Delay propagation through the 20 busiest US airports

Alicia Gonzalez (SES 598)

Abstract

A network of airports can be modeled by assuming aspects of the real network such as aircraft rotation, flight connectivity and airport congestion, along with different kinds of initial conditions, both random and from real data. We model a network consisting of 20 airports and 2318 flights between them that take place within a 24 h period with the information related the airports and the flights being obtained from real data. Within this framework, the spread of delays within the network is studied. The model follows that presented in *Fleurquin et al.* 2013 (doi: 10.1038/srep01159). We study the informational architecture of the network following the procedures described in *Kim et al.* 2016 (doi: 10.1098/rsif.2015.0944) and *Walker & Davies* (*The hidden simplicity of biology*, submitted). The results show that the system can be modeled by introducing constant delays in the initial legs of flight instead of real data and that information flows from the most delayed airports to the rest of the network.

1. Introduction:

Transportation systems constitute some of the largest socially-driven networks. These networks have not traditionally being considered biological as they are not part of living systems or directly related to the interactions among them; however, they are directly regulated and affected by human behavior, and they arise in the first place from the interaction between human beings. A deep analysis of these kinds of networks using the same tools and parameters as in biological systems can lead to a deeper understanding of these systems and the part that the biological component that they possess play in their behavior. This project is part of a group effort to quantify the informational architecture of biological and non-biological networks and identify, if existing, the parameters that make them intrinsically different. However, this analysis is out of the scope of this paper.

The system chosen for this analysis is the airport network within the continental US. The structure and dynamics of this network are ruled by pre-established schedules that constrain the timing of the interactions and the allocation of resources. The economic nature of the system also requires a great level of efficiency, essential to make the system profitable. Under this circumstances, it is important to acknowledge the factors that can introduce inefficiencies in the system (in this particular situation, delays) and how these problems spread through the networks, affecting not only the airports where the problems arise, but also the rest of the network.

Given the computational tractability of the full airport system, it was reduced to the 20 busiest airports and the flights that directly connect them. The system is then modeled following the model presented in *Fleurquin et al.* 2013 (doi: 10.1038/srep01159). The informational architecture of the network is quantified and compared for different sets of initial conditions using the procedure described in *Kim et al.* 2016 (doi: 10.1098/rsif.2015.0944) and *Walker & Davies* (*The hidden simplicity of biology*, submitted).

2. Model description:

The network structure and time series used in this project were developed for a study of delay propagation in the US airport network and published as “Systemic delay propagation in the US airport network” (*Fleurquin et al 2013)*. This study reproduced the delay propagation patterns observed in the US performance data and studied the most relevant internal factors contributing to delay spreading through the network. It created a network with 305 airports and 2318 connections between them. A simulation for each day of 2010 was then done, and the 365 results were compared among them and to the real performance of the airport network. The study found that the average delay (among the delayed flights) was 29 minutes. This value was the starting point to two different ways of analyzing the network: an airport was considered congested if the average delay of its delayed flights was greater than 29 minutes and a day was considered to have operational problems if the average delay of all the delayed flights in the network was greater than 29 minutes.

For tractability purposes, the study presented here only includes a subset of the network of airports. In order to determine which airports to include in the study, the number of passengers that departed from each airport of the network on the chosen day was calculated. The system in the end made of the 20 busiest airports in the continental US and the 3845 flights that connect those airports. The list of nodes (in order of decreasing number of passengers) and their degrees can be found in *Table 1*. As the project spanned a single day, the only congestion measurement employed was that of individual airports. Contrary to the original project, that employs the same average delay for all days and initial conditions, the results presented here have a different threshold for each simulation, as the average delay is calculated independently for each set of initial conditions.

The databases for both the original and this project were obtained from the Bureau of Transport Statistics (http://www.btw.gov). In particular, the flight information was extracted from the Airline On-Time Performance Data, which contains flight data provided by air carriers. This archive contains information about real flights, which might differ from the original schedule. Our study is done with the flights that took off on May 28th 2010.

Table 1: List of the airports (nodes) and the number of other airports they are directly connected to (edges).

Airport	City, State	# edges	Airport	City	# edges
ATL	Atlanta, GA	19	CLT	Charlotte, NC	18
ORD	Chicago, IL	19	MSP	Minneapolis, MN	19
DFW	Dallas, TX	19	MCO	Orlando, FL	19
DEN	Denver, CO	19	SLC	Salt Lake City, UT	17
LAX	Los Angeles, CA	18	BOS	Boston, MA	19
IAH	Houston, TX	18	EWR	Newark, NJ	17
PHX	Phoenix, AZ	18	JFK	New York, NY	16
DTW	Detroit, MI	19	LGA	New York, NY	11
LAS	Las Vegas, NV	18	BWI	Baltimore, MD	19
SFO	San Francisco, CA	18	SEA	Seattle, WA	18

The network was then build based the decision tree elaborated for the original project, spanning 24 hours in 1 minute intervals. The model dynamics are based on three key aspects of the real airport network:

- *Aircraft rotation*: The same aircraft might have to complete more than one flight in the same day. This means that a flight cannot take off if the previous legs haven't been fulfilled yet, as the aircraft won't physically be there. Also, once an aircraft has landed, it is not immediately available for another flight, but it needs to be serviced. The time required to service an aircraft depends on multiple factors but, in order to simplify this procedure, it is considered that all flights require exactly 30 minutes. It is also considered that the duration of the flight is as scheduled, with no chance of making up for the delay while in the air.

- *Flight connectivity*: For some passengers, a flight will not be the first in their itinerary, but they will be connecting flights. In real life companies wait for passengers whose previous flight have been moderately delayed, so it is important that a model includes this situation as a source of delay. The way this important feature is included in the network is by calculating a connectivity factor for each individual airport, equal to the fraction of connecting passengers that travel through that airport. This calculation is done from real data, as the companies keep track of what percentage of their passengers are connecting flights. Each flight will have a probability to have to wait for connecting passengers equal to the connectivity factor of the airport it departs from. For each of these selected flights, a flight from the same airline that was scheduled to land between 3 h and 0.5 h before is randomly chosen and the flight will not be able to take off until that previous flight has landed.

- *Airport congestion*: Real life airports have a finite capacity, which means an unlimited amount of flights can't take off per unit of time. This is introduced in the model by considering that the scheduling is done in an optimal way, so no more than the number of flights originally scheduled can depart per hour. Also, two flights are not allowed to take off simultaneously, so, in case of two planes scheduled to depart at the same time, the first one who finished its service procedure would be the first one scheduled to take off.

In summary, a flight can take of if all its previous legs are completed, the aircraft has been serviced, all the connections have landed, the airport hourly capacity has not been reached and if the aircraft is the first one in the airport's queue.

Another key parameter to both these studies are the initial conditions. This network allows for two different kinds of initial conditions: from the data and random. Initializing the model from the data implies that the first flight of each series departs at exactly the same time as it did in reality. Random initial conditions can, however, be setup in several different ways, with the delay being constant, randomly selected from a time interval or following a set distribution.

Three different kinds of initial conditions were chosen in the end for the project:

- From real data: each initial segment of each flight had a delay equal to that of that same flight on May 28th 2010.
- Constant delay: a fraction of the initial segments of each flight were imposed a delay equal to the average delay of the delayed flights in the real data (42 min). This fraction was equal to the fraction of initial legs that were delayed in the real data (17%).
- Real delay distribution: a fraction of the initial segments of each flight were imposed a delay following a distribution equal to that of the real data (*Figure 1).* Once again, this fraction was equal to that of the real data (17%).

Once the time series for the whole day was obtained it was transformed into a boolean network classifying each airport as delayed/congested (“1”) or not delayed/congested (“0”). The threshold between these categories was made by calculating the average delay among all the delayed flights over the day. An airport whose average delay was higher than the global average was then considered congested.

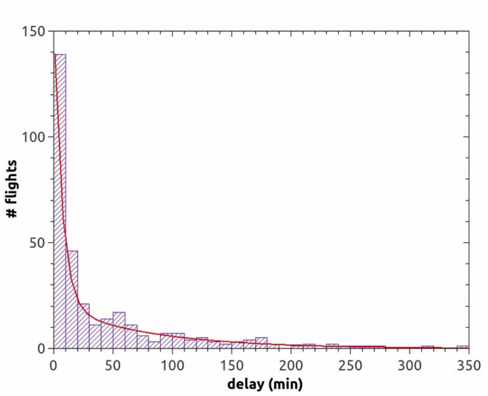

Figure 1: Distribution of the delay for the first leg of each flight during May 28th 2010 and exponential fit to the data.

In order to study the information processing, we utilize the concept of transfer entropy (TE) from a source node A to a destination node B. TE is defined as the reduction in uncertainty about the future state of airport A, knowing the state of airport B, above the reduction in uncertainty given by knowing the past states of airport A

$$TE_{B\rightarrow A} = \sum_{(a_n^{(k)},\, a_{n+1}, b_n)\in A_0} p\left(a_n^{(k)}, a_{n+1}, b_n\right) \cdot \log_2 \frac{p\left(a_{n+1} | a_n^{(k)}, b_n\right)}{p\left(a_{n+1} | a_n^{(k)}\right)}$$

where A_0 is the time series for the delays over the course of the day. The full definition and calculation are explained in detail in *Kim et. al.* 2016 and *Walker & Davies* (submitted).

3. Results:

- Scaling relation for information transfer:

We calculated the TE for each pair of nodes for the three different sets of initial conditions, and ranked the pairs according to their TE value. The scaling relations can be seen in *Figure 2*.

Figure 2 shows that networks where the delay distribution follows that of real data have more information transfer than the networks made with pure real data. In Figure 2(c) it can be appreciated that for different simulation of the same network with the same kind of initial conditions, the shape of the distribution is maintained even when the magnitude of the information transferred varies. In Figure 2(b) it is shown that for different instances of a constant delay introduced in the system the results are not identical, as they vary in both shape and magnitude. However, there is not enough data to claim that the distributions are indeed different. The data obtained from a constant delay is more similar in both shape and magnitude to the real data than the one obtained from an analogous but not real delay distribution.

These differences among different simulation with identically set initial conditions can be due to several factors and further measurements and analysis will have to be done to address this issue. It is important to note that the individual flights that are delayed are randomly selected in each simulation, so the initial conditions will only be identical for runs with the same random seed.

An important question we asked after studying the transfer entropy for each pair of nodes was if and how that scales with the average delay within each airport. For each airport we calculated the sum of all TE from itself to all the other airports and compared it to its average hourly delay (that is the average delay of all the delayed flights during that hour). *Figure 3* shows this results. It is important to note that the average delay is usually very low as most flights actually depart on time so for most hours of the day the average delay is zero. Also some of the airport don't experience any traffic during some hours of the day, making the delay automatically zero too.

- Distribution of information processing over causal structure:

The correlations between airports measured with TE can have two different origins: direct causal effect via a flight or statistical correlations with no direct connection. It is important to also take into account that TE measures correlations across both space and time.

In order to study the information transfer along edges, each pair of airports was classified as having a direct flight connection or not, along with as being correlated (TE>0) or not (TE=0). *Figure 4* shows the results for the three different kind of initial conditions: using the real delays for the first leg of each flight the day of the study, using a constant delay for those flights or including a delay distribution equal to the real data. In both the situations where the real data or constant delays were used as initial conditions, the majority of nodes are not correlated via information transfer even if they share an edge. On the other hand, when the initial conditions follow a distribution analogous but not identical to the real data, there is no significant difference in the number of pairs of nodes that are correlated via information with respect to those who are not. When it comes to nodes not directly connected by flights, no differences are appreciated for any of the different initial conditions.

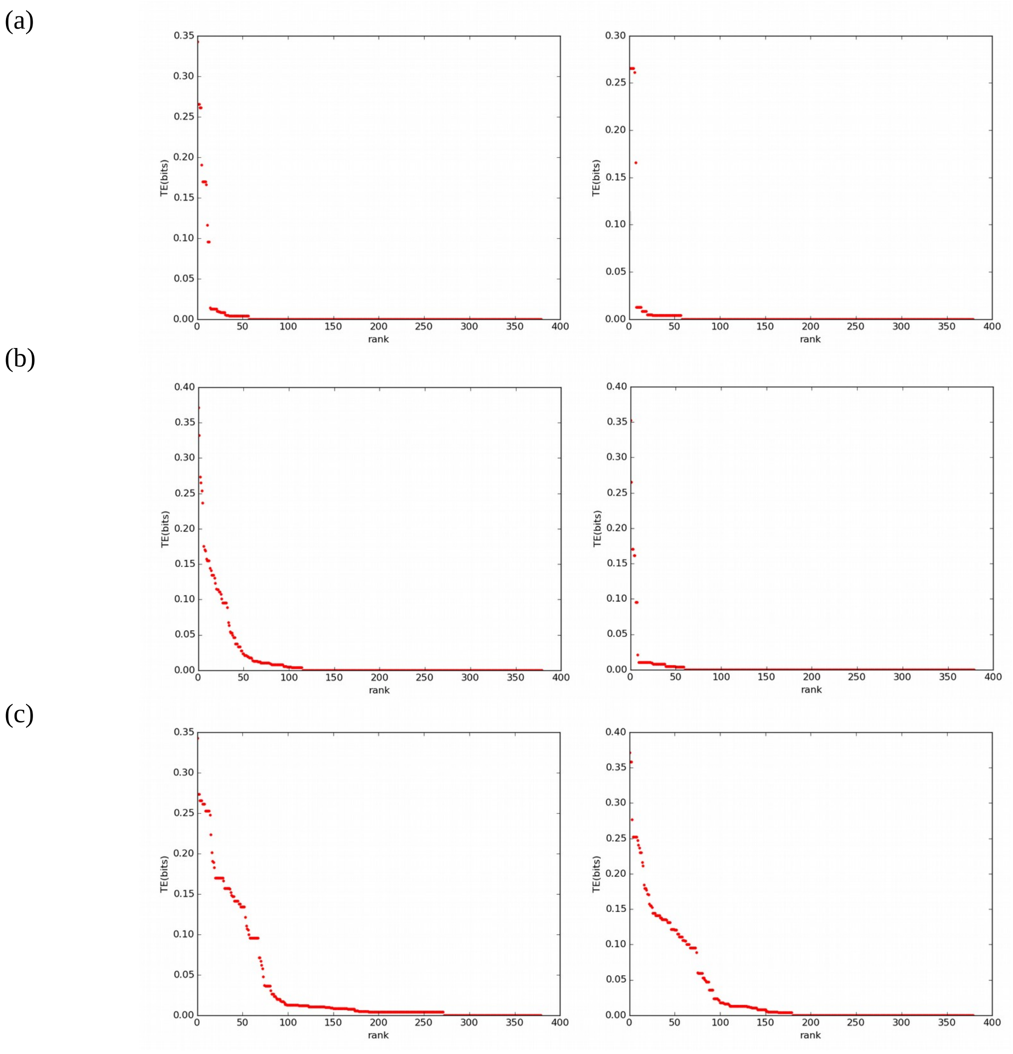

Figure 2: Scaling of transfer entropy among pairs of nodes for pairs of simulations with initial conditions: (a) from real data, (b) constant delay, (c) real delay distribution.

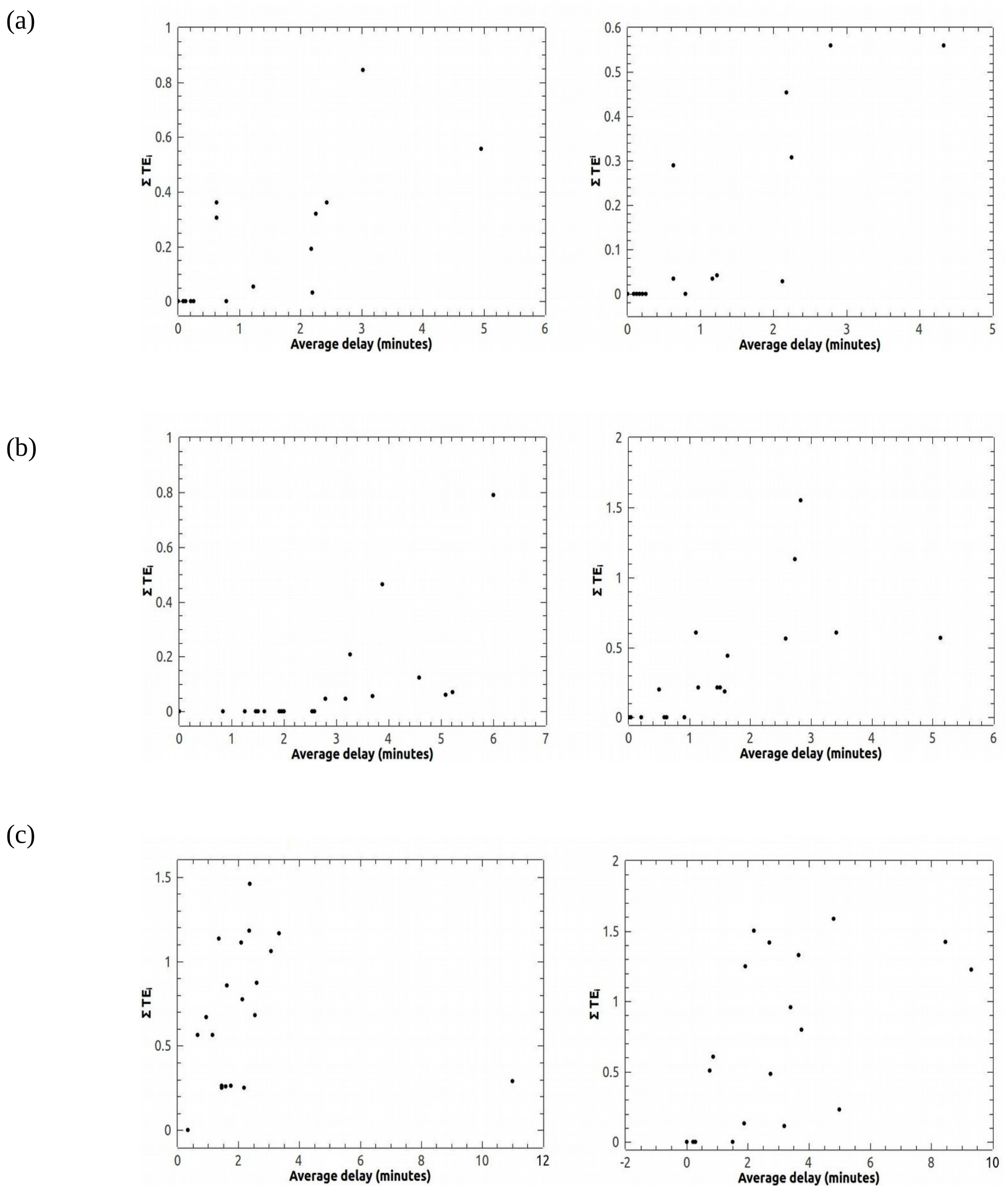

Figure 3: Total transfer entropy among pairs of nodes for pairs of simulations with initial conditions: (a) from real data, (b) constant delay, (c) real delay distribution.

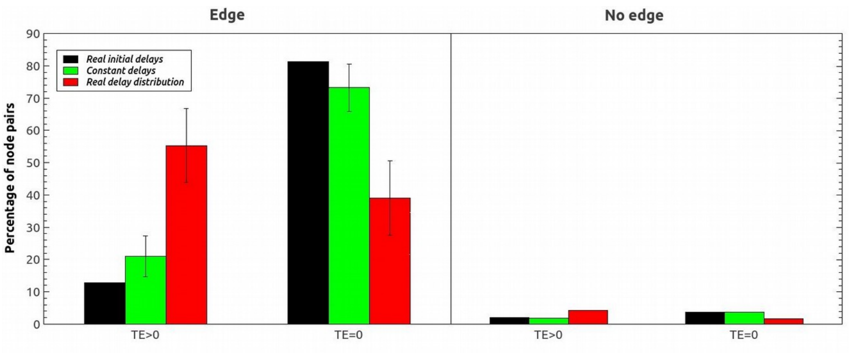

Figure 4: Classification of all pairs of nodes by correlation and causal interaction.

4. Discussion and future work:

The study of the information processing within the system and the correlation between the different nodes (airports) of the network shows that the inclusion of different initial conditions affect the dynamics of the system.

It is shown that a system where all the delayed flights are so a constant amount of time processes and transmits information in an analogous way to one where the initial conditions are that of real flights. However, some of the results obtained from the study of the information processing between nodes of the system are not conclusive and further analysis needs to be done in order to understand the inconsistencies.

The study of the total transfer entropy for each node shows that it correlates with its average delay, that is, the airports that show a highest transfer entropy also present longer delays. This result implies that information flows from delayed airports to the rest of the network, therefore propagating delays.

Looking at all the results results it can be seen that the spread of delays in a network formed by airports and the flights that connect them can dramatically depend on the initial conditions of the system. As the results show that the informational architecture of the system can be captured by the insertion of a constant delay, equal for all the delayed flights, further simulations of similar networks can be simplified, as constant initial conditions are easier to implement.

The next step for this project would be to increase of the number of airports included in the network. This will not only increase the number of nodes and edges, but also change the dynamics of the network, as the average degree of the nodes will decrease. When only considering the busiest airports, almost all of them are connected to every other airport. However, this degree of connection will dramatically decrease as less popular less connected airports are added.

Along with expanding the network, it will be important to increase the number of runs of each simulation, in order to get more reliable statistics and be able to tell apart the results due to the structure of the network and its initial conditions from statistical fluctuations.

This new features will make it possible to study control kernels in the network. This analysis could have real practical outcome as it can provide the airline companies with insights on which airports are the most critical when it comes to the spread of delays within the network.

5. References:

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

[2] P. Fleurquin, J.J. Ramasco, V.M. Eguiluz, Supplementary Information for Systemic delay propagation in the US airport network (2013)

[3] H. Kim, P. Davies, S.I. Walker, New scaling relation for information transfer in biological networks, Journal of The Royal Society Interface 12 (2015) 20150944

[4] S.I. Walker, P. Davies, The hidden simplicity of biology