Transportation Research Part E 94 (2016) 95–109

Contents lists available at ScienceDirect

Transportation Research Part E

journal homepage: www.elsevier.com/locate/tre

On the multi-dimensionality and sampling of air transport networks

Seddik Belkoura [a,b,*], Andrew Cook [c], José Maria Peña [b,d], Massimiliano Zanin [a,e]

[a] Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006 Madrid, Spain
[b] Centro de Investigación en Simulación Computacional, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón, Madrid, Spain
[c] Department of Planning and Transport, University of Westminster, 35 Marylebone Road, London NW1 5LS, UK
[d] University of Oxford, Department of Engineering Science, Parks Road, Oxford OX1 3PJ, UK
[e] Faculdade de Ciências e Tecnologia, Departamento de Engenharia Electrotécnica, Universidade Nova de Lisboa, Lisboa, Portugal

ARTICLE INFO

Article history:
Received 29 March 2016
Received in revised form 29 July 2016
Accepted 30 July 2016

Keywords:
Air transport
Complex networks
Network topology

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.

1. Introduction

During the last decade, the application of complex network theory (Newman, 2003; Boccaletti et al., 2006, 2014; Zanin et al., 2016) 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 (Costa et al., 2011); 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 (Zanin and Lillo, 2013; Lordan et al., 2014) 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

* Corresponding author at: Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006 Madrid, Spain.
E-mail address: sb@innaxis.org (S. Belkoura).

http://dx.doi.org/10.1016/j.tre.2016.07.013
1366-5545/

from the dynamics of the composing elements (the case of functional networks) (Papo et al., 2014). 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. Zanin and Lillo (2013) 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 Li and Cai (2004), 144 and 1018 in Wang et al. (2011), and 203 airports and 1877 connections in Du et al. (2016) - additional intermediate values can be found in Wang and Wen (2012), Zhang et al. (2010). Also of interest is the case of the USA air transport network, which was represented respectively by 215, 272, 305 and 732 airports in Chi et al. (2003), Xu and Harriss (2008), Fleurquin et al. (2013), Jia and Jiang (2012). 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 (CIA, 2010).

Notably, in some cases the number of airports is not even reported (Alderighi et al., 2007; Bagler, 2008). It is well known that considering a sampled version of a network (*i.e.* a sub-network) has important consequences for the observed topological features, as some properties may be lost and others may emerge in a spurious way (Stumpf et al., 2005; Lee et al., 2006). In the previously reported examples of China, for instance, the clustering coefficient varied between 0.69 (Wang et al., 2011) and 0.73 (Li and Cai, 2004); more startling is the case of the Italian network, which was reported to have clustering coefficients between 0.07 in Guida and Maria (2007) and 0.42 in Zanin et al. (2008). While the impact of a sampling process has been studied in different theoretical and applied contexts (Stumpf et al., 2005; Lee et al., 2006), it has largely been neglected in air transport.

The second problem emerges when one considers the multi-dimensional nature of the air transport system (Neal, 2014). 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 dimension is represented by different layers (Boccaletti et al., 2014; Kivelä et al., 2014). 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 dimensions may be included. Some of them, *e.g.* airlines, are trivial and have already been considered in past research (Cardillo et al., 2013); 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: