: Statistical Mechanics and its Applications 389 (2010) 3922–3931.

[14] L.-P. Chi, R. Wang, H. Su, X.-P. Xu, J.-S. Zhao, W. Li, X. Cai, Structural properties of us flight network, Chinese Physics Letters 20 (2003) 1393.

[15] Z. Xu, R. Harriss, Exploring the structure of the us intercity passenger air transportation network: a weighted complex network approach, GeoJournal 73 (2008) 87–102.

[16] P. Fleurquin, J. J. Ramasco, V. M. Eguiluz, Systemic delay propagation in the us airport network, Scientific Reports 3 (2013).

[17] T. Jia, B. Jiang, Building and analyzing the us airport network based on en-route location information, Physica A: Statistical Mechanics and its Applications 391 (2012) 4031–4042.

[18] CIA, The world factbook, Retrieved August 20 (2010) 2010.

[19] M. Alderighi, A. Cento, P. Nijkamp, P. Rietveld, Assessment of new hub-and-spoke and point-to-point airline network configurations, Transport Reviews 27 (2007) 529–549.

[20] G. Bagler, Analysis of the airport network of india as a complex weighted network, Physica A: Statistical Mechanics and its Applications 387 (2008) 2972–2980.

[21] M. P. Stumpf, C. Wiuf, R. M. May, Subnets of scale-free networks are not scale-free: sampling properties of networks, Proceedings of the National Academy of Sciences of the United States of America 102 (2005) 4221–4224.

[22] S. H. Lee, P.-J. Kim, H. Jeong, Statistical properties of sampled networks, Physical Review E 73 (2006) 016102.

[23] M. Guida, F. Maria, Topology of the italian airport network: A scale-free small-world network with a fractal structure?, Chaos, Solitons & Fractals 31 (2007) 527–536.

[24] M. Zanin, J. M. Buldú, P. Cano, S. Boccaletti, Disorder and decision cost in spatial networks, Chaos: An Interdisciplinary Journal of Nonlinear Science 18 (2008) 023103.

[25] J. Lin, Network analysis of chinas aviation system, statistical and spatial structure, Journal of Transport Geography 22 (2012) 109–117.

[26] J. Wang, H. Mo, F. Wang, Evolution of air transport network of china 1930–2012, Journal of Transport Geography 40 (2014) 145–158.

[27] Z. Neal, The devil is in the details: Differences in air traffic networks by scale, species, and season, Social Networks 38 (2014) 63–73.

[28] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, M. A. Porter, Multilayer networks, Journal of Complex Networks 2 (2014) 203–271.

[29] A. Cardillo, J. Gómez-Gardenes, M. Zanin, M. Romance, D. Papo, F. del Pozo, S. Boccaletti, Emergence of network features from multiplexity, Scientific Reports 3 (2013).

[30] L. d. F. Costa, F. A. Rodrigues, G. Travieso, P. R. Villas Boas, Characterization of complex networks: A survey of measurements, Advances in Physics 56 (2007) 167–242.

[31] M. E. Newman, Scientific collaboration networks. i. network construction and fundamental results, Physical Review E 64 (2001) 016131.

[32] M. E. Newman, Assortative mixing in networks, Physical Review Letters 89 (2002) 208701.

[33] B. Wang, H. Tang, C. Guo, Z. Xiu, Entropy optimization of scale-free networks robustness to random failures, Physica A: Statistical Mechanics and its Applications 363 (2006) 591–596.

[34] V. Latora, M. Marchiori, Efficient behavior of small-world networks, Physical Review Letters 87 (2001) 198701.

[35] M. Zanin, P. A. Sousa, E. Menasalvas, Information content: Assessing meso-scale structures in complex networks, EPL (Europhysics letters) 106 (2014) 30001.

[36] Y. Moreno, M. Nekovee, A. F. Pacheco, Dynamics of rumor spreading in complex networks, Physical Review E 69 (2004) 066130.

[37] P. Bonacich, Some unique properties of eigenvector centrality, Social Networks 29 (2007) 555–564.

[38] V. Colizza, A. Flammini, M. A. Serrano, A. Vespignani, Detecting rich-club ordering in complex networks, Nature Physics 2 (2006) 110–115.

[39] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani, Large scale networks fingerprinting and visualization using the k-core decomposition, in: Advances in neural information processing systems, 2005, pp. 41–50.

[40] A. Reggiani, S. Signoretti, P. Nijkamp, A. Cento, Network measures in civil air transport: A case study of lufthansa, in: A. Naimzada, S. Stefani, A. Torriero (Eds.), Networks, Topology and Dynamics, volume 613 of *Lecture Notes in Economics and Mathematical Systems*, Springer Berlin Heidelberg, 2009, pp. 257–282. doi:10.1007/978-3-540-68409-1_14.

[41] T. Gross, B. Blasius, Adaptive coevolutionary networks: a review, Journal of The Royal Society Interface 5 (2008) 259–271.

[42] M. Zanin, Can we neglect the multi-layer structure of functional networks?, Physica A: Statistical Mechanics and its Applications 430 (2015) 184–192.

[43] M. Barahona, L. M. Pecora, Synchronization in small-world systems, Physical Review Letters 89 (2002) 054101.

[44] M. E. Newman, Spread of epidemic disease on networks, Physical Review E 66 (2002) 016128.

[45] R. M. May, A. L. Lloyd, Infection dynamics on scale-free networks, Physical Review E 64 (2001) 066112.

[46] A. Cook, G. Tanner, M. Zanin, Towards superior air transport performance metrics–imperatives and methods, Journal of Aerospace Operations 2 (2013) 3–19.

[47] A. Cook, H. A. Blom, F. Lillo, R. N. Mantegna, S. Miccichè, D. Rivas, R. Vázquez, M. Zanin, Applying complexity science to air traffic management, Journal of Air Transport Management 42 (2015) 149–158.