D5.1 Metrics and analysis approach

Deliverable 5.1

Domino

Grant:	**783206**
Call:	**H2020-SESAR-2016-2**
Topic:	**SESAR-ER3-06-2016 ATM Operations, Architecture, Performance and Validation**
Consortium coordinator:	**University of Westminster**
Edition date:	**21 December 2018**
Edition:	**01.00.00**

SESAR JOINT UNDERTAKING

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Fabrizio Lillo / Università di Bologna	Project member	19 December 2018
Piero Mazzarisi / Università di Bologna	Project member	19 December 2018
Silvia Zaoli / Università di Bologna	Project member	19 December 2018
Luis Delgado / University of Westminster	Project member	19 December 2018
Gérald Gurtner / University of Westminster	Project member	19 December 2018

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Andrew Cook / University of Westminster	Project coordinator	20 December 2018
Lorenzo Castelli / Università degli studi di Trieste	Project member	20 December 2018

Approved for submission to the SJU By — Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
Andrew Cook / University of Westminster	Project coordinator	21 December 2018

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
N/A		

Document History

Edition	Date	Status	Author	Justification
01.00.00	21 December 2018	Release	Domino Consortium	New document for review by the SJU

Domino

NOVEL TOOLS TO EVALUATE ATM SYSTEMS COUPLING UNDER FUTURE DEPLOYMENT SCENARIOS

This deliverable is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 783206 under European Union's Horizon 2020 research and innovation programme.

Abstract

This deliverable presents the metrics proposed to assess the impact of innovations in the ATM system and a stylized ABM model, called a 'toy model', to be used as a test ground for the metrics. Existing network metrics are reviewed and their limitations are highlighted by applying them to real data. New metrics are then suggested to overcome these limitations. Their better results in measuring interconnections and causal relationships between the elements of the ATM system are shown for empirical case studies. The design of the toy model is presented and preliminary results of its baseline implementation are shown.

Table of Contents

List of figures

List of tables

Executive summary

The main goal of Domino is to characterize how different subsystems interact and to quantify the dependence structure of the interactions, and especially to understand whether they become stronger or weaker with the implementation of certain ATM innovations. This Deliverable presents metrics relevant for the evaluation of the impact of innovations in the ATM system and a simple agent-based model (ABM) model, called the 'toy model', to be used as a testing ground for the metrics.

We review some standard metrics that are usually used in the ATM community for performance monitoring and to understand the impact of different ATM solutions on the different stakeholders. We also examine network-level indicators to better capture the complex relationships among the elements of the system. To this end, we focus on two types of network metrics, namely: centrality and causality metrics. We propose several new network metrics and methods to assess centrality and causality in the ATM system. Domino will use these metrics (together with basic, classical measures) to compare simulations of the ATM system obtained from the Domino ABM under different innovation scenarios. For example, in addition, the outgoing centrality of a flight could be a tool for airlines to decide which flights to prioritize.

In a network system, centrality is a measure of the importance of a node and relies on the concept of the connectivity of the node in terms of links, paths or walks joining it to the other nodes of the network. Regarding centrality, the difference between scheduled and actual centrality using newly-defined metrics 'Trip Centrality' and 'TripRank' gives an indication of the loss of centrality of airports and of flights. When aggregated across the different elements of the system (e.g., airports), these measure the loss of performance of the whole ATM system. When considering single airports or flights, Domino could highlight the effect of innovation implementation on the airlines/airports adopting them. In the case of the simulations of the model with ATM innovations, Domino could verify whether actual and scheduled centralities of the airline implementing them become more similar, i.e., the extent to which disruption is mitigated.

Considering causality, the density of the Granger causality network is a measure of interconnection and tightness of the system (or of a part of it). Domino will investigate whether the introduction of innovations modifies the causality network making it less dense and/or more modular, i.e., how resilience is impacted. The number of reciprocated causal links and feedback triangles is a simple measure of causal feedback in the system, and its monitoring under different innovations can provide insights into their effects on the stability of the ATM system. The multilayer causality network, and specifically the density and level of reciprocity of its components, give indications of the tightness of its elements and their mutual interconnection. We implement for the first time in ATM the 'Granger in tail' metric.

Whilst we have considered the empirical application of the proposed metrics to a network of flights and airports, they can be applied much more generally, especially in other parts of ATM that can be mapped by a network or monitored with multivariate time series. One could consider as nodes the DMAN and AMAN at different airports and their queue size as the variables describing their state.

We plan to extend the existing metrics by weighting the links by considering: (i) the number of passengers traveling in the flight represented by the link or in the considered walk; and, (ii) the cost of delay to study how the network structure affects the propagation of cost and to identify the most central nodes (flights or airports) in terms of airline cost.

Two specific features of the causality network emerge: the overexpression of feedback triangles and of mutual linkages with respect to the randomized graphs. The interpretation of this result is that delay tends to propagate in both directions between a couple of airports and in triangle loops. Both these patterns tend to amplify delays in the system, and therefore innovations should aim at their reduction. Hence, we conclude that these two metrics, i.e., feedback triangles and mutual linkages in the Granger causality networks, are two important metrics to assess the impact of innovations on the system's performance. According to our initial results, large airports are more informative regarding the prediction of the state of delay of the whole system and more central for the process of delay propagation. However, when the focus is on extreme events, i.e., the states of distress of airports, information on small and medium airports is central for predictions and the characteristics of the corresponding causality network are not strongly correlated with traffic.

The results prove that the generalized metrics which we have introduced are able to tell apart different delay conditions, differently from the existing centrality metrics, and therefore they represent a suitable tool to compare the different scenarios simulated by the ABM in terms of disruptions caused by delays. Additionally, the loss of centrality provides more specific information with respect to delay statistics, as it measures the real impact of delays in terms of missed connections, a central point to assess because it entails the highest costs for airlines and disutility for passengers.

To the best of our knowledge no metrics have been previously formulated which account for the time-ordering of scheduled links nor for the different meaning of inter- and intra-layer walks. This is the reason why we have introduced a set of new centrality metrics which answer to the needs of the project. Also, to the best of our knowledge, multiple hypothesis test correction has not been applied before to Granger networks in ATM, although it has a large impact on the inferred causality networks.

A detailed description of a simple, yet controllable model of the interactions between airlines, the agents of the model, and other elements of the ATM system, such as departure and arrival managers and passengers is also being developed in Domino. This model, referred to as the 'toy model', is much simpler with respect to the ABM model, as it currently contains less details on the processes and considers only one type of agent, the airline. Preliminary results of its baseline implementation are presented.

In the next deliverables, the causality metrics presented here will be applied to both the outputs of the agent-based model and the toy model under different innovation scenarios.

1 Introduction

This Deliverable contains the results of the Task 5.1 of the Domino project. The Domino project will produce simulations from agent-based models (ABMs) of the European air traffic under current conditions and for the scenarios which have been defined in D3.2. These scenarios consider different mechanisms (4D trajectory adjustments, flight prioritisation, and flight arrival coordination) with the aim of studying the impact of the introduction of such mechanisms on ATM. To this end, the first objective of WP5, to which this Deliverable belongs, is to identify the relevant network metrics used to quantify such impacts.

For these reasons, as specified in the proposal, this Deliverable can be broadly divided in three parts:

- the first one concerns network metrics for ATM. In particular, weaknesses and strengths of existing network metrics are reviewed. Of note, new network metrics are introduced and their superior performance in capturing interconnection, tightness, and causal relationships between elements of the ATM systems is justified theoretically.
- the second one shows the performance of the existing and new network metrics by applying them to large datasets of European and US air traffic. The superior performance of the new metrics is confirmed by these empirical case studies.
- the third part contains the development of a stylized ABM, termed a “toy model”, which is calibrated on scheduled traffic data and can be used as a testbed for the developed metrics.

1.1 ATM networks

Air traffic can naturally be described as a networked system and in recent years a large number of studies have applied concepts from network science to ATM (for a recent review, see [1]. The advantage of a network description is clearly that it is possible to focus on the role of interactions between the elements or between the parts of the system, the role of network topology in the propagation of signals, distress, congestion, etc. Since the focus of Domino is to understand the impact of the new mechanisms at a global level, the use of network science is a natural choice.

The ABM developed in Domino will consider the different elements of the airspace, i.e., passengers, airlines, aircraft, network manager, arrival and departing manager, etc. The interaction between these different elements will be modeled with networks whose nodes are used differently for different stakeholders. The links between these nodes describe the interaction between them or the indirect effect of the state of a node on the state of another node. Metrics will be used to extract relevant information from these networks.

In order to test the existing metrics and to develop new ones, in this Deliverable we decided to focus on a specific, yet important, ATM network, namely the one whose nodes are airports and the links are the flights between them. This choice is in part dictated by the data availability. However, this is also an important system, which has been investigated in the past and that share the main property of being composed by a planned network and a actual network. We are confident that the metrics and methods developed for this system will be very useful when we will investigate the more heterogeneous network obtained from the simulation of the Domino ABM. The reason is that we believe the concepts of interconnection, centrality, and causality (see below) that we will consider in the following are generally important for the research questions of the Domino project.

Specifically, in this Deliverable we consider two different airspaces (ECAC and US) and we review a large number of existing network metrics by investigating their application on the two datasets. In particular, in the Deliverable, apart from the standard interconnection metrics (density, degree distribution, clustering coefficient, assortativity, etc), we focus on two types of network metrics that we believe are relevant for the Domino project: centrality and causality metrics.

Centrality. Centrality is a key property of a node in the network. Generally speaking, the centrality of a node describes its 'importance', either in terms of the number of connections, or in terms of its role in getting the network connected, or in terms of its role in the shortest paths connecting other nodes. Centrality describes several different properties and therefore many different metrics have been proposed. In the context of the Domino project, it is important to study the centralities of the nodes of the ATM networks to investigate whether the introduction of the new mechanisms (possibly in a subset of the nodes, e.g., airports) drastically changes the centralities of all the nodes. This is particularly relevant when one considers that there are two networks which can be associated with air traffic: the first one considers the scheduled flights (including the times of departing and landing) and the other one considers the actual flights. The relevant question is whether the introduction of the new mechanisms makes the system (of parts thereof) more robust, in the sense that an airport preserves its centrality also when the system is perturbed and many flights are delayed.

In reviewing the existing network metrics, we actual that many of them are not suitable for such comparison [2]. The main reason is that most of such metrics were developed for static and single layer networks, while air traffic is naturally described by a temporal multilayer network. The air traffic network is temporal because links (flights) appear and disappear, and connections between two flights are possible only if the corresponding links appear in time in the right order. Moreover, the network is multilayer because each layer describes a single airline (or alliance) and the effect of a missing connection on the centrality decrease of an airport is very different if the two flights belong to the same or different airlines (for example because of reactionary delays or passengers changing airlines). Also, the existing network metrics for temporal networks and multilayer networks do not seem to be suitable for ATM.

For these reasons, in this Deliverable we develop three new centrality metrics for temporal multilayer networks, which we termed 'Trip centrality', 'TripRank' and '2-legs Trip centrality'. Their advantage is that they have a clear operational meaning in terms of paths in the air system.

Causality. As mentioned above, networks are the natural structure to study the propagation of disturbances (or information) in a complex system. In the ATM case, where the nodes of the network are the airports, an interesting problem is to study how the congestion state of an airport, for

example because there is a large number of delayed flights, propagates to the other airports. In the context of Domino, the relevant question is whether the introduction of the new mechanisms makes the propagation of the congestion state more or less likely, and the paths of propagation. For example, if UDPP is implemented in few airports, will the congestion propagate less likely in the connected airports?

In this Deliverable we decided to study the propagation of congestion by using a well established concept in time series analysis, namely Granger causality. In a nutshell, a time series 'Granger causes' another one if the past history of the former helps to predict the future of the latter. Since we are interested in direct and indirect contagion effects, we will perform a statistical test for causality between all the pairs of airports. The detected causality between the nodes (airports) constitute a Granger causality network. This approach has been considered in ATM in some recent papers (see the literature review in the corresponding Section).

In this Deliverable we introduce several innovations to the Granger causality network methods. First, since these methods are based on statistical testing, we show that taking into account the fact that a large number of tests is performed is critical to obtain clear results. Second, differently from the standard Granger method (also applied in ATM) which considers the possibility of predicting the average behavior, we implement for the first time in ATM the 'Granger in tail' metric (which we define in Section 2.2.2.1). The idea is to focus on extreme, rather than common, events and to identify causality relationships between them. In the ATM context, this means to identify whether the congested state of an airport helps predicting the congested state of another one. The third innovation is the construction of multilayer Granger causality networks: the ATM system is composed by many interacting parts and each part can be modeled as a network, thus the whole system can be seen as a multilayer network. For example, in the investigated system, each layer is the network of an airline. We extend the concept of Granger causality to the case of a multilayer system, where therefore the causal connection between the different layers can be seen as a measure of the tightness of their mutual interaction.

1.2 Toy model

In order to test the performance of the metrics not only with real observations but also with synthetic data, we developed a simplified model of the airports-flights network. The idea is to minimize the number of parameters and assumptions, but by calibrating the model on real scheduled data, simulate the actual flights with simple decision making processes of the airlines/pilots. If the model is able to reproduce the statistical properties of the real actual flights, we can test on simulated data the performance of the proposed metrics in the current and future scenarios. This will constitute a testbed for the application of the proposed metrics to the simulations of the full Domino ABM.

In the version of the model presented in this Deliverable, only the current scenario is considered. The whole ECAC space is simulated by calibrating it with the scheduled times for one day in September 2014. With a minimum amount of behavioral assumptions and by calibrating the statistical distribution of the percentage delay in the en-route phase, we are able to reproduce remarkably well the statistical properties of delays and airport congestions. Several of the baseline metrics investigated in the data analysis part of this Deliverable show good agreement with the ones obtained from the simulations of the toy model. We are therefore confident that our model is a

significant asset to test the new and more sophisticated metrics, by considering the current and the future scenarios with the implementation of the different mechanisms. This will be done in future Deliverables.

Finally, the choice of using delays and not, for example, costs for the validation of the model is dictated by the fact that delays are observable while costs are not (at least to us) and must be estimated. We believe that to validate a model is much better to use first observable quantities. However, even if we consider here only delays, we have calibrated airline costs by using data from recent literature. In the next Deliverables we will consider explicitly statistics and metrics related to costs, by comparing the estimations from real and simulated data.

1.3 Organization of the Deliverable

The Deliverable is organized as follows. In Section 2 we review existing metrics and present the new ones. In Section 2.1 we review classical basic metrics used in ATM, while Section 2.2 is devoted to network metrics. Specifically, Section 2.2.1 concerns centrality metrics, and Section 2.2.2 causality metrics. In Section 3, existing and newly proposed network metrics are applied to two datasets, presented in Section 3.1. Section 3.2 presents a simple descriptive analysis of the data, while Sections 3.3 and 3.4 focus, respectively, on the application of centrality and causality metrics. Section 3.3.1 highlights the limits of the existing centrality metrics, and the following Section 3.3.2 shows the superior performance of the proposed ones. Section 3.4.1 specifies how the causality methods presented in Section 2.2.2 are applied to the specific case study, and the results of the analysis are shown in Section 3.4.2, with particular attention to the difference between existing methods and the improved ones proposed in this deliverable. Section 3.3, finally, summarizes the new metrics proposed and indicates future perspectives. Section 4 is devoted to the toy model. The model designed is illustrated in Section 4.1, by explaining, e.g., who are the agents and subsystems modelled (Section 4.1.2), what are the model inputs (Section 4.1.3), how is a day of operation simulated in the model (4.1.4). Section 4.2 presents a baseline implementation of the model, which results are presented and analysed in Section 4.3.

2 Metrics

To assess the results of the agent-based model (ABM) presented in WP4, comparing its outcomes in different investigative and adaptive case studies, it is necessary to establish a set of tools able to characterise such outcomes and highlight and quantify their differences. To this end, we first consider, in Section 2.1, classical metrics strictly pertaining to the ATM field. Such metrics are a useful starting point to evaluate a model outcome. However, in order to analyse the network impact of the different mechanisms, additional and more detailed insight on the outcomes can be obtained from the theory of complex networks. In fact, from the outcome of the ABM we can obtain a network description of the ATM system, where nodes may be the different system's elements (e.g., airport, E-AMAN) and links are flights or passengers' itineraries. While the network of scheduled flights and passenger itineraries is fixed for a chosen day, the results of the ABM determine the actual interaction of the network's nodes, e.g., the actual delays and costs of flights, in the case study considered. Delays and cancellations influence the network connectivity by disrupting possible connections. Additionally, delays and costs propagate through the network, creating the observed situation.

Complex network theory provides metrics able to characterize the functionality of the network from different points of view. In Sections 2.2.1 and 2.2.2 we focus on two types of network metrics of particular relevance to the ATM network, i.e., centrality and causality metrics. In this Section, existing metrics are reviewed, and their limits are highlighted. New metrics are then proposed to overcome such limitations.

2.1 Classical basic metrics

There are a set of metrics that are usually used in the ATM community. These metrics can be grouped by different areas and stakeholders and are considered when assessing the performance of the system. However, they lack a network view of the system. Also, it is common to consider average values when it has been shown that the distribution of the values is critical to understand the system performance. This is particularly true in the case of delay and cost of delay due to the non-linearity between them [3].

SESAR identifies 6 different key performance areas (KPAs) with different key performance indicator (KPIs) that need to be monitored in order to assess the impact of the different solutions. Table 1 summarises these.

Table 1. SESAR performance KPAs/KPIs (adapted from [4])

Key performance area	Key performance indicator

Cost efficiency: ANS productivity	• Gate-to-gate direct ANS cost per flight
	  ○ Determined unit cost for en-route ANS
	  ○ Determined unit cost for terminal ANS

Operational efficiency	• Fuel burn per flight (tonne/flight) • Flight time per flight (min/flight)
Capacity	• Departure delay (min/dep) • En-route air traffic flow management delay • Primary and reactionary delays all causes • Additional flights at congested airports (million) • Network throughput additional flights (million)

Environment	• CO2 emissions (tonne/flight)
	  ○ Horizontal flight efficiency (actual trajectory)
	  ○ Vertical efficiency
	  ○ Taxi-out phase

Safety	• Accidents with ATM contribution
Security	• ATM related security incidents resulting in traffic disruptions

These indicators allow us to monitor the performance of the ATM system at a very high level and define political goals. The Domino model will be able to generate low level indicators that will help to describe trade-offs and the performance of the system for the different stakeholders in the ATM. The different stakeholders that have been identified are:

- ANSPs
- Airports
- Airspace users
- Passengers
- Environment

Table 2 summarises different metrics per stakeholder that will be generated by the ABM model and that have been explored in previous research [5], [6].

Table 2. Basic metrics per stakeholder

Stakeholder	Metrics
ANSP	• En-route airspace charges revenues

Airport	• departing queue delay
	• arrival queue delay
	• number of operations
	  ○ departures
	  ○ arrivals

Airspace users	• flight departure delay
	• flight arrival delay
	• fuel
	• delay per flight segment
	• reactionary delay
	• ATFM delay
	• gate-to-gate time
	• cost of delay
	  ○ non-passenger related
	  ○ passenger related (hard and soft)
	• cost
	  ○ en-route charges
	  ○ fuel cost

Passengers	• departure delay • arrival delay • missed connections • connecting time • gate-to-gate time
Environment	• fuel kg • CO_2 tonnes

As identified in previous research and previously mentioned, average metrics are not always representatives of the performance of the system, and therefore, for each metrics not only their average but their distribution needs to be considered.

Also, it has been pointed out several times how similar metrics (e.g., delay) could be experienced very differently by different stakeholders and in particular the differences between flight-centric and passenger-centric metrics. For example, reductions in flight arrival delay with passenger arrival delay map close to a 1:1.3 ratio [6]. That is, on average, one minute of flight delay corresponds to 1.3 minutes of delay per passenger (in this model; other projects have reported similar values). This is due to the fact that the delay experienced by passengers is higher due to missed connections (ibid.).

This is one of the main reasons why Domino will focus on describing not only flight-centric metrics but also passenger-centric ones. We also suspect that the relationship between the elements in the ATM system might be different for different stakeholders and the link between elements might be different from the flight or passenger perspective.

The metrics described here are useful for performance monitoring and to understand the impact of different ATM solutions on the different stakeholders and their trade-offs. However, they do not address the complexity of the network or provide information on how the different elements are related in the system.

2.2 Network metrics

In this Section, we move to network-level indicators. With respect to the classical metrics presented in the previous Section, network metrics often provide more specific understanding of the system-wide implications of changes in the elements of the system implemented in the different case studies. In particular, we will focus on two types of network metrics: centrality metrics and causality metrics.

Centrality (of a node, a link, or a group of nodes) is a key property in network science, quantifying the 'importance' of an element in the overall network architecture. Examples of questions related to the importance are: how influential is a person in a social network? How critical is an element in an infrastructure network? What is the disease spreading capacity of an individual? What is the most systemically important financial institution? Since the definition of importance depends on the context and on the questions raised, many different metrics have been proposed in the literature. In this deliverable, we will review existing centrality metrics, both for static and for temporal networks (i.e., where links appears and disappears in time). In particular, we are interested in developing new centrality metrics for temporal networks, which explicitly take in into account the temporal order of links and therefore consider network paths which are time consistent. As an important example, as explained in detail in Section 2.2.1, we shall consider centrality metrics measuring the connectivity of airports, i.e., the passenger's potential of moving through the network passing through a particular airport. Therefore, centrality metrics help to assess if innovations improve the network connectivity from the passenger point of view.

Causality is an important, yet elusive, concept in many disciplines. Knowing that the state of a particular subsystem 'causes' a state in another subsystem is important, even if there is no consensus on the way of identifying causality. Since many ATM variables, both from real data and from the simulations of the ABM, can be cast in time series, in this Deliverable we will consider a widespread and operative definition of causality, termed Granger causality. Broadly speaking, the idea is that a time series Granger causes another one if the past history of the former helps to predict the second one, even when one considers as explanatory variable the past history of the second. This very general definition can be applied to many different contexts, depending on the forecasting model adopted (e.g., linear or non-linear models), and moreover can be specialized by choosing which quantity to forecast: for example, are we interested in forecasting the average behavior of the second variable or the occurrence of an extreme event? Finally, when, as in ATM, the system is composed of many interacting units, each of them represented by a time series, one can test for causality between all the pairs of variables. Each statistically significant causality is associated with a directed link between two nodes representing the two tested variables. One obtains a Granger causality network, whose topology gives important information on the existence of 'causal hubs', i.e.,

variables which cause the behavior of many others and the presence of causality loops which can amplify disturbances. We believe that understanding causality networks is critical to test the degree of interconnection and the tightness of the ATM system. In this Deliverable we provide a concrete empirical example by considering the propagation of delays between airports, with the aim of measuring whether the fact that a certain airport is congested causes other airports to be congested. Such metrics will therefore help evaluate if innovations are successful in decoupling the system elements, reducing the propagation of delay.

Both for centrality and for causality, we first review existing metrics and highlight their limits in view of this project's scope. This is clearly shown by considering as an important case study the delays in the airport network, backing up our conclusions with specific applications to air traffic data. Then, new metrics are proposed to overcome the identified limits and validated on historical data.

2.2.1 Centrality metrics

2.2.1.1 Existing centrality metrics

Centrality is a measure of the importance of a node in a network. While several different definitions of centrality exist, all centrality metrics are based on some concept of connectivity of a node in terms of links, paths or walks joining it to the other nodes of the network. For example, the larger the centrality of an airport is, the higher is the potential to move through the network passing through that node. As a consequence, the loss of centrality of an airport, between the scheduled and the actual network, signals a diminished potential of moving through the network passing through that node, which means, from both the passenger and the airline point of view, a diminished performance of the network. In the light of Domino's scope, this loss of centrality should reflect not only the missing links due to cancellations but also the disrupted paths due to delays, since innovations mainly aim at reducing such disruptions. Provided a centrality metrics satisfying these requirements, comparing the loss of centrality between the actual and the scheduled flight network among case studies implementing different mechanisms would allow us to assess the impact of innovations on the network performance. In particular, an innovation diminishing the centrality losses between the scheduled and actual network represents an improvement both from the passenger point of view and from the airline point of view, as it means that less itineraries were disrupted by delays, implying less inconveniences for the passengers and smaller costs for the airlines.

A network is a set of nodes and links. A link between two nodes is said to be directed if it has a defined direction. A link can also be weighted, when it is associated to a number describing its properties. For example, the network of airports and flights is a directed network, where a flight from airport *i* to airport *j* is represented by a link going from node *i* to node *j*. While centrality metrics can be applied to any network, here we will introduce them focusing for clarity on this specific network.

As we will explain in more detail in the following, the most general representation of such network is dynamic in time, as links appear and disappear according to the schedule. Moreover, it might have a multi-layer (or multiplex) structure, where each layer contains a certain type of links between nodes. For example, in the case of the network of flights and airports, each layer contains the links

corresponding to flights of a different company or alliance, and inter-layer links connect nodes corresponding to the same airport.

However, commonly used centrality metrics apply to single-layer static networks. To review existing metrics, therefore, we start by neglecting the temporal and multiplex structure. In our example, this corresponds to considering the network of flights and airports aggregated across layers, i.e., across airlines, and across time frames, i.e., where all flights operated on the day chosen for the analysis are present at the same time regardless of their schedule. In this network, an edge is present from *i* to *j* if at least one flight connects them. A weight *k* is assigned to the edge, where *k* is the number of flights going from *i* to *j*.

Let A be the weighted adjacency matrix of the network, such that A_{ij} =k if there are *k* links going from *i* to *j*, and $\tilde{A}$ be the non-weighted adjacency matrix, such that $\tilde{A}_{ij} = 1$ if there is at least one link going from *i* to *j*, and zero otherwise. Here, we consider some among the most common and well known centrality metrics: degree, strength, Katz, PageRank, closeness and betweenness centralities. When the network is directed, a distinction should be made between incoming and outgoing centrality.[1]

The incoming (outgoing) degree centrality of a node *i* is given by the number of incoming (outgoing) edges, $d_i^{IN} = \sum_j \tilde{A}_{ji}$ (and $d_i^{OUT} = \sum_j \tilde{A}_{ij}$), where the index *j* runs on all nodes. This centrality measure determines from how many nodes node *i* can be reached (respectively, how many destinations can be reached from node *i*) with one link. The incoming (outgoing) strength of a node *i*, instead, is given by the total weight of incoming (outgoing) edges, $s_i^{IN} = \sum_j A_{ji}$ (and $s_i^{OUT} = \sum_j A_{ij}$). It measures with how many links node *i* can be reached (respectively, how many links depart from node *i*). However, an important feature for network connectivity are paths which make use of two or more links. A commonly used metric which considers a node's centrality to depend on the walks of any length arriving to (or departing from) that node is Katz centrality [7] [8]. The incoming Katz centrality of node *i* is

$$k_i^{IN} = \sum_j (\mathbb{I} - \alpha A)^{-1}{}_{ji} = \sum_j \sum_{n=0}^{\infty} \alpha^n (A^n)_{ji},$$

that is, each walk of length *n* from any node *j* of the network to *i* contributes α^n to the centrality of *i*. Note that, by using the weighted adjacency matrix in this calculation, for each different link present between two nodes a different walk is counted. Since $\alpha < 1$, longer walks contribute less and its value determines the relative importance of walks of different length[2]. The coefficient α must be smaller than the inverse of the largest eigenvalue of A for the expression to converge [8]. Correspondingly, the outgoing Katz centrality of node *i* is

$$k_i^{OUT} = \sum_j (\mathbb{I} - \alpha A)^{-1}{}_{ij} = \sum_j \sum_{n=0}^{\infty} \alpha^n (A^n)_{ij},$$

PageRank is a generalisation of Katz centrality, developed by Google, that introduces an additional weight to the paths, depending on the in- (or out-) degree of the nodes they cross. Specifically,

[1] Note that this is not a complete review of centrality metrics.

[2] Specifically, 1/α walks of length n-1 are needed to give the same contribution to centrality as a single walk of length n.

$$pr_i^{IN} = \sum_j (\mathbb{I} - \alpha D^{-1} A)^{-1}{}_{ji},$$

where $D_{ij} = \delta_{ij} {d_i}^{OUT}$, so that a link from *j* to *k* is weighted by the inverse of the out-degree of *j*, $1/{d_i}^{OUT}$. In terms of airports and flights, the meaning of this generalization is that an airport with an inbound flight coming from a large airport, with a large out-degree, will inherit a fraction of its centrality proportional to the inverse of such out-degree. In other words, the more outbound flights an airport has, the less of its centrality the destination airports inherit.

Betweenness centrality, instead, measures how often a node is used in the shortest paths joining any two nodes. The shortest path between two nodes is the path using the smaller number of links. The betweenness centrality of node i is computed as

$$b_i = \sum_{j,k \neq j} \frac{{\sigma^i}_{jk}}{\sigma_{jk}},$$

where $\sigma_{jk}$is the number of shortest paths between the nodes *j* and *k*, and ${\sigma^i}_{jk}$is the number of such paths which pass by *i*.

Finally, closeness centrality measures how close a node is, on average, to the other network nodes. Calling $d_{ij}$the distance between node *i* and node *j* (the length of the shortest path between them), the closeness centrality of node *i* is obtained as the inverse of the average distance of node *i* from other network nodes:

$$c_i = \frac{n-1}{\sum_{j \neq i} d_{ij}}.$$

Therefore, the higher the closeness centrality of a node is, the closer it is to other nodes. However, if the network is formed by more than one component, the distance $d_{ij}$is infinite when the nodes *i* and *j* fall in two different components. Therefore, for network with more than one component, $c_i = 0$ for all nodes. To overcome this issue, closeness centrality can be defined using inverse distances:

$$c_i = \sum_{j \neq i} \frac{1}{d_{ij}}.$$

These existing centrality metrics neglect the dynamic structure of the network and its multiplex structure. As we will show in Section 3.3, for the case of the network of airports and flights, this means that the centrality metrics are not able to characterize the effects of the delays on the network connectivity. This result makes clear that these metrics are not suitable to evaluate the effect of the innovations addressed by Domino on the network performance, as they would not be able tell apart a situation where delays disrupt connections to one where they do not. In any network where links evolve in time, to discern these situations, a node centrality should reflect its participation to walks that can actually be travelled, i.e., respecting the schedule, so that disrupted connections imply a centrality drop. This, in turn, requires accounting for the temporal structure of the network. Katz and PageRank centrality, in our specific case, count walks on the network which are not time ordered and therefore have no relationship with the trajectories that passengers could travel. As a consequence, these metrics cannot reflect the effect of delays on the network's connectivity. An additional limitation of Katz and PageRank centrality is that the weight assigned to

each walk does not consider to which airline each flight composing the walk belongs, therefore a walk using only flights of one airline has the same weight of a walk of the same length using several airlines. However, a more realistic assumption would be that the latter contributes less to centrality, or not at all, as it is travelled with a smaller probability. Accounting for this requires considering the multiplex structure of the network.

2.2.1.2 Proposed centrality metrics

In order to formulate a centrality metric quantifying the loss of connectivity of a node due to changes in the link dynamics (for example, delays in the actual network of airports and flights), the temporal and multiplex structure of such network must be considered. A short introduction to temporal metrics and to multiplex, sufficient for the understanding of the following analyses, is presented in the next section. While metrics applying to temporal network and metrics for multiplexes have been proposed (see e.g., [9], [10], [11], [12]), and will be succinctly reviewed in the following, to the best of our knowledge no metrics have been formulated which account for the time-ordering of scheduled links nor for the different meaning of inter- and intra-layer walks. None of the existing metrics, therefore, is appropriate for our scope. This is the reason why we introduce a set of new centrality metrics which answer to the needs of the project.

Temporal network

A temporal network is a network where links are dynamic, i.e., they appear and disappear depending on the time at which the network is observed. The network will therefore look different at different times (see Figure 1). In the case of the network of airports and flights, the presence or absence of a link is determined by the schedule of the corresponding flight. In general, a link in a temporal network is defined by the time t^* at which it appears and by its duration Δt^*, which can be zero if the contact is instantaneous. A way to work out a graph representation which allows us to easily generalise standard results of network theory consists in defining a time resolution Δt and building a time-labeled adjacency matrix $A^{[t]}$ containing all links such that at least one of the following conditions is satisfied:

$$t \leq t^* < t + \Delta t$$

$$t \leq t^* + \Delta t^* < t + \Delta t$$

$$t^* < t \text{ and } t^* + \Delta t^* \geq t + \Delta t$$

i.e., all links such that either the link appears in the interval $[t, t + \Delta t)$, or it disappears in that interval, or it is present during the entire interval. With this representation, the temporal network is defined by a time series of adjacency matrices $\{A^{[t]}\}_{t-1,\ldots,T}$.

In a temporal network, walks and paths are time-oriented. A temporal walk from node *i* to node *j* is defined as a sequence of M edges $\{i,k\}_{t_1},\{k,l\}_{t_2},\ldots.\{m,j\}_{t_M}$ where the times $t_1,\ldots,t_M$ are increasing, i.e., $t_1 < t_2 < \ldots < t_M$. A temporal path is a walk for which each node is visited at most once.

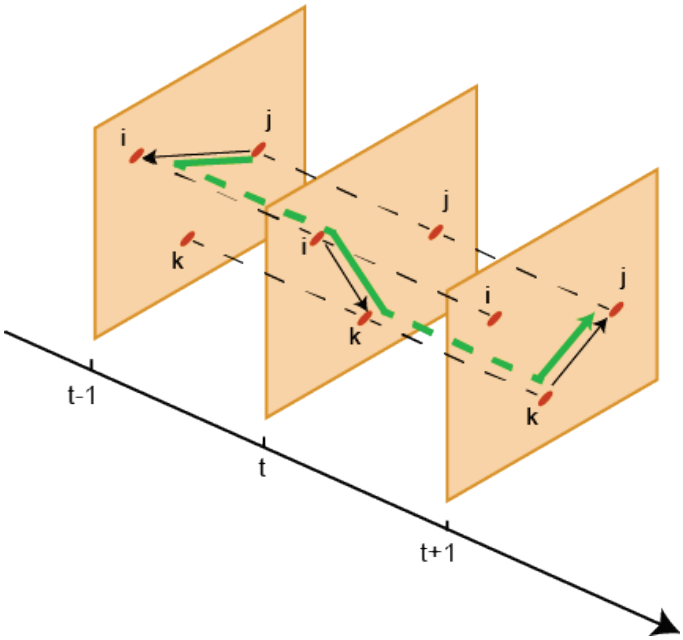

Figure 1. Example of temporal network .The network is represented at three subsequent time-steps of the discretized timeline. A time-ordered walk from node j to i, then k, and finally back to j is highlighted in green.

Multiplex

A multiplex is a network which consists of many layers, on each of which the same set of nodes are present (see Figure 2). A multiplex has two types of links, intra-layer links and inter-layer links. An intra-layer link joins two nodes of the same layer, while an inter-layer link joins two copies of the same node on two different layers. In the case of the network of airport and flights, one layer is present for each airline (or each alliance, as motivated in the following), and directed intra-layer links on the layer λ represent the flights of airline λ. Non-directed inter-layer links connecting each airport to all its 'copies' in another layer allows us to have walks which use more than one layer, i.e., flights of several airlines. When considering walks on a multiplex, we call intra-layer walk a walk that consists of only intra-layer links, while we call inter-layer walk a walk that includes at least one inter-layer link. In the example of Figure 2, node *i* can be reached by node *k* either by an intra-layer walk on layer λ , or by an inter-layer walk using the link $k \rightarrow j$ on layer μ , then passing to layer λ with the inter-layers link joining the two copies of node j and finally using the link $j \rightarrow i$ on layerλ.

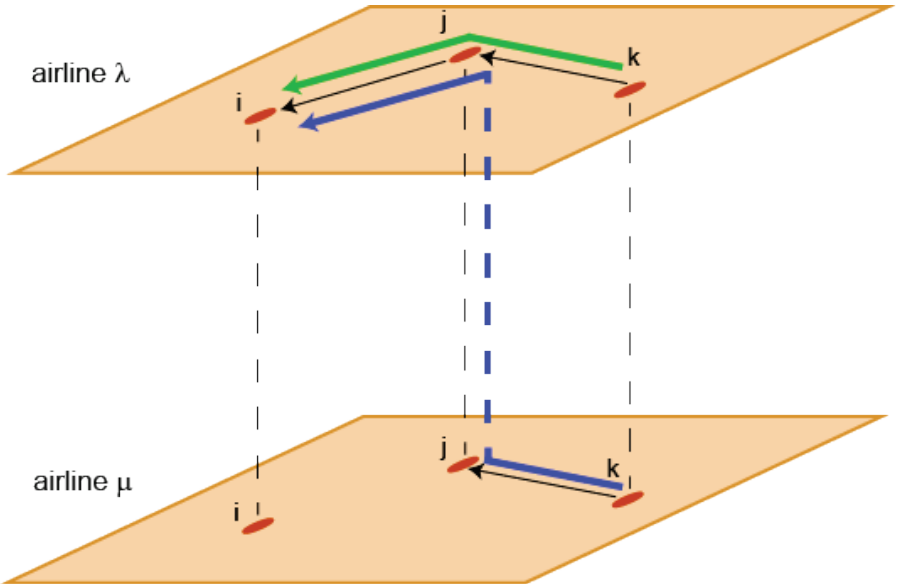

Figure 2. Example of multiplex network with two layers. Inter-layer links are represented by solid lines, intra-layer links by dashed lines. An intra-layer walk is highlighted in green, and in inter-layer one in blue.

Metrics on temporal network and on multiplexes

This subsection contains a review of existing metrics on temporal networks and on multiplexes.

In a temporal network [13], as we have seen, walks and paths are time oriented. In addition to the standard notion of distance between two nodes, given by the number of links used by the shortest path joining them, it is possible to define also the temporal distance d_{ij}between two nodes as the temporal length of the shortest path between them, where now the shortest path is the shortest in terms of duration. The *time-diameter* of the network is, then, the largest time distance between two nodes of the network. The *characteristic temporal path length* is the average time distance between two nodes in the network. However, in the network of airports and flights for a given day, there are certainly couples of airports that are not linked by a time-ordered path, therefore many of these temporal distances are infinite. Therefore, it is useful to introduce the *temporal global efficiency*, which is the average of the inverse of the time distances for all couples of flights. The largest the efficiency, the shortest time it takes on average to go from one node to another.

Centrality metrics have also been proposed for temporal networks. In an analogy with the standard *betweenness centrality*, a betweenness centrality for temporal network can be defined as

$$b_i = \sum_{j,k \neq j} \frac{{\sigma^i}_{jk}}{\sigma_{jk}},$$

where $\sigma_{jk}$is the number of shortest paths between the nodes *j* and *k*, and ${\sigma^i}_{jk}$is the number of such paths which pass by *i*. *Closeness centrality* can be extended to temporal networks as well, but recalling that in the network of airports and flights many temporal distances are infinite, it is convenient to define it using inverse distances:

$$c_i = \sum_{j \neq i} \frac{1}{d_{ij}}.$$

Finally, in [10] a generalization of Katz centrality for temporal networks is suggested, such that the incoming and outgoing Katz centrality of node *i* are $k_i^{in} = \sum_j Q_{ij}$ and $k_i^{out} = \sum_j Q_{ji}$, where the matrix Q is defined as

$$Q = (1 - \alpha A^{[0]})^{-1} \cdot \ldots \cdot (1 - \alpha A^{[T]})^{-1}.$$

The matrix Q contains all products of the form $\alpha^l A^{[t_1]} \cdot \ldots \cdot A^{[t_l]}$, where $t_1 \leq t_2 \leq \ldots \leq t_l$. However, this way of counting walks still does not fully account for the scheduled nature of the network. For example, if a flight from *i* to *j* departs at t_0 and lands at $t_2$and a flight from *j* to *k* departs at t_1 such that $t_0 < t_1 < t_2$, the path from *i* to *k* will be counted even if it cannot be actually traveled, as at time $t_1$both links are present. This problem is solved by the generalization of Katz centrality which we present in the following.

Regarding multiplex network, [11] propose several generalizations of standard network metrics. A vectorial *degree* can be defined as $\overline{d}_i = (\boldsymbol{d_i}^{[\mathbf{1}]}, \ldots, \boldsymbol{d_i}^{[\boldsymbol{M}]})$, where $\boldsymbol{d_i}^{[\mu]}$is the degree of node *i* on layer$\mu$and M is the number of layers. In order to rank nodes, a single *overlapping degree* o_i can be obtained for node *i* by summing its degree on all layers. However, the overlapping degree would not distinguish a node which is a hub on a certain layer but has few links on the others (i.e., an airport which is a hub only for a particular airline) from a node that has similar degrees on all layers. To distinguish these situations, the *multiplex participation coefficient* is introduced:

$$p_i = \frac{M}{M-1}[1 - \sum_{Q_{ji}} (k_i^{[\mu]}/o_i)^2].$$

This coefficient is maximum if the degree of *i* is the same for all layers, and it is 0 if the node has link on only one of the layers.

In order to define distance on a multiplex network, it should be decided whether inter-layer links count and, if they do, how much. In the case of the network of airports and flights, for example, a path using two flights of the same airline or alliance could be considered shorter than a path using two flights of two airlines not belonging to the same alliance. Therefore, the length of a path made of *n* links could be computed as $l = n + \beta n_{inter}$, where $\beta > 0$and $n_{inter}$is the number of inter-layer links used in the path. Once it has been decided how to measure distances, all metrics related to distances and shortest paths can be applied (network diameter, characteristic path length, efficiency, betweenness centrality, closeness centrality). Additionally, to measure to what extent the interconnection of different layers affects the connectivity of the network, the notion of *node interdependence* is introduced:

$$\chi_i = \sum_{j \neq i} \frac{\Psi_{ij}}{\sigma_{ij}},$$

where $\sigma_{ij}$is the number of shortest paths between *i* and *j*, and $\Psi_{ij}$is the number of those shortest paths which use inter-layer connections. The network level indicator is obtained averaging χ on all nodes.

In [11] several ways to compute centrality metrics in multiplexes are also suggested, focusing on eigenvector centrality (although the same considerations apply to Katz or PageRank centrality). The

first possible generalization is to compute centralities separately for each of the M layers, so that for each airport we have a vector of centralities $c_i = (c_i{}^{[1]}, \dots c_i{}^{[M]})$. Then, a scalar centrality measure can be defined as the sum of the $c_i{}^{[\mu]}$ or as their maximum, in order to rank the airports. This generalization considers only intra-layer connections to determine the centralities. Another possible generalization is that of computing the centralities on the *superposition network*, i.e., the network obtained by aggregating all the links present on every layer on a single one. This is equivalent to using the matrix $A_{sup} = \sum_{\mu=1}^{M} A^{[\mu]}$ to compute the centralities. In this case, no difference is made between inter- and intra-layer walks in computing the centralities. These two choices represent two particular cases of the generalization that we propose in the following (however, our generalization adds the temporal structure as well).

Finally, it could be interesting to know if nodes tend to have the same role in all layers or not, e.g., if an airport which is a hub on one layer tends to be a hub on the others as well. This can be assessed, for example, by computing the Pearson correlation coefficient of the airports' degrees on the different layers, or by computing the Kendall or Spearman correlation coefficient between the rankings obtained according to the degree (or any other centrality measure) in the different layers.

Temporal generalization of Katz centrality and PageRank

Katz centrality sums the contribution of walks outgoing from or incoming to a node, weighting them according to their length. In the case of the static network considered in Section 2.2.1.1, where the weighted adjacency matrix A contains all links present at any time, the number of paths of length *n* outgoing from node *i* are given by $\sum_j (A^n)_{ij}$, where the index *j* runs on all $N_A$nodes of the network, and contribute $\alpha^n \sum_j (A^n)_{ij}$ to the outgoing centrality of node *i*. Summing then the contribution of walk of any length, the outgoing centrality of node *i* is obtained as

$$k_i{}^{out} = \sum_n \alpha^n \sum_j (A^n)_{ij} = \sum_j (\sum_n \alpha^n A^n)_{ij} = \sum_j ((\mathbb{I} - \alpha A)^{-1})_{ij}.$$

Calling $\vec{k}^{out}$ the vector of outgoing centralities,$\vec{k}^{out} = (\mathbb{I} - \alpha A)^{-1}\mathbb{1}$- , where $\mathbb{1}$ is a vector of all ones. A similar derivation holds for the incoming case.

Here, we want to introduce a generalization of this centrality such that (i) passing from one node to another though an active link takes a non-zero time and (ii) only time-oriented walks are counted. For example, in the case of the airports and flights network, an itinerary $i \to j \to k$ is only counted if the flight from *j* to *k* departs after the landing of the flight from *i* to *j*. To do this, first we discretise the observation window in time frames of length Δt. Then, to overcome the limit of the temporal generalization introduced in [10], explained in the previous subsection, we introduce a set of secondary nodes, one for each of the N_F links (flights of the day) (see Figure 3). For each time frame $[t, t + \Delta t)$ we define an adjacency matrix $A^{[t]}$of size $(N_A + N_F) \times (N_A + N_F)$ such that $A_{ij}{}^{[t]} = 1$ either if a link outgoing from node *i* starts to exist during that time frame, and *j* is the secondary node associated to that link, or if a link incoming to node *j* stops to exist during that time step, and *i* is the secondary node associated to that link. In other words, in our specific example, for each flight a directed link between the origin airport and its secondary node is present in the time frame in which the flight's departure time falls, while a directed link between its secondary node and the destination airport is present in the time frame in which the arrival time falls. The introduction of these secondary nodes is necessary to correctly account for the time-ordering of walks. In fact, a walk joining airports *i, j* and *k* should not exist if the flight from *i* to *j* lands later than the departure of the one from *j* to *k*. This is ensured by introducing secondary nodes and by adding the rule that a walk

should not make more than one jump in each time frame. Note that the idea of secondary nodes to describe scheduled temporal networks had already been introduced previously [14] [15], but never applied to the computation of centrality metrics.

The length Δt of one time frame must be shorter than the duration of the shortest link. For example, for a flight of length larger than Δt, the departing and landing links will appear in two different time frames and can be both used in a walk, while for a flight of length smaller or equal to Δt both links appear in the same time frame, therefore a walk will remain stuck in the secondary node. With this definition of the time series of adjacency matrices, the contribution of time-ordered walks of length *n* outgoing from node *i* to the centrality of node *i* is obtained as the sum of all the contributions of the form

$$\alpha^n \sum_j (A^{[t_1]} A^{[t_2]} \dots A^{[t_n]})_{ij}$$

where $t_1, \dots, t_n$ are successive time frames, $t_1 < t_2 < \dots < t_n$ (with no repetition, so that a walk can use, at most, one link per time-frame), for all possible choices of $t_1, \dots, t_n$. The sum of all contributions of this form, for *n* that goes from 1 to infinity, is obtained as the product

$$\sum_j [(\mathbb{I} + \alpha A^{[1]})(\mathbb{I} + \alpha A^{[2]}) \dots (\mathbb{I} + \alpha A^{[T]}) - \mathbb{I}]_{ij}$$

Then, the vector of temporally generalized outgoing Katz centralities, termed Single-Layer Trip outgoing centrality, is

$$\vec{t}_s{}^{out} = [(\mathbb{I} + \alpha A^{[1]})(\mathbb{I} + \alpha A^{[2]}) \dots (\mathbb{I} + \alpha A^{[T]}) - \mathbb{I}]\mathbb{1}.$$

With a similar derivation, for the incoming centralities we have

$$\vec{t}_s{}^{in} = \mathbb{1}^T [(\mathbb{I} + \alpha A^{[1]})(\mathbb{I} + \alpha A^{[2]}) \dots (\mathbb{I} + \alpha A^{[T]}) - \mathbb{I}].$$

The generalization of PageRank centrality, here termed Single-Layer TripRank, is obtained analogously as

$$\overrightarrow{tr}_s{}^{out} = [(\mathbb{I} + \alpha A^{[1]} {D_{in}}^{-1})(\mathbb{I} + \alpha A^{[2]} {D_{in}}^{-1}) \dots (\mathbb{I} + \alpha A^{[T]} {D_{in}}^{-1}) - \mathbb{I}]]\mathbb{1}$$

and

$$\overrightarrow{tr}_s{}^{in} = \mathbb{1}^T [(\mathbb{I} + \alpha {D_{out}}^{-1} A^{[1]})(\mathbb{I} + \alpha {D_{out}}^{-1} A^{[2]}) \dots (\mathbb{I} + \alpha {D_{out}}^{-1} A^{[T]}) - \mathbb{I}]$$

where D^{in} and D^{out} are the usual diagonal matrix such that ${D_{ij}}^{in} = \delta_{ij} {d_i}^{in}$ and ${D_{ij}}^{out} = \delta_{ij} {d_i}^{out}$, with δ_{ij} the Kronecker delta and d_i the degree of node *i*. Note that, on the temporal network, we define incoming (outgoing) degree of a node the number of incoming (outgoing) links that it has during the observation window. This corresponds to its strength in the static network.

Note that, differently from the static case, there is no upper bound on the parameter, as the sum is always bounded. In fact, there are no walks longer than the number of time-frames.

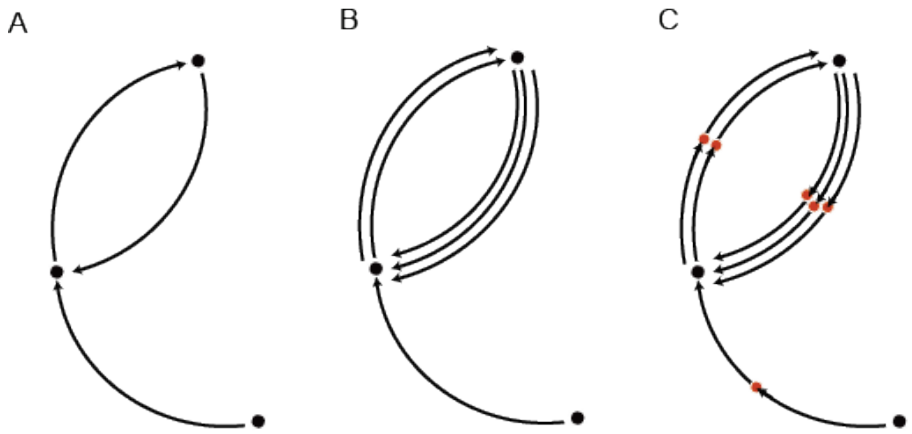

Figure 3. Illustration of the introduction of secondary links needed for the temporal generalization of Katz and PageRank centrality in the case of airports and flights. A) Network where two airports (black dots) are joined by a direct link if there is at least one flight between them; B) Network where each flight is a link, therefore multiple link are present between each airports; C) One secondary link (red dot) is introduced for each flight, so that each flight is represented by two links, one from the origin airport to its secondary link, and one from the secondary link to the destination airport.

Multiplex generalization of Katz centrality and PageRank

Let us now add the multiplex structure to the temporal network. We consider one layer for each type of link existing in the network. For example, in the case of the airports and flights network, each layer corresponds to an alliance or a single airline not part of any alliance. The scope of this further generalization is to distinguish the walks made of links of the same type from those made of links of different types, which might give a different contribution to the centrality (length being equal). For example, flights of the same airline or alliance, which are preferentially used by passengers, should give a larger contribution to centrality with respect to the ones using flights of different airlines which do not belong to the same alliance. As the multiplex has a copy of each node on each layer, the adjacency matrix A is of size $(N_A N_L + N_F) \times (N_A N_L + N_F)$, where N_L is the number of layers. Now, for each link of type λ a directed link is present, in the time frame corresponding to appearance (departure), between the copy of the origin node on layer λ and its secondary node. A directed link between its secondary node and the copy of the destination node on layer λ is present in the time frame corresponding to its disappearance (landing).

Let us introduce the parameter $\varepsilon \leq 1$, such that the contribution to Katz centrality of a walk of length n using m intra-layer links is $\varepsilon^m \alpha^n$. The effect of this parameter is that, the more changes of layer a walk has, the less is it contribution to centrality. Let us then introduce the matrix K, of the same size of A, as the matrix with elements $K_{ii} = 1$ and $K_{ij} = \varepsilon$ if i and j are two copies of the same nodes on different layers. Now, the products of the form $A^{[t_1]} K A^{[t_2]} K A^{[t_3]} \ldots$ count walks by introducing a factor ε every time that there is a change of layer. Therefore, the outgoing and incoming Trip centrality on the temporal multiplex are written as

$$\vec{t}^{out} = [(\mathbb{I} + \alpha A^{[1]} K)(\mathbb{I} + \alpha A^{[2]} K) \ldots (\mathbb{I} + \alpha A^{[T]} K) - \mathbb{I}] K^{-1} \mathbb{1}$$

$$\vec{t}^{in} = \mathbb{1}^T [(\mathbb{I} + \alpha A^{[1]} K)(\mathbb{I} + \alpha A^{[2]} K) \ldots (\mathbb{I} + \alpha A^{[T]} K) - \mathbb{I}] K^{-1}$$

TripRank centrality can be generalized analogously as

$$\overrightarrow{tr}^{out} = [(\mathbb{I} + \alpha A^{[1]} {D_{in}}^{-1} K)(\mathbb{I} + \alpha A^{[2]} {D_{in}}^{-1} K) \dots (\mathbb{I} + \alpha A^{[T]} {D_{in}}^{-1} K) - \mathbb{I}] K^{-1} \mathbb{1}$$

and

$$\overrightarrow{tr}^{in} = \mathbb{1}^T [(\mathbb{I} + \alpha {D_{out}}^{-1} A^{[1]} K)(\mathbb{I} + \alpha {D_{out}}^{-1} A^{[2]} K) \dots (\mathbb{I} + \alpha {D_{out}}^{-1} A^{[T]} K) - \mathbb{I}] K^{-1}$$

where D^{in} and D^{out} are the diagonal matrix such that ${D_{ij}}^{in} = \delta_{ij} {d_i}^{in}$ and ${D_{ij}}^{out} = \delta_{ij} {d_i}^{out}$, with δ_{ij} the Kronecker delta and d_i the degree of node i. Note that now the degree of a node is given only by the links incoming to or outgoing from that node on that layer, and not by the one belonging to its copies on other layers. Note that the matrix K is invertible for all $\varepsilon \neq 1$.

With this procedure, we obtain a centrality measure for each copy of each node, summing the contributions of walks outgoing from (or incoming to) that node with a specific type of link. In our example, such specific information is interesting if we want to measure the importance of an airport for a particular airline or alliance. However, if we are interested in the role of the airport for the entire network, an aggregated measure is needed. The aggregated centrality of a node is obtained by summing the centralities of all its copies. In fact, this will give the contribution of all walks outgoing from (or incoming to) any copy of that node.

Additionally, the centrality of a secondary node can be thought of as the centrality of the link corresponding to that secondary node. For example, the outgoing centrality of a flight could be a tool for airlines to decide which flights to prioritize. This centrality can be easily computed from the above formulas, however in this deliverable we will not show the empirical results, which might be investigated in future deliverables.

The parameter ε determines how much inter-layer walks are penalized. In particular, the case $\varepsilon = 0$ corresponds to not allowing inter-layer walks. In this case, the layer-specific centralities are those that consider only the links of one type. The aggregated centralities consider links of all layers, but only intra-layer walks contribute. When instead $= 1$, inter-layer walk gives the same contribution of intra-layer walks. For $0 < \varepsilon < 1$, inter-layer walks are considered, but contribute less than an intra-layer walk of the same length.

Note that, when ε=1, the matrix K is not invertible. In this case, it is not possible to multiply by K^{-1}in the computation of centralities. However, neglecting this multiplication only changes the aggregated centralities by a multiplicative factor, therefore it does not influence the ranking of nodes according to the aggregated centralities.

In our example, given that passengers tend to follow itineraries composed of flights of the same airline or alliance, the case ε=0 is probably the most realistic in evaluating the network functioning from the passengers' point of view. However, in the empirical part we will compare also results obtained with $\varepsilon > 0$, to show how the network functioning changes when inter-layer walks are permitted.

In conclusion, we obtained two original centrality metrics generalizing Katz centrality and PageRank centrality to the case where links on the network are dynamic and have a non-zero duration and the network has a multiplex structure. These new metrics are tailored for the ATM case or for other

transportation networks, as they focus on counting only walks that can actually be traveled by passengers.

Comparing the scheduled and the actual network

In many situations, especially in transportation, there exists a scheduled and an actual network. The Trip centrality that we introduced is able to quantify the loss of connectivity of a node (or of a link) due to the difference in the temporal link structure of the two networks.

Centrality metrics can be applied both to the scheduled network and to the actual ('real') one, by changing the adjacency matrices from $\{A^{[t]}{}_{sched}\}_{t=1,\dots,T}$, which are calculated according to the former, to $\{A^{[t]}{}_{real}\}_{t=1,\dots,T}$, which are calculated according to the latter. In the actual network, in fact, the timing of links is changed due to delays, therefore some of the time-oriented walks that existed in the scheduled network will not be present anymore, causing losses of centralities and changes in the ranking of nodes with respect to the scheduled network.

Such losses of centrality or rank, in absolute or percentage terms, can be used to assess to what extent the network, or a particular node, link or layer is affected by delays in a certain scenario.

Note that, in principle, delays could also create new walks, by allowing connections that were impossible in the scheduled network. However, in our specific example, these walks would only be used by passengers in cases of rerouting (e.g., after a missed connection). Therefore, they currently give no contribution to an airport's connectivity and should be excluded from the computation of centrality[3]. The new walks can be opened in two cases. The first case is when a negative arrival delay (early arrival) allows a connection with a departing flight. In this case, the problem is solved by setting the negative arrival delay to zero, i.e., setting the arrival time equal to the scheduled arrival time. The second case is when the delay of a departing flight allows the connection with an incoming flight, which was originally landing too late to make the connection. A partial solution to this can be implemented, and is explained in Annex 2.

2-legs Trip centrality

The parameter α determines the weight of a walk made of *n* jumps (i.e., *n/2* jumps between primary nodes, such as airports; recall that a single flight comprises 2 links, as per Figure 3). By choosing a sufficiently small α, we can give less importance to long walks relatively to short ones. Specifically, in our example, if we express the parameter α in an exponential form, in terms of a new parameter L, as $\alpha = e^{-1/(2L)}$, we can see that the weight of a walk made of $l < L$ flights is very small with respect to walks made of one flight. However, longer walks are still considered and have a role in determining the centralities. As passengers mainly travel one- and two-legs itineraries, it is interesting to compare the results with a centrality measure that only considers walks of one or two legs, which we will call 2-legs Trip centrality. As these walks are a small number (the number of flights plus the number of connections), the computation of centrality can be done simply by summing the contribution of each walk, without the need of a matrix formulation. The outgoing and incoming centralities of airport *i* are obtained respectively as

[3] In other applications, however, it could be important to consider these new walks, e.g., in the bus or metro network.

$$t_{[2],i}{}^{out} = \alpha d_i{}^{out} + \alpha^2 c_i{}^{out}$$

and

$$t_{[2],i}{}^{in} = \alpha d_i{}^{in} + \alpha^2 c_i{}^{in}$$

where $d_i{}^{out}$ and $d_i{}^{in}$ are the number of outgoing and incoming flights and $c_i{}^{out}$ and $c_i{}^{in}$ are the number of incoming an outgoing two-legs itineraries. It is possible to count, or not, two-legs itineraries made by flights belonging to different airlines or alliances. The weights can be adapted according to the nodes degree to mimic PageRank. An additional advantage of this two-legs centrality is that we can choose to consider only two-legs itineraries with a connecting time which does not exceed a threshold. In the following, the results of this two-legs centrality will be compared with the results of Trip centrality.

2.2.2 Causality metrics

Causality is what connects one process, i.e., the cause, with another process or state, i.e., the effect, where the first is partially responsible for the second, and the second is partially dependent on the first. In the ATM case, the cause and effect processes could be, for example, the state of delay of two different airports, where the state of delay quantifies the amount of delays at that particular airport. A causality relationship between these two processes could arise, for instance, when a flight departing with a delay from the first airport arrives at destination with a primary delay which induce delays in other flights because of rotational effects. Thus, the state of delay at the destination airport partially depends on that of the origin airport. This represents a process of delay propagation between two airports mediated by a one-leg effect, i.e., one flight connecting the two airports. However, a causality relationship might be mediated by more than one leg.

In scientific investigations of causality, cause and effect can be conceived as time processes whose realisations represent their state at a given time. In the example given above, the realization at time *t* of the state of delay of an airport reflects the delays in that airport at time *t*, as will be better specified in the following. In general, in a complex system which is formed by several subsystems interacting with each other, we can quantify at any time the state of subparts of the system as a time series and causality metrics allow to infer the causality structure among these subparts, starting from the observations of their states. The main idea behind the causality metrics introduced below lies in assessing whether the information about an element, the 'causer', is statistically useful in forecasting the state of another element.

The ATM system includes several subsystems, ranging from flights to the airline companies, the Network Manager, the Departure and Arrival Managers, airports and, last but not least, passengers. All subsystems interact with each other during both the strategic and the tactical phases in order to plan the operations and work to fulfil the program, respectively. Causality inference can be used to identify interactions among subsystems monitored with time series. In the following, as a case study, we consider the ATM system aggregated at the airport level, i.e., the network of airports and flights. Nevertheless, future applications will refer to the interactions of other subsystems, in line with Domino's scope, and the causality metrics presented here can be applied broadly to the outputs of both the toy model and the full ABM model, where several subsystems are modelled. Signals, such as delays and costs, propagate through the ATM network due to the stochastic interactions of the

elements. Detecting the channels of propagation and quantifying the tightness of interactions is fundamental to assess the system performance and, in the light of Domino's scope, the impact of innovations, i.e., whether or not the novel mechanisms make the ATM system more interdependent.

The detected causality relationships between a large set of elements form a second network, named causality network, where directed links are the causal relationships themselves. In view of Domino's goals, the study of causality networks, whose topology may change depending on implemented scenarios, allows to investigate the impact of innovations at the micro level on the delay and cost dynamics and propagation. For example, the detection of a smaller number of causal links and causal feedbacks loops can be seen as an improvement of the system, as it signals a diminished coupling of the systems' elements and therefore a reduced risk of systemic spreading events. For this purpose, in the next deliverables, the causality metrics presented here will be applied on both the outputs of the agent-based model and the toy model under different innovation scenarios.

A method to detect causality in time series analysis was introduced for the first time by Granger in his pioneering paper [16]. His original approach has then been generalized and applied in diverse fields. In finance, for example, causality metrics have been successfully applied to define instability indicators which can identify periods of turbulence in the market and determine what the channels of propagation of systemic risk are. This has been done, for example, by considering measures of connectedness for the causality network built by means of the Granger causality tests [17], [18]. Recently, the standard Granger causality test has been applied to investigate delay propagation in the Chinese ATM system [19] in order to identify which airports have a more important role in the process. The Granger causality measure has also been applied to the European ATM system with a particular focus on the passenger perspective, by studying the propagation of both flight and passenger delay in a network simulation model [20], [21]. More recently, several studies have been devoted to the detection of causality for extreme events and Hong et al. proposed to use an extension of the Granger causality test, namely: 'Granger causality in tail' [22]. With the same aim, other measures of causality for extreme events have been introduced [23] and considered for a first application in the ATM system [24], opening novel research questions regarding the propagation of extreme delays.

2.2.2.1 Causality metrics: methods

In the following, we review the existing methods to assess causality between two elements of a system, which are represented by two time processes described by two discrete time series. As a specific example, the two elements could be two airports described by the time series of their state of delays.

First, we review the most commonly adopted causality measure, i.e., Granger Causality (GC) in mean test. One of the limits of the Granger causality in mean metrics lies in the assumption of linearity for the dynamics of time processes. Hence, some complex behaviours might not be fully captured by linear models. For example, in the case of airports and flights, departing delays which are small with respect to flight time are probably not relevant for delay propagation, as they are easily absorbed in the en-route phase or by buffers. These small delays are, nevertheless, considered by the Granger causality test. For this reason, we propose to use the extension of [22] for a novel application to the ATM system, which we review.

Granger causality in mean

The original method proposed in [16] to test whether there is a causal relationship between two time series is based on the idea that, if the knowledge of past observations of one time series allows us to forecast future observations of the other time series better than without considering them, then there exists a directional causal relationship. We refer to this metric as Granger causality in mean.

Let X and Y be two time processes, which might represent, e.g., the state of delay of two airports, having realisations x_t and $y_t$at time t. The time process $Y = \{y_t\}_{t=1,\dots,T}$ is said to Granger-cause $X = \{x_t\}_{t=1,\dots,T}$ if we reject the null hypothesis that the past values of Y do not provide statistically significant information about future values of X by assuming VAR(p) as the predictive model [25], i.e.

$$x_t = {\phi_0}^1 + \sum_{j=1}^{p} {\phi_j}^{11} x_{t-j} + \sum_{j=1}^{p} {\phi_j}^{12} y_{t-j} + {\varepsilon_t}^1$$

$$y_t = {\phi_0}^2 + \sum_{j=1}^{p} {\phi_j}^{21} x_{t-j} + \sum_{j=1}^{p} {\phi_j}^{22} y_{t-j} + {\varepsilon_t}^2$$

where ${\varepsilon_t}^1, {\varepsilon_t}^2$ are taken to be two uncorrelated white-noise series. The goal of the Granger test is to assess the statistical significance of $\{{\phi_j}^{12}\}_{j=1,\dots,p}$ by considering as null hypothesis that they are zero, i.e., ${H_0}^{mean}$: $\{{\phi_j}^{12} = 0\}_{j=1,\dots,p}$. The null hypothesis ${H_0}^{mean}$ is equivalent to considering that x_t evolves according to a AR(p) process, i.e.,

$x_t = {\varphi_0}^1 + \sum_{j=1}^{p} {\varphi_j}^{11} x_{t-j} + \varepsilon_t$.

After estimating both VAR(p) and AR(p) models, an F-test is applied in order to test if VAR(p) outperforms statistically AR(p) in fitting the observations $\{x_t\}$ [26]. If it does, ${H_0}^{mean}$is rejected, meaning that Y 'Granger-causes (in mean)' X.

The parameter *p,* the order of the autoregressive process, represents the number of past observations considered for the forecasting. This can be interpreted as the system's 'memory'. In order to identify the optimal value of *p* for a specific problem, the Granger test is performed for different values of *p* and the optimal one is then selected according to the Bayesian information criterion (BIC)[4] [25].

In the case of airports and flights, a directional causal relationship between two airports exists when the state of delay of an airport X has a significant statistical dependence on the past states of delay of another airport Y. The optimal value of *p* represents the time scale of the detected causal relationships. In principle, this depends on the types of effects which mediate the propagation, e.g., one-leg or two-legs effects.

[4] The Bayesian Information Criterion (BIC) is a criterion for model selection among a finite set of models, but considering that it is possible to increase the likelihood of a model by adding parameters. In the case of bivariate VAR(p) models, we introduce four new parameters any time p is increased by one. BIC selects the model which provides the best fit of the data but penalizing for the number of parameters.

Finally, from a theoretical point of view, the statistical test weights equally the prediction performance on both large and small values of the time series. However, when events with an extremely large value have a particular importance, as e.g., in the case of delays, it would be preferable to consider an extension of the Granger causality test, namely Granger causality in tail [22], which only considers extreme events. This extension is reviewed in the following section.

Granger causality in tail

With the same spirit of the original test, 'Granger causality in tail' aims to evaluate whether extreme events of a time process cause (in the sense of Granger, i.e., help to predict) extreme events for another element of the system [22]. For example, the state of an airport is now described by a binary variable, the state of congestion, which is one if its state of delay is *extreme*, zero otherwise. The state of delay is *extreme* if it falls in the tail of the distribution. The causality test is then applied to the binary time series describing the states of 'distress' of airports[5].

In detail, the Granger causality in tail test works as follows. Assume to know at time t the probability density function of the state X conditional on past values $\{x_s\}_{s=1,\dots,t-1}$ and let us define $V_t = V(x_1,\dots,x_{t-1},\beta)$ as the $(1-\beta)$-quantile of the conditional probability distribution of X, i.e., $P(X > V_t|x_1,\dots,x_{t-1}) = 1-\beta$ almost surely with $\beta \in (0,1)$ defines V_t implicitly. The null hypothesis ${H_0}^{tail}$ of the test is that predicting an extreme event of X with or without the past information on Y is statistically equivalent, i.e. $P(X > V_t|\{x_s\}_{s=1,\dots,t-1}) = P(X > V_t|\{x_s\}_{s=1,\dots,t-1},\{y_s\}_{s=1,\dots,t-1})\ a.s.$

A rejection of the null hypothesis ${H_0}^{tail}$ means that Y 'Granger causes in tail' X at level β.

${H_0}^{tail}$ can be formulated in a similar fashion to ${H_0}^{mean}$, after a proper transformation of the continuous state variables $\{X,Y\}$ to the binary state variables $\{\hat{X},\hat{Y}\}$ by means of the indicator function, i.e.

$$\hat{x_t} = I_{x_t > V_t}, \hat{y}_t = I_{y_t > Q_t}$$

where V_t and Q_t are the $(1-\beta)$-quantile of the conditional probability distribution of X and Y, respectively. The indicator function takes value 1 when the observation x_t exceeds the quantile, i.e., it falls on the right tail, and takes value 0 otherwise. In the case of airports, $\{\hat{X},\hat{Y}\}$ describes thus the states of distress. Then, the null hypothesis ${H_0}^{tail}$ can be stated as

$${H_0}^{tail}: E(\hat{x}_t|\{x_s\}_{s=1\dots,t-1}) = E(\hat{x}_t|\{x_s\}_{s=1\dots,t-1},\{y_s\}_{s=1,\dots,t-1})\ a.s.$$

Thus, one would want to test Granger causality in tail between X and Y as Granger causality in mean between $\hat{X}$ and $\hat{Y}$. However, the standard regression-based test proposed by Granger can not be used here, because the quantiles have to be estimated, and the quantile estimation uncertainty has a nontrivial impact that should be taken care of properly. Furthemore, the regression analysis of the Granger in mean causality test cannot be applied straightforwardly to the case of binary random variables $\hat{X}$ and $\hat{Y}$. To overcome this problem, the use of the cross-spectrum of $\{\hat{X},\hat{Y}\}$ is suggested in [22]. Indeed, under ${H_0}^{tail}$, $\rho(j) = 0$ for all $j > 0$ where $\varrho(j) = corr(\hat{x}_t, \hat{y}_{t-j})$ is the lagged cross

[5]This definition of 'distress' aims to highlight the difference between normal operations at the airport and occurrence of 'extreme' delays, which is not necessarily related to congestion, according to the standard definition.

correlation. For further information on how to make testable this last version of the hypothesis ${H_0}^{tail}$ by means of spectral methods see [22].

Finally, there are many methods to estimate the quantiles from empirical observations of a time series, ranging from historical simulation methods, autoregressive conditional density model [27] to conditional autoregressive VaR (CAViaR) models [28]. Here, we adopt the autoregressive conditional density model in [27]. Specifically, we assume a AR(p) model for X with independent and identically distributed Gaussian innovations ε_t and, once the autoregressive process has been estimated on $\{x_s\}_{s=1,\dots,T}$, we project at each time t the Gaussian density conditional on past observations of X to obtain the time series of $\hat{X}$. When ${H_0}^{tail}$ is rejected, we say that Y 'Granger-causes in tail' X.

Causality networks and correction for multiple hypothesis testing

Having established how to detect a causal relationship between two elements of the system, we consider the network of causal relationships where a link $i \to j$ is present if i 'Granger causes' j. This approach has already been considered in ATM by Zanin et al. [19]. Given *N* time series, representing the states of the subsystems, a Granger causality test is performed on all the possible *M=N(N-1)* pairs and the network of direction causal relationships is obtained. Then, the Granger causality network is described by an adjacency matrix whose generic entry is one if there exists a directional causal relationship from node *i* to node *j*, zero otherwise. The number of links incident to a node determines its degree. The adjacency matrix is not symmetric because a causal relationship from i to j does not imply a causal relationship from j to i. Thus, a node's in- and out-degree might differ.

Since to build the causality network we perform *M=N(N-1)* tests we must take into account that, in multiple hypothesis testing, even if the null hypothesis is correct, a fraction α of tests will be rejected at a significance level α. For this reason, the statistical literature has proposed several methods to correct the significance level by taking into account the fact that multiple tests are performed. The choice of the correction is still matter of debate and depends on the fraction of false positives (i.e., test rejections when the null hypothesis is correct) one is willing to accept.

In this Deliverable we use a very restrictive correction, named Bonferroni correction [29] (see also [30] for the application to networks): when one tests simultaneously *M* hypotheses, in order to achieve a significance level α, a significance level of $\alpha' = \alpha/M$ should be apply to each test. This correction decreases dramatically the number of false positives, even if it might increase the number of false negatives, i.e., some test can be erroneously not rejected. In our context, this means that the observed causality relationships are very likely present, while there might be other undetected causality relations. For this reason, in the near future we plan to apply less restrictive multiple hypothesis test corrections, such as the False Discovery Rate.

3 Evaluation of metrics

We here apply the centrality and causality metrics described in the previous section to an important case study, that of the network of airports and flights. The empirical analysis is performed on two datasets containing, respectively, the scheduled and actual flights of the ECAC and the US airspaces. The ECAC dataset will be used in the calibration of the ABM.

In Section 3.1, we describe the two datasets. In Section 3.2 we present some baseline statistics related to delays and missed connections in the two datasets, with the aim of characterizing the considered days in terms of their delays, allowing us to better interpret the results that we will obtain from the applications of the network metrics. We underline that the aim of such simple analysis is not a comparison of the two ATM systems. In Section 3.3 we apply the existing and the new centrality metrics to the two datasets. In Section 3.4 we consider the causality metrics for the US dataset.[6]

3.1 Data

Two different datasets are used for the evaluation of existing metrics and the validation of the new proposed ones.

The first dataset contains all flights passing through ECAC airspace on 1 September 2017, therefore both within-ECAC flights and flights which only cross the ECAC air space for part of their route. For this day, IFPS and DDR data are available. The former contains the m0 flight plans, while the latter contains the m1, m2 and m3 flight plans. The m0 are all the submitted flight plans prior the last one submitted to the Network Manager. The m1 is the final submitted flight plan. If an ATFM regulation has been applied to a flight, then a trajectory shifted in time by the amount of delay assigned is stored in the m2 flight plan. Finally, m3 contains the flight plan which has actually been flown. If a flight plan is submitted for a flight which is subsequently cancelled, m3 does not exist. If only one flight plan was submitted to the Network Manager, it is contained in the m1 (and the m0 will be empty) [31].

Since scheduled departure and arrival times are not available, in this deliverable, scheduled departure is estimated as the earliest time between the earliest off block time of m0 flight plans and the actual off block time (from m3). (It might happen that the actual off block time is earlier than the earliest m0 off block time if the first flight plan submitted was already delayed with respect to the schedule). The scheduled arrival time is then obtained adding to the scheduled departure time the

[6] In this deliverable we are not applying the causality metrics to the ECAC airspace, as the currently available data cover a time window too short to assess causality relationships with enough statistical significance. We plan to perform such an analysis when the data will be available.

flight plan duration according to the flight plan (m0 or m1, if m0 is not present) plus an estimation of the taxi in times and a buffer of 10 minutes (see [5]). Figure 4 shows the different segments of a flight and the buffers defined strategically. This ensures that schedules times used in this deliverable are realistic instead of considering the landing time as the schedule inbound time.

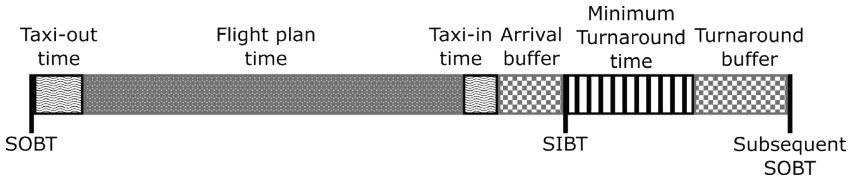

Figure 4. Schedules and buffers for a flight [5]

The list of available information for each flight in the dataset is shown in Table 3. All times are expressed in UTC. The dataset contains 41 656 flights, of which 29 939 are scheduled flights (72%) while the rest are non-scheduled flights (this includes charter flights but also military, helicopter, search and rescue, etc.). 28 221 of all the flights are within-ECAC flights, of which 83% are scheduled (23 489 flights). The remaining 13 435 flights are either overflying ECAC or flights with origin or destination outside ECAC. In the following, this dataset is termed “ECAC”. The intra-day pattern of the number of departing flights, in 5A), shows that most flights are operated between 5AM and 9PM UTC.

Additionally, we consider a second dataset, obtained from the US Department of Transportation's (DOT) Bureau of Transportation Statistics, containing flights operated in 2015 by 14 major US airlines. The reason to consider this additional dataset is two-fold. First, it covers an entire year, making it possible to compare network metrics applied to different days, with different delay conditions, and to perform the causality analysis, which needs a time-window of several days. For the ECAC dataset, at the moment, only one day was available. Secondly, ATM in the US differs from ATM in the ECAC making for an interesting comparison between the two. For example, in Europe the number of Level 2 and Level 3 slot coordinated airports is much higher than in the US, in Europe congestion is usually experience in the airspace while in the US is more common at airport level, etc. [32].

The information available for each flight in the US dataset is reported in Table 4. All times are converted from local time to Eastern Standard Time (EST). In the following, this dataset is termed “US”. The intra-day pattern of the number of departing flights, in Figure 5.B), shows that most flights are operated between 6AM and 12PM EST. The slightly longer time window of activity in the US with respect to Europe is due to the presence of time zones differing up to 5 hours.

Table 6 lists the 14 airlines present in the dataset with the corresponding number of flights in the period from January to March 2015, which give an idea of the size of each airline.

For the analyses presented in the following, different subsets of the ECAC and US datasets are used in different cases. All the used sub-datasets are described in Table 5. Additionally, for some centrality analyses, airlines are grouped by alliances. The alliance considered are Star Alliance, SkyTeam and Oneworld. Airlines that do not belong to any alliance are considered singularly in these analyses.

Table 3. Information available for each flight in the ECAC dataset

Field of dataset	What it contains
Date	Day, month, year of the scheduled departure time
Flight number	Unique flight id
Tail number	Registration number of the aircraft
Flight type	"S" for Scheduled Air Service, "N" for Non-scheduled Air Transport Operation, "G" for General Aviation, "M" for Military and "X" for everything else
Origin airport	ICAO code of origin airport from DDR
Destination airport	ICAO code of destination airport from DDR
Scheduled off block time	Estimated as explained in the text
Actual off block time	Actual off block time from m3
Scheduled in block time	Estimated as explained in the text
Actual in block time	Actual in block time from m3
Airline	ICAO code of airline operating the flight
Alliance	If the airline belongs to an alliance, it is specified
Cancellation	1 if the flight was cancelled, 0 otherwise

Table 4. Information available for each flight in the US dataset

Field of dataset	What it contains
Date	Day, month, year of the scheduled departure time
Flight number	Unique flight id
Tail number	Registration number of the aircraft
Origin airport	IATA code of origin airport from schedule
Destination airport	IATA code of destination airport from schedule
Scheduled off block time	Scheduled off block time
Actual off block time	Actual off block time from m3

Scheduled in block time	Scheduled off block time
Actual in block time	Actual in block time from m3
Airline	Scheduled off block time
Cancellation	1 if the flight was cancelled, 0 otherwise
Diverted	1 if the flight was diverted, 0 otherwise
Weather delay	Minutes of delay due to weather

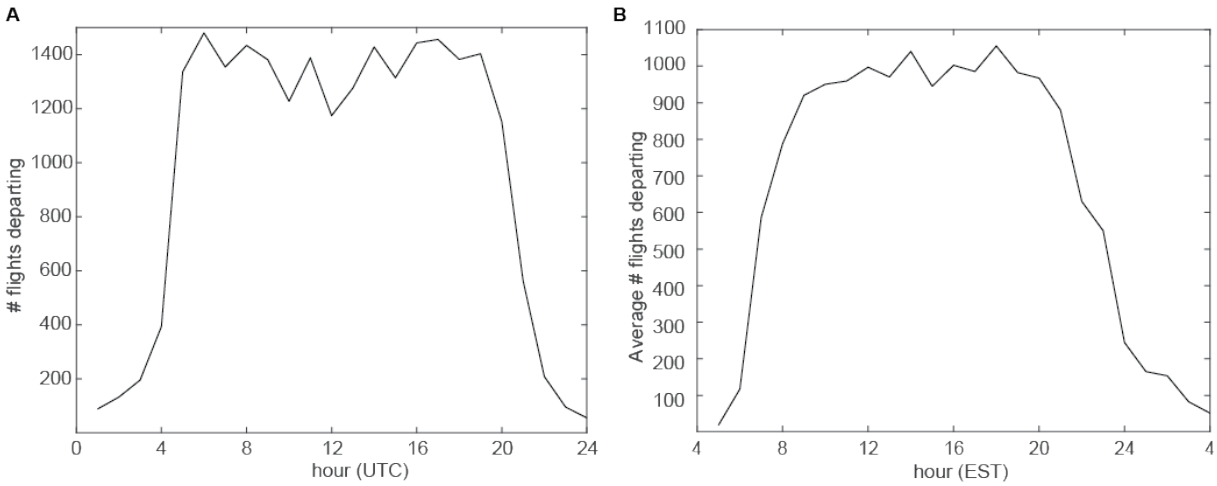

Figure 5. Intra-day pattern of departing flights. A) Number of flights departing per hour in the ECAC1 dataset; B) Average number of flights departing per hour in the US_April dataset.

Table 5. Description of sub-datasets of the ECAC and US dataset used in the analysis.

Dataset name	Description	Number flights	Number airports	Number airlines
ECAC1	Within-ECAC, scheduled, passenger	23 365	523	183
ECAC2	Subset of ECAC1, obtained by keeping only airlines which number of flights and number of destinations are above the average, or which are part of an alliance[7]	19 648	435	55
US_April	Subset of US dataset containing only the month of April (30 days) [8]	16 042 (per day on average)	322	14
US_JM	Subset of US dataset containing the months of January, February and March	1 403 471 (in total) 15 594 (per day on average)	315	14

Table 6. List of airlines in the US dataset, with their IATA code and the number of flights operated in the period from January to March 2015.

Airline	IATA code	Number flights
Southwest Airlines	WN	299 459
Delta Air Lines	DL	199 471
ExpressJet Airlines	EV	149 253
SkyWest Airlines	OO	142 181
American Airlines	AA	129 860
United Airlines	UA	118 233
US Airways	US	98 158
Envoy Air	MQ	84 986
JetBlue Airways	B6	63 964

[7] The list of airlines in the ECAC2 dataset is in Annex 1.

[8] Note that days are considered to start at 4AM EST, as this is the time of minimum traffic across the entire US.

Alaska Airlines	AS	39 727
Spirit Airlines	NK	26 232
Frontiers Airlines	F9	19 588
Hawaiian Airlines	HA	18 532
Virgin America	VX	13 827

3.2 Data description analysis

In this Section we present a simple analysis of the two datasets, considering delay statistics, potential reductions of connections and simple delay metrics scaled up to the network level.

3.2.1 Flight-centric delay statistics

Here, we present some simple delay statistics computed on the European and North American datasets described in Section 3.1. The computations have been done on the ECAC1, US_April and US_JM datasets. The aim of this analysis is, first, to show an example of how, when analysing delay, not only average values are relevant to characterize different ATM networks, but also their distribution and how they are generated. Secondly, these baseline statistics will provide us with a general idea of the delay situation of the datasets which we will use in the following analyses.

Figure 6 shows some statistics for the month of April in the US dataset, namely the histograms of the fraction of cancelled flights, the average delay, the average delay of delayed flights and the fraction of delayed flights. Such histograms show that the considered month includes a heterogeneity of ATM situations, ranging from days with many delays to days with few delays.

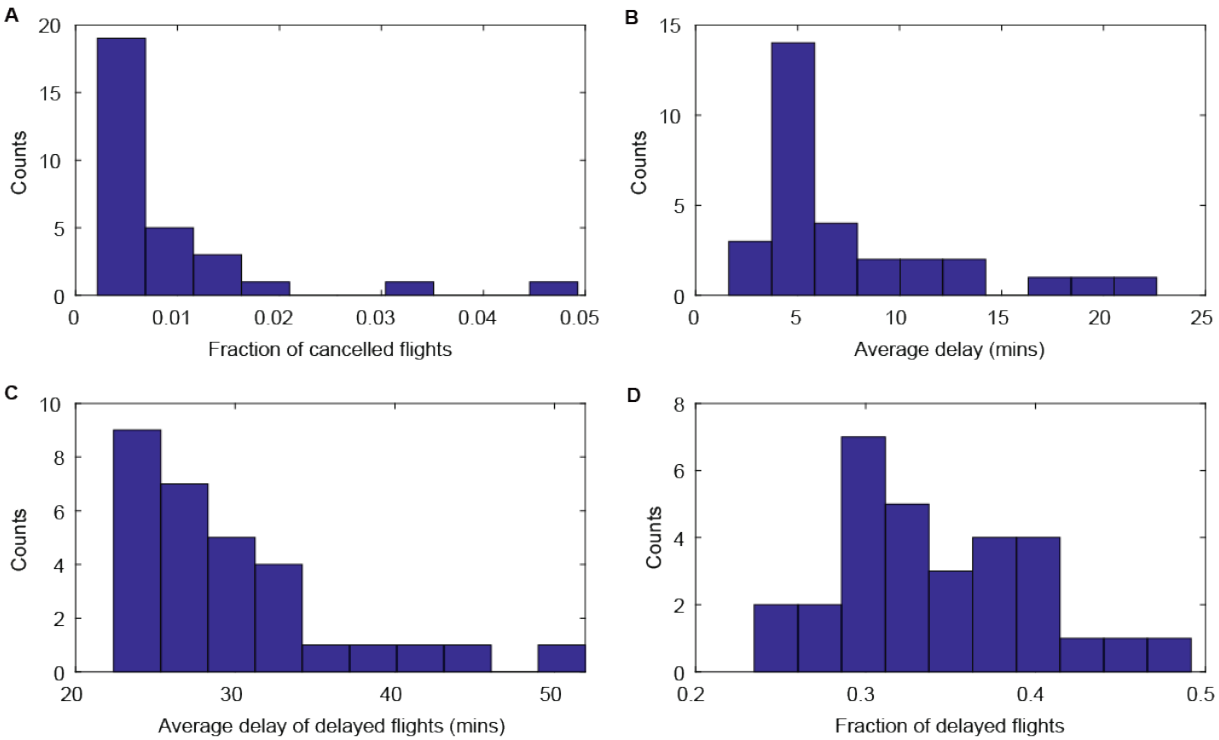

Figure 6. Delay and cancellation statistics for the US_April dataset; "Counts" = days. A) Histogram of the fraction of cancelled flights; B) histogram of the average departure delay; C) histogram of the average departure delay of delayed flights; D) histogram of the fraction of delayed flights.

Table 7 reports some statistics related to delays and cancellations. For the US_April dataset, statistics are averaged over all the days in the considered period. For each flight, the departure delay is obtained as the difference between the departure time and the scheduled departure, the arrival delay as the difference between the arrival time and the scheduled arrival, and the gate-to-gate delay as the difference between the arrival delay and the departure delay. The statistics have meaningful differences between the two cases.

In the European dataset, departure delay affects most flights, however less than half have an arrival delay at the gate with respect to their scheduled arrival time. The average arrival delay is smaller than the average departure delay. This is due to the fact that delays are absorbed thanks to the buffers. Also, note that we have defined the scheduled departure time as the earliest scheduled departure time from all the m0 and m1 flight plans, which might have an impact on the amount of flight that are counted as delayed at off-block time.

In the US data, the percentage of flights with departure delay and arrival delay is similar, and so are also the average departure and arrival delays. Additionally, the percentage of flights having extra delay between departure and arrival is higher than in the European dataset.

Departure delays, therefore, tend not to be recovered in the US, contrary to Europe. However, in terms of average arrival delays the results are similar. As mentioned before, we must note that the average gate-to-gate delay in the European dataset is affected by the choice of adding 10 minutes of buffer when estimating the scheduled inbound time, since real schedules were not available. However, independently of the choice of buffers, it remains that the US dataset has a much smaller

percentage of departure delays, which is in line with the fact that, in US tactical management of delay during the airborne phase is more common than in Europe. See [32] for more information on these.

Table 7. Statistics related to delays and cancellations in the two datasets.

	ECAC1	US_April
Percentage of cancelled flights	1.75 %	0.9%
Percentage of diverted flights	NA	0.3%
Percentage of flights with departure delay	71%	34%
Average departure delay	17 mins	8 mins
Average departure delay of delayed flights	24 mins	30 mins
Percentage of flights with arrival delay	43 %	35%
Average arrival delay	6 mins	3 mins
Average arrival delay of delayed flights	26 mins	29 mins
Percentage of flights with gate-to-gate delay	6 %	28%
Average gate-to-gate delay	-11 mins	-5 mins
Average gate-to-gate delay of delayed flights	6 mins	10 mins

The histograms of departure, arrival and gate-to-gate delays are reported in Figure 7. For the US dataset, histograms for all 30 days are shown. In the ECAC case, departure delays, in black, are always positive and peaked at zero. On the contrary, arrival delays (in red) and en-route delays (in blue) peak at a negative value. While the distribution of gate-to-gate delay is symmetrical, arrival delay has a right-skewed distribution. Note that in general departure delays with respect to the scheduled departure can also be negative, as they are in the US case. However, in the ECAC dataset, the estimation of the scheduled departure, which is not available, from the m1 makes all negative departure delay appear as zeros. Figure 8 shows the distribution of departure delays for a day of 2014, in Europe, for which schedules are available, and in this case the distribution is similar to the US one. Table 8 presents some descriptive statistics of the delay distributions.

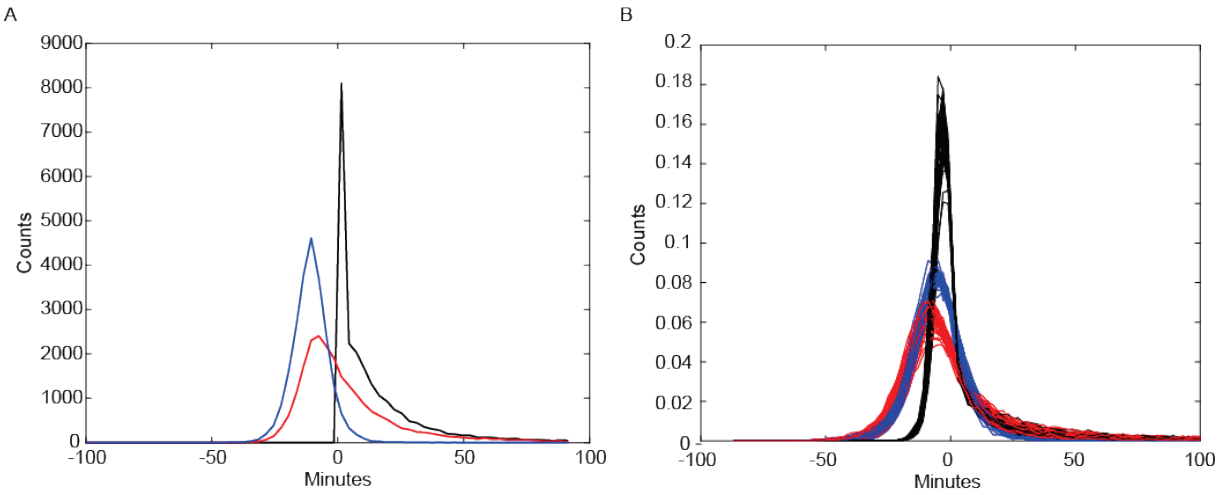

Figure 7. Histograms of delays for ECAC and US_April datasets. Departure (black), arrival (red) and en-route (blue) delays.ECAC1 dataset in panel A and US_April dataset in panel B. For the latter, histograms for each of the 30 days are shown.

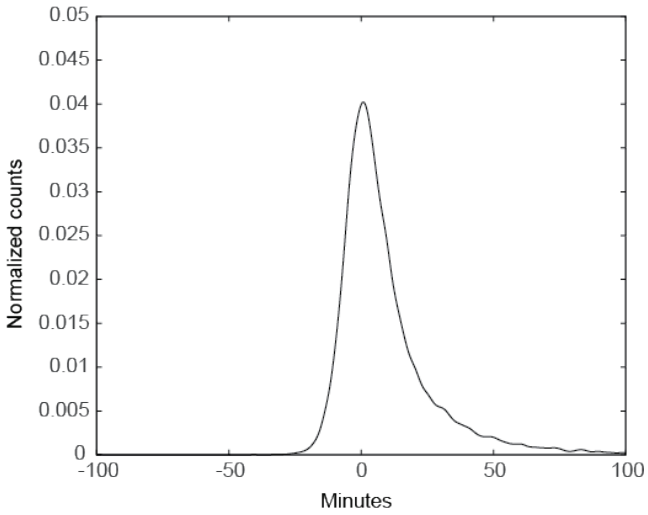

Figure 8. Histograms of departure delays for a day of 2014 in Europe.

Table 8. Descriptive statistics relative to the distributions delays: departure, arrival and en-route delays in the two datasets. For the US_April dataset, the reported values are obtained as the average of the corresponding statistics over the 30 days contained in the dataset, associated to its standard deviation.

	ECAC1			US_April		
	Departure delay	Arrival delay	Gate-to-gate delay	Departure delay	Arrival delay	Gate-to-gate delay
Mean	17	6	-11	8 ± 5	3 ± 6	-5 ± 2
Median	7	-2	-11	-2 ± 1	-5± 3	-5 ± 1
Mode	0	-8	-11	-4 ± 1	-9 ± 2	-6 ±1
Max	1 229	1 223	139	883 ± 248	884 ± 244	120 ± 34
Min	0	-116	-186	-31 ± 9	-63 ± 8	-63 ± 13
Stand. Dev.	31	32	8	33 ± 8	35 ± 8	13 ± 1

3.2.2 Reduction of potential passenger connections

In this Section an analysis of potential passenger connections at airports and how these get reduced when delay arises is performed.

As a first approximation, it is considered that there is a potential passenger connection between two flights f1 and f2, if the destination of f1 is the origin of f2 and the scheduled arrival time of f1 is earlier than the departure of f2. Note that we are assuming a minimum connecting time of zero as this computation is done just to show how the potential connections are affected by delay. Also, in order to reduce the number of connections to feasible selected ones by passengers, we consider that if the connecting time between two flights is too large, that itinerary will not be selected. I.e., the time between the scheduled arrival of f1 and the scheduled departure of f2 should be smaller than *n* hours, and for testing purposes *n* varies from 1 to 5 hours.

Again, simplifying for now, a potential connection is broken when the delay of flights produce that f1 arrives to the airport later than the departure time of f2.

Finally, two case studies are computed, one where connections are possible between flights of the same airline or alliance (within-airline potential connections) and one considering any two flights (any-airlines potential connections).

Table 9 reports the number and percentage of broken potential connections for n=5 for both datasets. Figure 9 shows how the percentage of the two types of broken potential connections changes with n. In both datasets, for maximum connecting times $n \leq 2$ hours the percentage of within-

airline broken potential connections is significantly lower (1 to 3 percentage points) that the percentage of any-airlines case. This might be due to two factors. First, part of the within-airline connections use the same aircraft, therefore their connection can never be disrupted, because if the incoming flight is late so will be the outgoing one. Secondly, this difference might depend on an additional effort of airlines to preserve within-airline connections, for example by organising a wave structure at a hub. This difference becomes small, however, for *n≥3*. This is probably due to the fact that delays as large as to cause the disruption of a connection with connecting time larger than 3 hours are difficult to recover, therefore for these connections the additional effort to preserve within-airline connections has no effect. Additionally, there are probably very few within-airline connections using the same aircraft which have a waiting time *n≥3*, therefore this factor pays a smaller role. Note that in the dataset that will be actually used in Domino, information on minimum connecting time will be used along with specific passenger itineraries allowing a more detail analysis of connections.

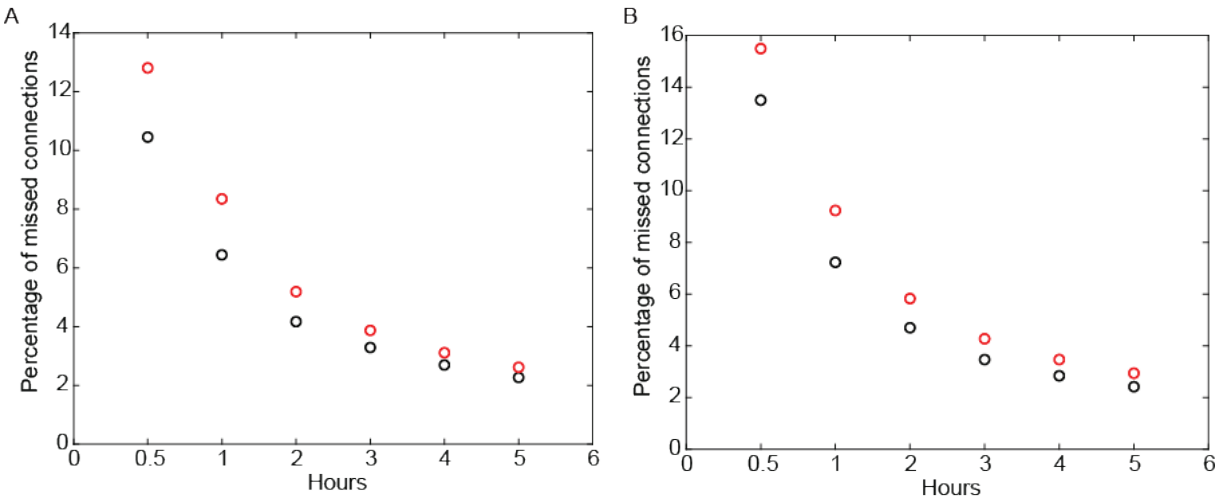

Figure 9. Percentage of broken potential connections in function of the connecting time between the two flights used to define a connection for the ECAC1 dataset (panel A) and the US_April dataset (panel B). Black circles represent connections between flights of the same airline or alliance, red dots represent connections between flights of any two airlines

Table 9. Statistics related to missed connections for the two datasets.

	ECAC1	US_April
Any-airlines missed connections (n=5)	29 442	38 861
Percentage of any-airlines connection missed (n=5)	2.7%	2.9%
Within-airline missed connections (n=5)	8 486	10 474
Percentage of within-airline connections missed (n=5)	2.6%	2.4%

3.2.3 Other delay statistic examples

Delays can be characterized geographically by identifying airports which were particularly affected by delays on the analysed day. Let us call an airport 'distressed' when the fraction of flights having a departure delay larger than the average delay of delayed flights for that day is greater than the mean fraction of delayed flights in an airport (averaged over all airports in the network). Note that the presence of many strongly delayed departing flights might be due to congestion in the arrival airports or in the airspace, and not necessarily in the departing airport. However, an airport in such a condition can be considered as distressed because of the large differences between the scheduled and the actual departing queue.

Figure 10 shows the position of the distressed airports for the day of the ECAC dataset. The map shows that delays are concentrated in central and southern Europe, particularly in Italy. Figure 11 displays the same for 4 days of the US dataset chosen among the ones with a particularly high number of distressed airports. The colour of the circles represents the fraction of delays due to the weather. The figure shows that on distressed days the geographical pattern of delays varies. Often distress is focused in a particular area (e.g., in the examples in the figure, East Coast, Chicago area or Dallas area), but in some cases it is more distributed.

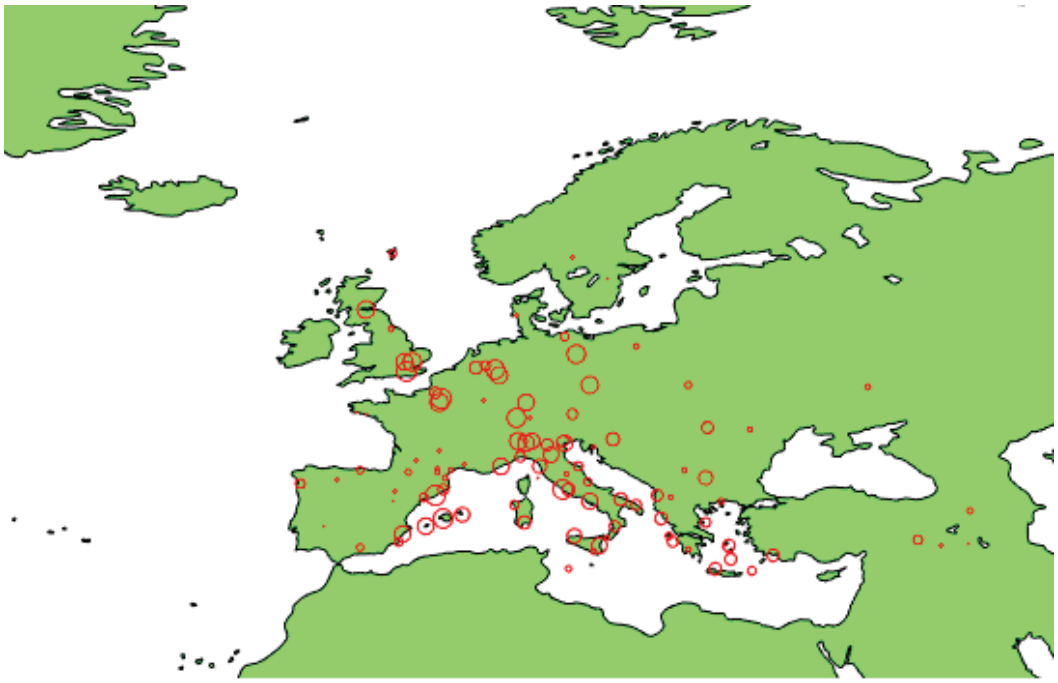

Figure 10. Geographical characterization of delays in the ECAC1 dataset. Red circles represent airports where more than 25% of flights have a departure delay larger than 24 minutes, which is the average delay of delayed flights in the dataset. The size of the circle is proportional to the logarithm of the total number of flights departing from the airport.

Geographical clustering of distressed airports might be due to a localized weather disturbance or to a localized over-occupation of the airspace, but might also be due to the tight interconnection of airports in the same geographical area, connected by many flights. For example, in panel C, the large

number of flights delayed due to weather in Dallas probably induced delays in the smaller neighbouring airports. In general, it might be interesting to know if distressed airports tend to be linked by flights, i.e., if they tend to cluster in the network of airports and flights. To answer this question, we replicate the analysis done in [33] on the period from January 1st 2015 to March 31st 2015 of the US_JM dataset. For each of the considered days, we determine the airports which are distressed according to the definition given above, and then we consider the network formed by these airports plus all the links between them on that day (i.e., a link is present if there is at least one flight connecting the two airports on that day). The results are shown in the top panels of Figure 12 for the days February 24th and 26th 2015. On a total of 315 airports, on day February 24th there are 43 distressed airports and none of these have been connected by a flight during the day, while on day February 26th we observe a smaller number of distressed airports, i.e., 11, but for about the 80% of these there was at least one flight connecting them. Note that the distressed airports are not necessarily the airports with the largest traffic.

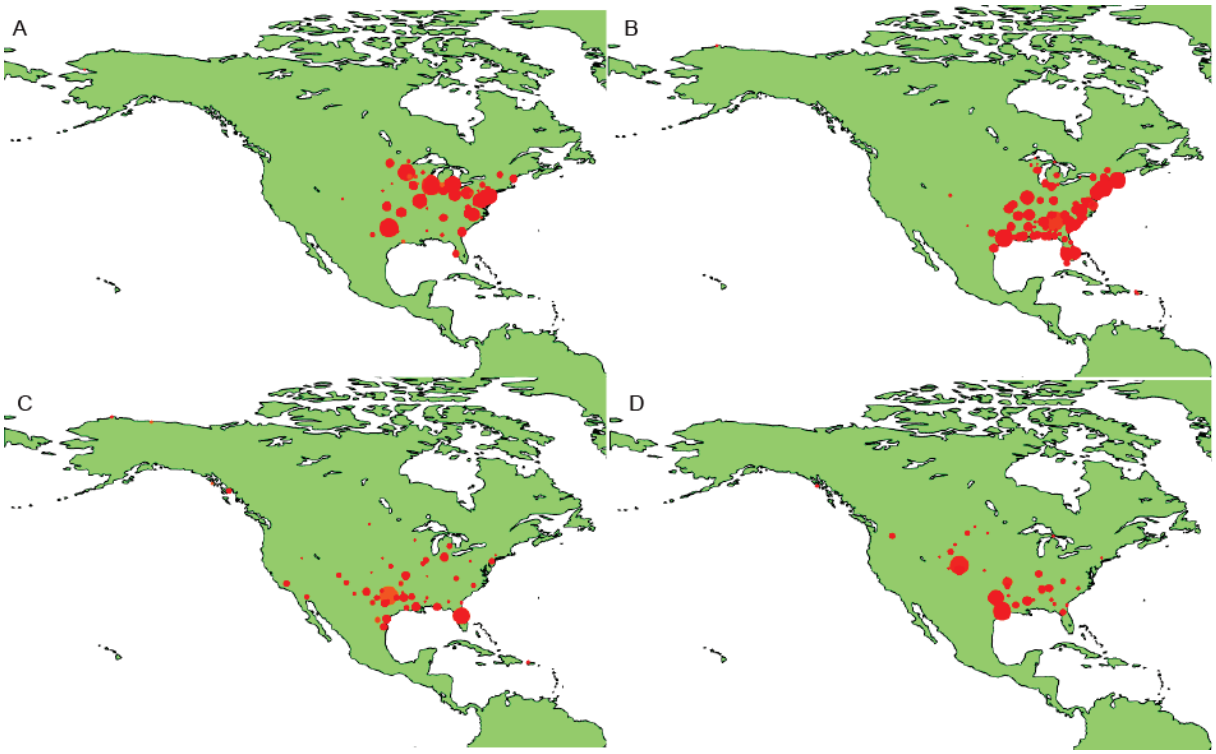

Figure 11. Geographical characterization of delay for four particularly distressed days in the US_April dataset. Circles represent airports where more than 25% of flights have a departure delay larger than 30 minutes, which is the average departure delay of delayed flights in the dataset. The circle color indicates the fraction of delays which are due to the weather (red= none, yellow= all). The circle size is proportional to the logarithm of the total number of flights departing from the airport.

The fact that, for instance on day February 26th 2015, distressed airports tend to be linked seems to suggest that flights are a channel of delay propagation. Nevertheless, there may exist other kinds of channels, e.g., two-legs effects which are described by paths of two flights with intermediary

airports, as the network of distressed airports on day February 24th 2015 suggests. An overall picture of this behaviour for all days can be captured looking at both the number of distressed airports, see the left bottom panel of Figure 12, and the relative size of the connected component, i.e., the fraction of connected nodes among distressed airports, see right bottom panel of Figure 12.

However, note that whether distressed airports tend to cluster or not, it is a sign of spatio-temporal correlation of distress but not necessarily of a cause-effect relationship. In fact, the process of delay propagation between two airports is not trivially related to the presence of a flight connecting them at any time, for two reasons. Firstly, for a flight to be a possible channel of delay propagation, its schedule must agree with the dynamics of delay propagation, i.e., the origin airport must be distressed before the flight departure and the destination airport must become distressed after the flight landing. Secondly, the presence of one flight might be a primary channel of delay propagation, but other mechanisms are possible, e.g., two-legs effects. We will delve further into these issues in Section 3.4, where we will apply the methods proposed in Section 2.2.1.2 to identify the channels of delay propagation taking into account these remarks.

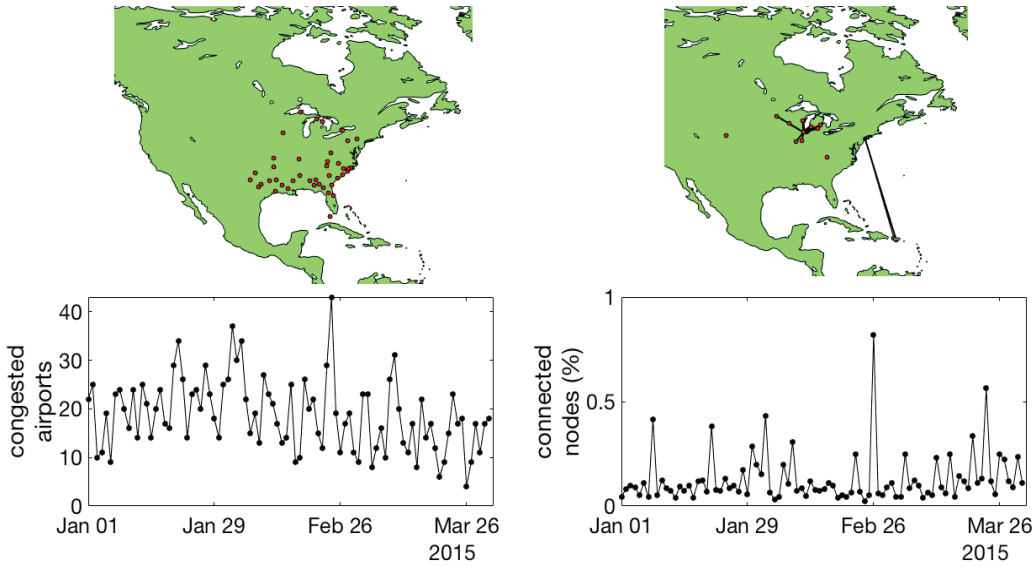

Figure 12. Network of distressed airports and daily number of distressed airports. Top panels: Networks of distressed airports for the days February 24th 2015 (left top panel) and day 26th February 2015 (right top panel). A link is present between two distressed airports if at least one flight connects them during that day. Bottom panels: Daily number of distressed airports (left bottom panel) and the fraction of connected nodes among them (right bottom panel) on the period from January 1st 2015 to March 31st 2015.

3.3 Centrality metrics analysis

In this Section, we first apply the existing centrality metrics presented in Section 2.2.1.1 to the two datasets. We show that these metrics are not able to identify changes of centrality resulting from delays. As we discussed, this is due to the fact that such metrics are static and do not consider the time ordering of walks. Therefore, we proceed by applying the new Trip centrality metrics, showing

that they succeed in identifying those airports that experience a large loss of centrality (or of centrality ranking) as a consequence of delays. To perform comparisons between rankings of airports according to different metrics or computed on different networks (e.g., scheduled and actual), we use the Kendall rank correlation coefficient τ, which measures the similarity of two ranked sequences of data. The coefficient takes values in [-1,1], with the value 1 corresponding to two identical sequences and the value -1 to two sequences that are one the inverse of the other.

3.3.1 Baseline centrality metrics

For Katz centrality, we choose α = 0.004 for the ECAC dataset and α = 0.003 for the US dataset, the largest values which assures the convergence of the metric. Note that these small values of α penalise strongly long walks. We use the same value of α for PageRank, to allow comparison, although a large value of α would have still assured convergence.

By applying the four metrics to the two datasets, we find that for each metric the rankings according to the outgoing metric and to the incoming metric are very similar. In the ECAC dataset, they display values of τ respectively of 0.96, 0.98, 0.96, 0.90 on the scheduled network and 0.97, 0.97, 0.96, 0.90 on the actual one. For the US dataset the correlation coefficients, averaged over all days, are respectively 0.98, 0.99, 0.99, 0.99 on the scheduled network and 0.97, 0.98, 0.97, 0.95 on the actual one.

The ranking produced by the different centrality metrics are correlated but not identical. A visual comparison of the rankings on the scheduled network in the incoming case is shown in Figure 13 for the ECAC dataset, and in Figure 14 for one specific day of the US dataset. On the scheduled network the rankings according to incoming degree centrality and incoming strength centrality have a correlation coefficient τ=0.83 (ECAC) and τ=0.84 (US, average value), those according to incoming degree centrality and incoming Katz centrality have τ=0.74 (ECAC) and τ=0.76 (US, average value), those according to incoming strength centrality and incoming Katz centrality τ=0.85 (ECAC) and τ=0.89 (US, average value) and those according to incoming Katz centrality and incoming PageRank centrality τ=0.59 (ECAC) and τ=0.67 (US, average value). Results are similar in the outgoing case and in the actual network. These differences highlight the fact that different centrality metrics describe different aspects of the network structure, and care should be taken in their comparison. For example, degree and strength consider only direct links, therefore they are appropriate if we are interested in assessing the potentiality of an airport to provide direct connections to other airports of the network but are not able to evaluate the role of flight connections. Additionally, an airport with many different destinations might be very central according to degree centrality, but according to strength it might be surpassed by one with less destinations, but more flights directed to each of these destinations. Katz centrality and PageRank, instead, take into account also walks of any length on the network. While walks on the aggregated, static network do not correspond to real itineraries that can be followed, accounting for longer walks means attributing centrality to an airport if it is connected to other central airports. Therefore, these two metrics are more appropriate when we want to assess the potentiality of an airport to provide connections to other airports of the network with walks of any length. As a consequence of the different ways of weighting walks in the two metrics, Katz centrality favours airports linked to large airports (with many links), as they will have many walks departing or arriving, while PageRank rather tends to favour airports with more links to smaller sized airports. Table 10 and Table 11 compare the top ten airports according to the different centralities in the two datasets. Note that the dataset ECAC1 on which the metrics were applied

contains only intra-ECAC flights, which explains why airports as, e.g., the Manchester airport enter the top ten.

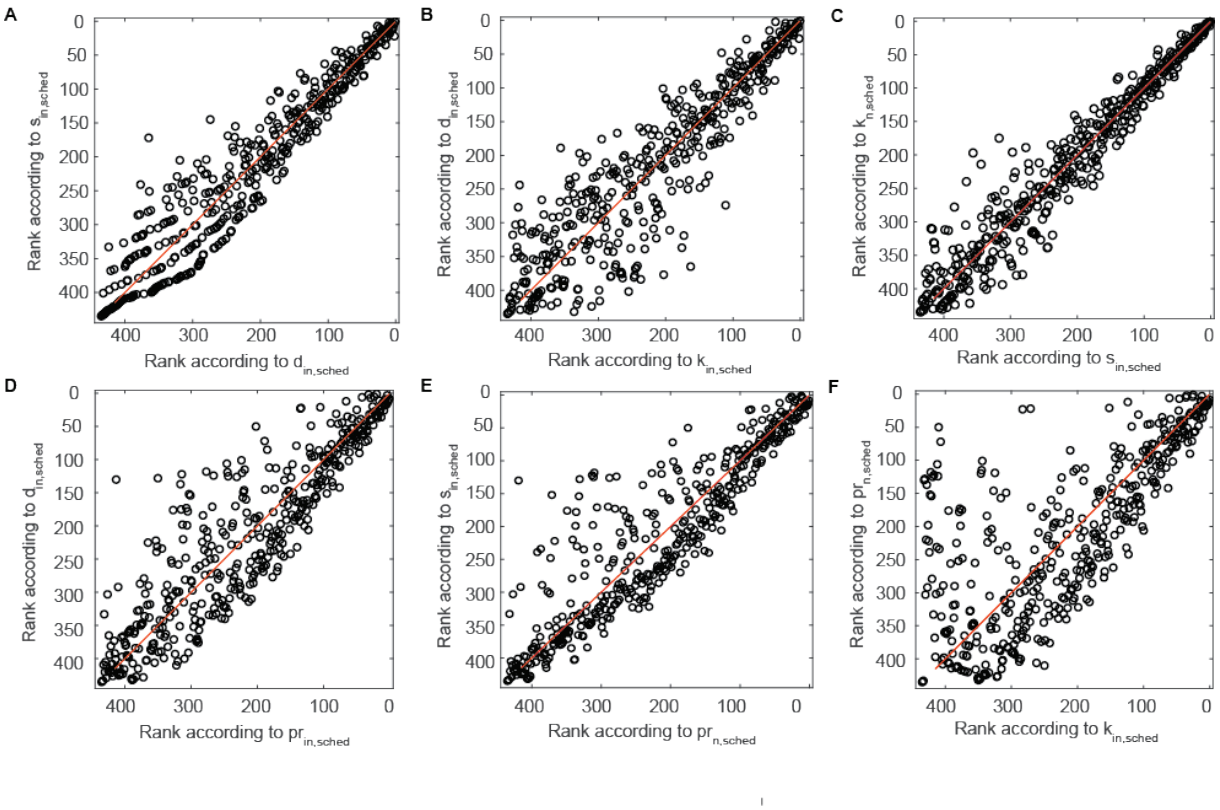

Figure 13. ECAC dataset, comparison of airports' rankings according to different incoming centrality metrics on the scheduled network.A) Degree versus strength; B) Degree versus Katz; C) Strength versus Katz; D) Degree versus PageRank; E) Strength versus PageRank; F) Katz versus PageRank.

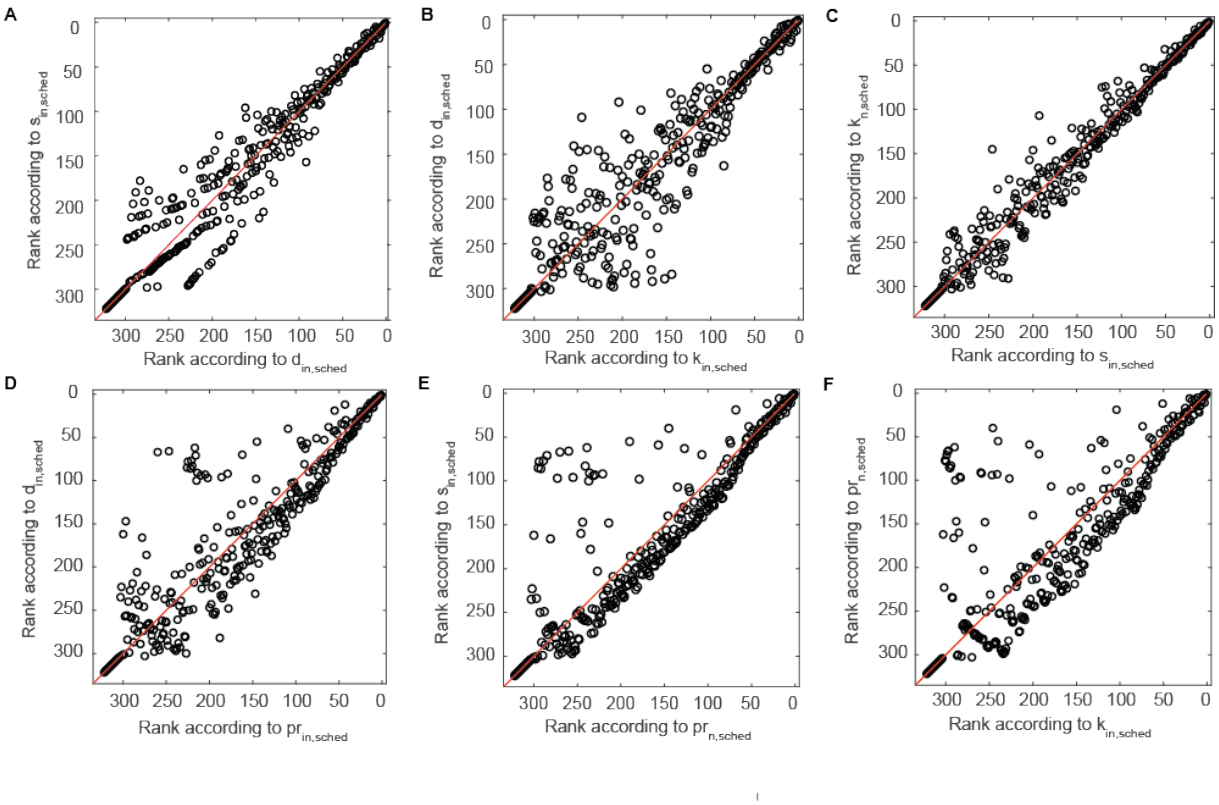

Figure 14. US dataset, comparison of airports' rankings according to different incoming centrality metrics on the scheduled network. A) Degree versus strength; B) Degree versus Katz; C) Strength versus Katz; D) Degree versus PageRank; E) Strength versus PageRank; F) Katz versus PageRank.

Table 10. Top-ten airports in the ECAC1 dataset according to the four metrics.

Degree	Strength	Katz	PageRank
Amsterdam Airport Schiphol	Amsterdam Airport Schiphol	Amsterdam Airport Schiphol	Stockholm-Arlanda Airport
London Stansted Airport	Frankfurt am Main Airport	London Heathrow Airport	Helsinki Vantaa Airport
Barcelona International Airport	Munich Airport	Frankfurt am Main Airport	Atatürk International Airport
Atatürk International Airport	Adolfo Suárez Madrid–Barajas Airport	Munich Airport	Eleftherios Venizelos International Airport
London Gatwick Airport	Barcelona International Airport	Barcelona International Airport	Oslo Gardermoen Airport
Munich Airport	Charles de Gaulle International Airport	Charles de Gaulle International Airport	Paris-Orly Airport
Adolfo Suárez Madrid–Barajas Airport	London Heathrow Airport	Adolfo Suárez Madrid–Barajas Airport	Sabiha Gökçen International Airport
Charles de Gaulle International Airport	Leonardo da Vinci–Fiumicino Airport	Leonardo da Vinci–Fiumicino Airport	Amsterdam Airport Schiphol
Manchester Airport	London Gatwick Airport	Copenhagen Kastrup Airport	Adolfo Suárez Madrid–Barajas Airport
Frankfurt am Main Airport	Oslo Gardermoen Airport	Zürich Airport	London Stansted Airport

Table 11. Top-ten airports in the US_April dataset according to the four metrics.

Degree	Strength	Katz	PageRank
Hartsfield-Jackson Atlanta International Airport	Hartsfield-Jackson Atlanta International Airport	Hartsfield-Jackson Atlanta International Airport	Hartsfield-Jackson Atlanta International Airport
Chicago O'Hare International Airport	Chicago O'Hare International Airport	Chicago O'Hare International Airport	Dallas/Fort Worth International Airport
Dallas/Fort Worth International Airport	Dallas/Fort Worth International Airport	Los Angeles International Airport	Chicago O'Hare International Airport
Denver International Airport	Los Angeles International Airport	Dallas/Fort Worth International Airport	Denver International Airport
George Bush Intercontinental Airport	Denver International Airport	Denver International Airport	George Bush Intercontinental Airport
Minneapolis-Saint Paul International Airport	George Bush Intercontinental Airport	San Francisco International Airport	Salt Lake City International Airport
Detroit Metropolitan Airport	San Francisco International Airport	McCarran International Airport	Detroit Metropolitan Airport
Salt Lake City International Airport	Phoenix Sky Harbor International Airport	Phoenix Sky Harbor International Airport	Minneapolis-Saint Paul International Airport
McCarran International Airport	McCarran International Airport	Gen. Edward Lawrence Logan International Airport	Phoenix Sky Harbor International Airport
Phoenix Sky Harbor International Airport	Detroit Metropolitan Airport	LaGuardia Airport (Marine Air Terminal)	San Francisco International Airport

Most importantly, for all four metrics the ranking of airports in the scheduled network is extremely similar to the ranking in the actual network. For the three incoming metrics, τ is, respectively, 0.996, 0.991, 0.985 and 0.990 for the ECAC data set and 0.994, 0.980, 0.986, 0.986 for the US dataset. Similar results hold for the outgoing metrics. The reason of this similarity is that the metrics neglect the dynamic structure of the network, therefore the presence of delays changing the links' schedule does not have any effect on the rankings. The only differences between the scheduled network and the actual one which these metrics capture are the cancelled flights.

This result makes clear that these metrics are not suitable to evaluate the effect of the innovations addressed by Domino on the network performance, as they would not be able tell apart a situation where delays disrupt connections to one where they do not. To discern these situations, an airport's centrality should reflect its participation to walks that can actually be travelled, i.e., respecting the schedule, so that disrupted connections imply a centrality drop. This, in turn, requires accounting for the temporal structure of the network. Katz and PageRank centrality, in particular, count walks on the network which are not time ordered and therefore have no relationship with the trajectories that passengers could travel. As a consequence, these metrics cannot reflect the effect of delays on the network's connectivity.

An additional limitation of Katz and PageRank centrality is that the weight assigned to each walk does not consider to which airline each flight composing the walk belongs, therefore a walk using only flights of one airline has the same weight of a walk of the same length using several airlines. However, a more realistic assumption would be that the latter contributes less to centrality, or not at all, as it is travelled with a smaller probability. Accounting for this requires considering the multiplex structure of the network.

3.3.2 Trip centrality metrics

The Trip Centrality, TripRank and 2-legs Trip Centrality metrics are applied to both the ECAC2 and the US_April dataset. The choice to use the reduced dataset for the European case is due to the fact that, for each additional layer, adjacency matrices increase their size considerably. Including all the 183 airlines present in the ECAC1 dataset would be computationally very heavy. The length Δt of a time frame is chosen as a rounded number of minutes, corresponding to the shortest flight in the dataset, that is, 15 minutes for the ECAC2 dataset and 20 minutes for the US_April dataset. In future iterations, extreme outliers such as these may be removed, and a percentile cut-off used. Each airline constitutes one layer, except for airlines belonging to the same alliance, which are aggregated in a single layer. This choice produces 32 layers for the ECAC2 dataset and 14 for the US dataset. Table 12 summarizes this information and reports how many airports and flights each dataset has.

Table 12. Summary of the two datasets ECAC2 and US_April.

	ECAC2	US_April
Number of flights	19 648	16 042 (on average)
Number of layers	32	14
Number of airports	435	322
Δt(min)	15	20

The parameter α was chosen to be $= e^{-1/4}$, so as to assign a negligible weight to paths composed of more than 2 flights (4 links, counting the jumps to the secondary nodes). In fact, with this choice, a path of k legs, with k>2, has a weight $e^{-2k/4} \ll 1$. This choice is motivated by data on passenger itineraries for September 1st, 2014, containing the number of passengers following one, two or three legs itineraries (with the same airline). A total of 1 272 477 passengers follow a one leg itinerary, 227 445 follow a two-legs itinerary and 10 429 a three-legs one. We can estimate the maximum number of legs L for which we should assign a non-negligible weight by assuming that the number of passengers following an itinerary with i legs, n_i, decreases exponentially when i increases, i.e., $n_i = n_0 e^{-i/L}$, where L is therefore interpreted as the maximum number of legs of itineraries with a non-negligible number of passengers. We therefore have $n_{i+1}/n_i = e^{-1/L}$. Subtituting $n_1 = 1\,758\,655$ and $n_2 = 227\,445$we obtain an estimate L=1.35, while substituting $n_2 = 227\,445$ and $n_3 = 10\,429$ we obtain L=0.74. However, these calculations underestimate the real L, because we are only considering itineraries within the same airlines or alliance, while two- and three- legs itinerary could

also be formed by flights of different airlines, which do not belong to the same alliance. Therefore, we take the rounded up estimate L=2 as our estimate of the maximum number of legs to consider.

For the parameter ε , instead, we compare the results relative to the value$\varepsilon = 0, e^{-1}, e^{-1/2}, e^{-1/3}, 1$, describing respectively a situation where inter-layer jumps are forbidden ($\varepsilon = 0$), where paths with up to 1, 2 or 3 inter-layer jumps give a non-negligible contribution to centrality and where there is no difference between inter-layer and intra-layer jumps ($\varepsilon = 1$).

In the following we compare the results of the different metrics on the two datasets. Unless explicitly stated, we refer to the aggregated centralities (and not to the layer-specific ones).

3.3.2.1 Change of airports' ranking for different values of ε

Figure 15 and Figure 16 show the evolution of the airports' ranking according to their incoming and outgoing Trip centrality on the scheduled network and to incoming and outgoing TripRank when ε changes from 0 to 1, for the ECAC2 dataset. Figure 17 and Figure 18 show the same for one day (April 9th) of the US_April dataset. It appears clear that major changes in the ranking occur when passing from $\varepsilon = 0$ to $= e^{-1}$, while further increases of the parameter do not introduce large changes. To quantify this observation, in the ECAC2 dataset, the correlation coefficient between the ranking induced by incoming Trip centrality with $\varepsilon = 0$ and that with $\varepsilon = e^{-1}$ is $\tau = 0.73$, while that between $\varepsilon = e^{-1}$ and $\varepsilon = e^{-1/2}$ is already 0.95. For the US dataset, the two correlations coefficients are, respectively, 0.82 and 0.95 (average on all days). Similar results hold for the outgoing case. For TripRank, in the incoming case the correlation coefficients are respectively 0.90 and 0.96 for the ECAC dataset, and 0.88 and 0.96 for the US one (average over all days), and similar results hold for the outgoing case. The presence of major ranking changes when passing from $\varepsilon = 0$ to $\varepsilon = e^{-1}$ means that the possibility of using across-layer paths influences strongly the connectivity of the network, increasing the centrality of a good number of airports, which therefore climb the ranks. However, the smaller changes with further increases of ε mean that the weight given to an inter-layer jump does greatly influence the ranking, meaning that paths with one inter-layer jumps are the most important for the connectivity increase. With TripRank, the ranking change when passing from $\varepsilon = 0$ to $\varepsilon = e^{-1}$ is smaller than with Trip centrality, signifying that many of the newly allowed inter-layer paths pass from large hubs, having a large number of outgoing/incoming flights, and are therefore given a small weight by TripRank. The results are very similar for the actual network, meaning that such conclusions are quite robust. In the following, unless explicitly specified we consider the case $\varepsilon = 0$. In some cases, we will compare the results with the ones obtained with $\varepsilon = e^{-1/2}$, as a representative instance of the cases where inter-layer walks are permitted.

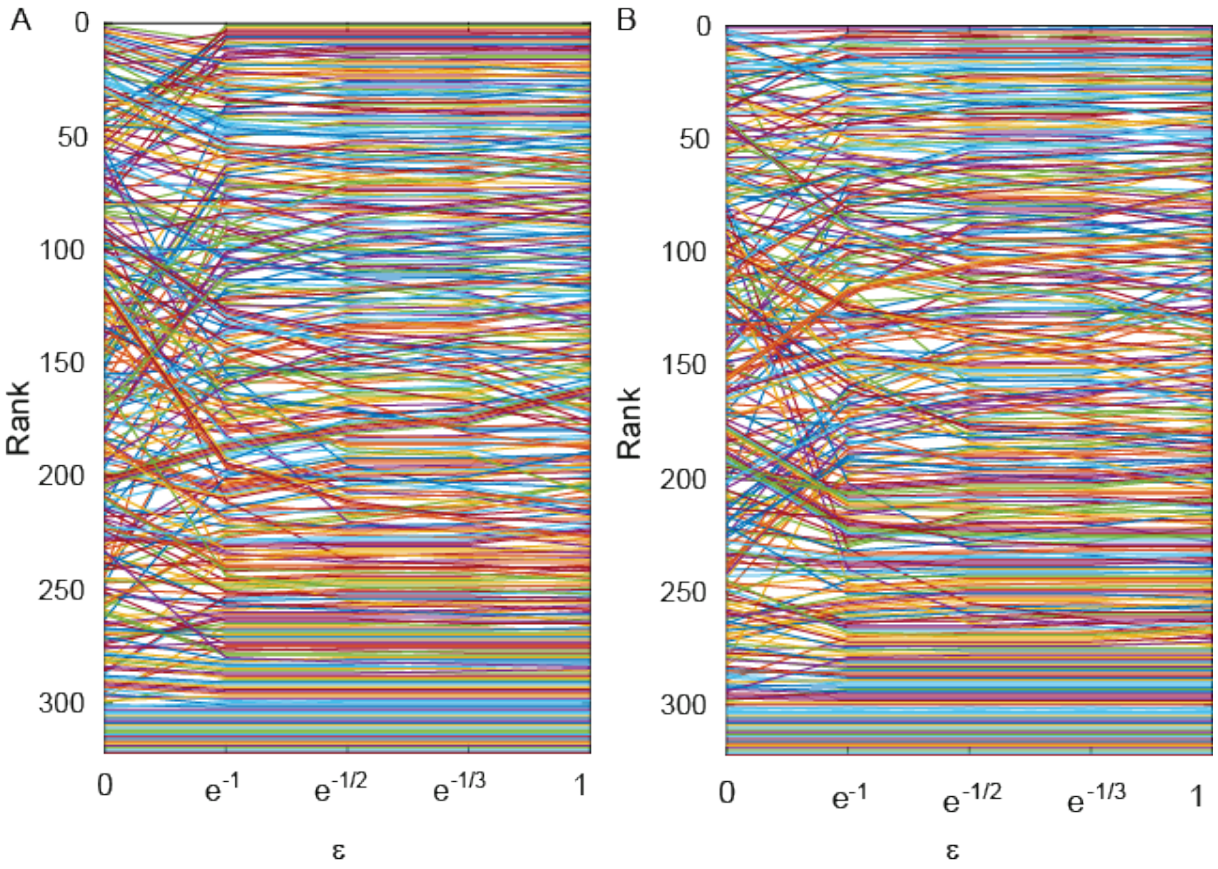

Figure 15. Trip centrality evolution of airports' ranking in ECAC2 dataset. Evolution as a function of parameter ε from 0 to 1.Shown are the ranking according to incoming Trip centrality (A) and outgoing Trip centrality (B). Each line represents one airport.

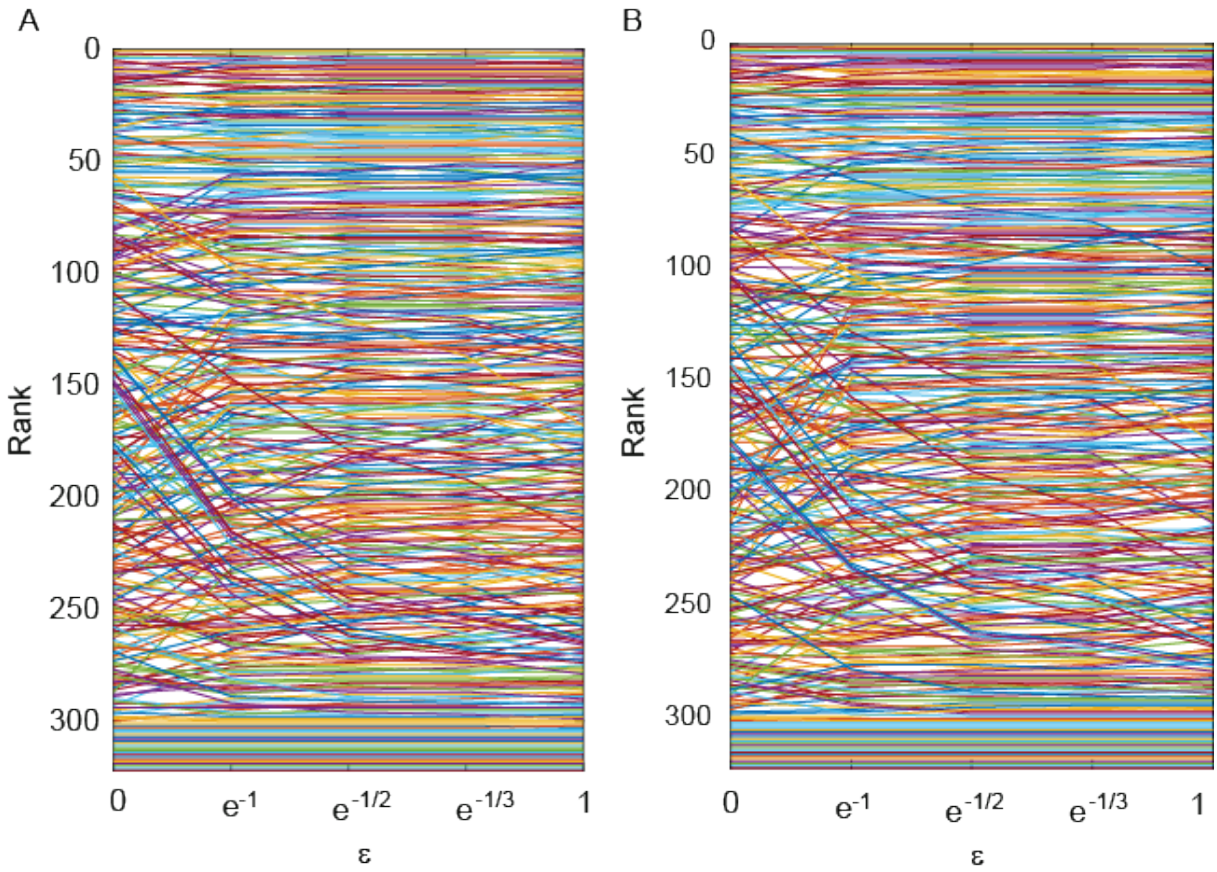

Figure 16. TripRank evolution of airports' ranking in ECAC2 dataset. Evolution as a function of parameter ε from 0 to 1. Shown are the ranking according to incoming TripRank (A) and outgoing TripRank (B). Each line represents one airport.

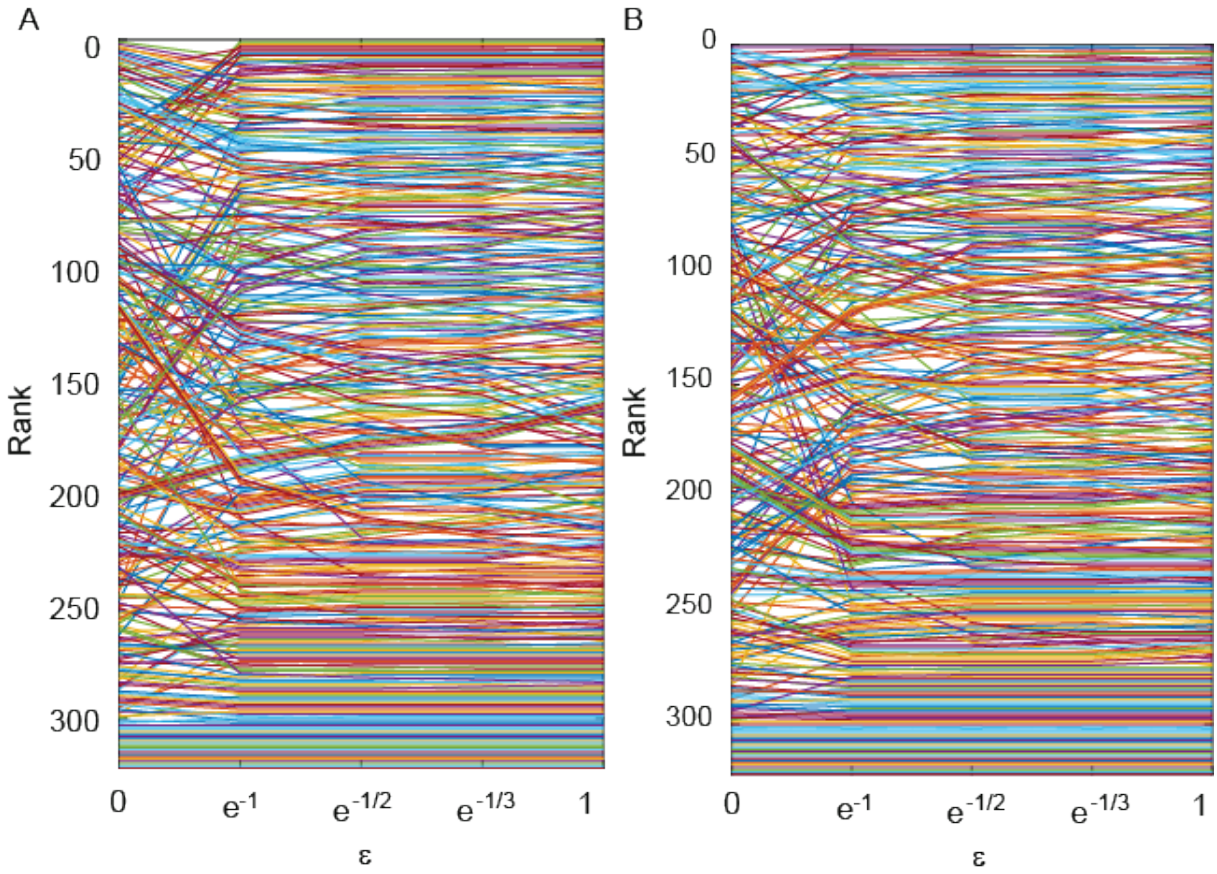

Figure 17. Trip centrality evolution of airports' ranking in US_April dataset. Evolution as a function of parameter ε from 0 to 1.Shown are the ranking according to incoming Trip centrality (A), outgoing Trip centrality (B). Each line represents one airport.

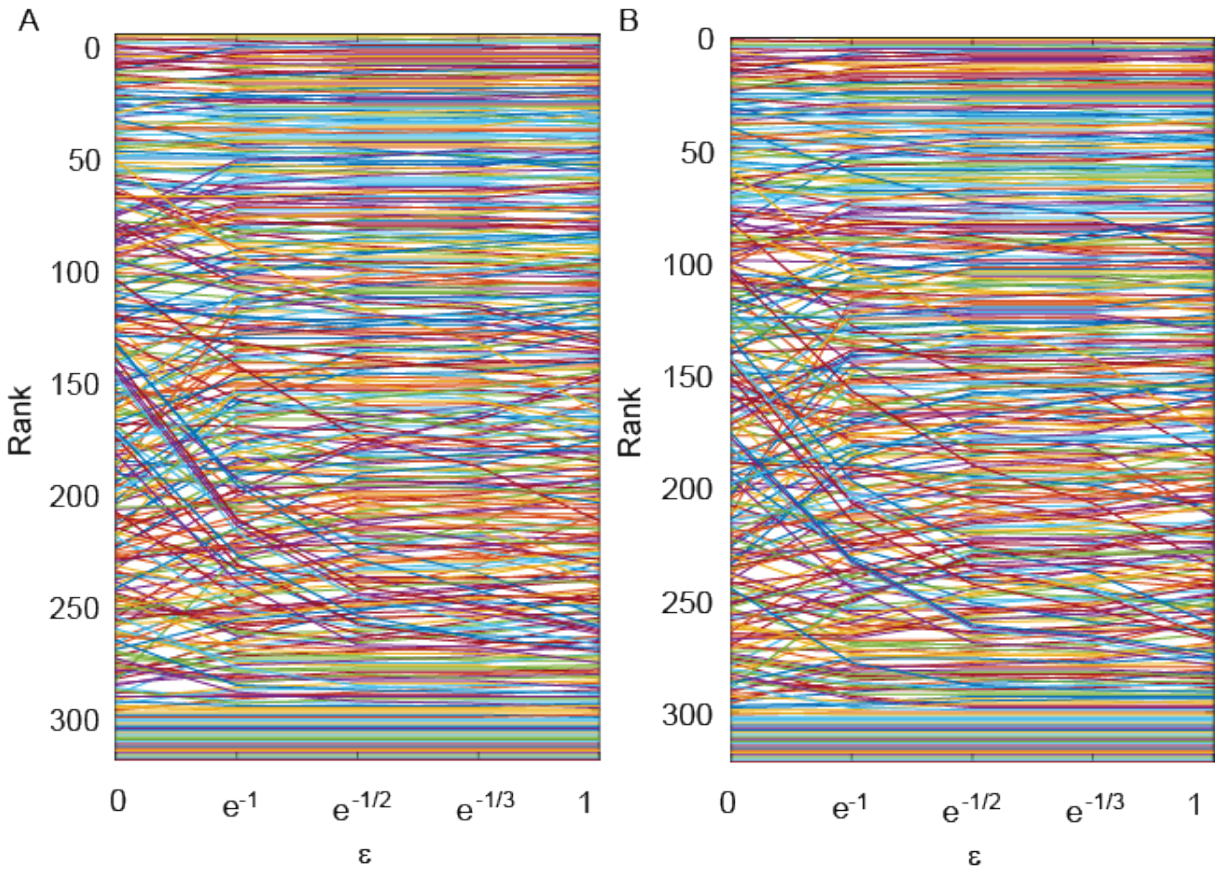

Figure 18. TripRank evolution of airports' ranking in US_April dataset. Evolution as a function of parameter ε from 0 to 1.Shown are the ranking according to incoming TripRank (C) and outgoing TripRank (D).

3.3.2.2 Comparison between incoming and outgoing centrality

Differently from the centrality metrics on the time-aggregated network, on the temporal network there is a noticeable difference between the ranking according to incoming centrality and that according to outgoing centrality. In the ECAC2 dataset, the Kendall correlation coefficient between the two is 0.77 with Trip centrality, and 0.89 with TripRank, both on the scheduled and on the actual network. The rankings are compared visually in Figure 19. This difference between the incoming and outgoing centrality of airports is due to the fact that we are accounting for flights schedules. In fact, although the number of incoming and outgoing flights from an airport are very similar (as we have seen in Section 3.3.1), the incoming and outgoing connections can differ. For example, imagine that an airport A has one incoming flight and one outgoing flight. If both flights depart early in the morning, the outgoing flight will have more available outgoing connections than the incoming connections of the incoming flight, due to its early time of departure. Therefore, this couple of flights will provide to A more outgoing centrality than incoming one. This is just one example of how this asymmetry emerges. In Figure 20 each airport is coloured according to its change of rank from the incoming to the outgoing case. In the case of Trip centrality (panel A), there is a noticeable geographical dependence of the rank change, with south eastern airports typically gaining ranks and western airport typically losing ranks. Such geographical asymmetry is not evident in the TripRank case (panel B).

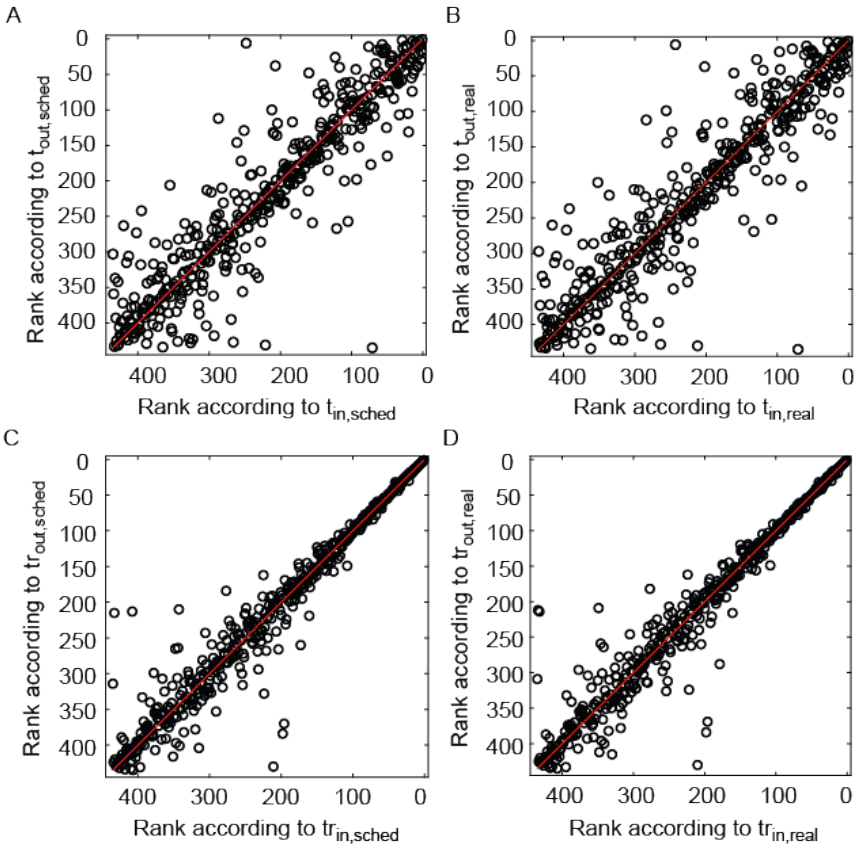

Figure 19. Comparison of incoming and outgoing airports' rankings in the ECA2 dataset. A) Trip centrality on scheduled network; B) Trip centrality on actual network, C) TripRank centrality on scheduled network; D) TripRank centrality on actual network. Red lines are 1:1 lines.

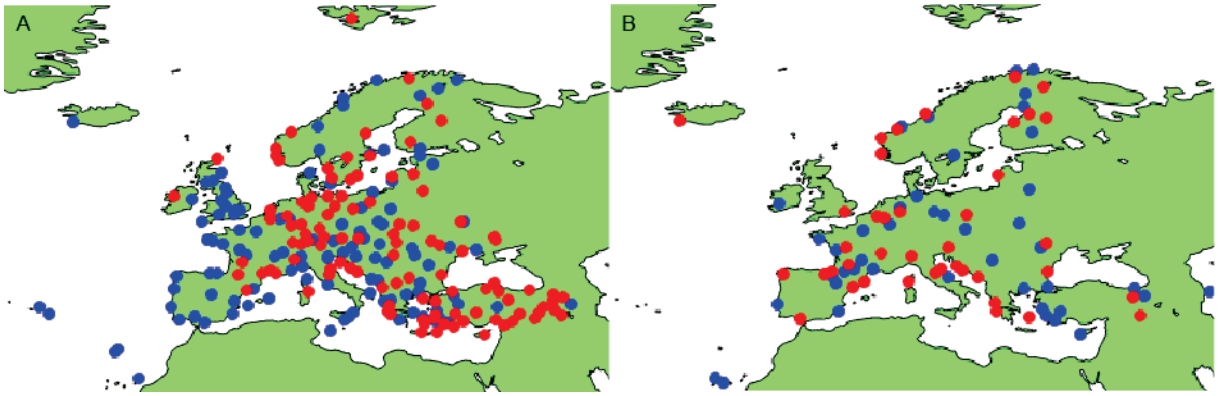

Figure 20. Comparison of ranking according to incoming and outcoming Trip centrality for ECAC dataset (panel A) and TripRank (panel B). The airports' colour corresponds to the difference of rank between incoming and outgoing centrality: an airport is coloured in red if its outgoing centrality is more than 10% larger than its incoming one, in blue if it is more than 10% smaller. Airports with smaller rank differences are not plotted.

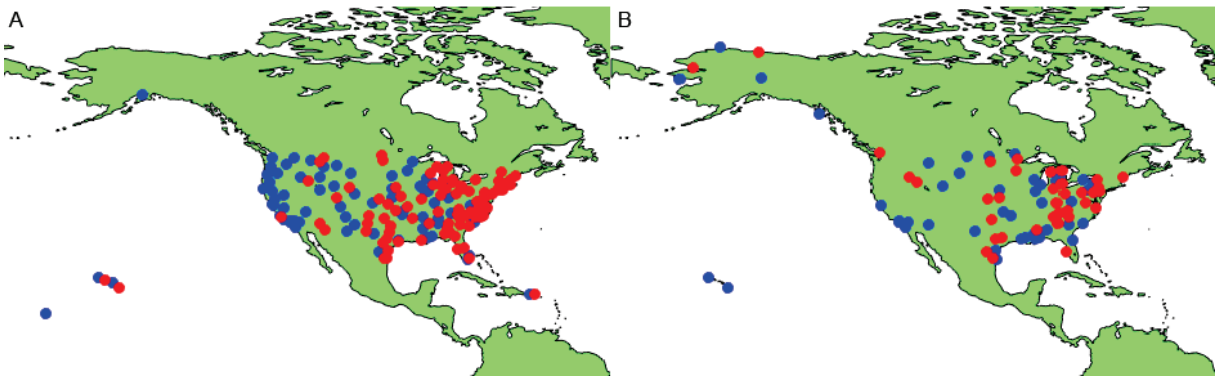

Figure 21. Comparison of ranking according to incoming and outcoming Trip centrality for US dataset (panel A) and TripRank (panel B). The airports' colour corresponds to the difference of rank between incoming and outgoing centrality: an airport is coloured in red if its outgoing centrality is more than 10% larger than its incoming one, in blue if it is more than 10% smaller. Airports with smaller rank differences are not plotted.

The geographical asymmetry between incoming and outgoing centrality is even more evident in the US_April dataset, as shown by Figure 21, relative to April 9th. On average, for the schedule network, the Kendall correlation coefficient between the two is 0.81 with Trip centrality, and 0.89 with TripRank, for both the scheduled and the actual network. The reason for this geographical asymmetry is that the time difference between the East Coast and the West Coast is such that the first flights of the day on the East Coast precede of roughly 3 hours the first flights on the West Coast, therefore producing higher outgoing centralities on the East Coast. On the contrary, incoming centralities are higher on the West Coast, where the last flights of the day land roughly 3 hours later than on the East Coast. A similar reasoning explains the asymmetry in the European case.

3.3.2.3 Comparison between Trip centrality and TripRank

The ranking produced by Trip centrality and by TripRank centrality are correlated but quite different, as is shown Figure 22 for the ECAC2 dataset. For the scheduled incoming case, the correlation coefficient is $\tau = 0.43$ when $\varepsilon = 0$ and 0.64 when $= e^{-1/2}$, therefore increasing ε makes the two metrics more similar. Similarly, for the US case, on average the two correlation coefficients are 0.64 and 0.68. Table 13 and Table 14 compare the top ten airports according to the two centralities, in the scheduled incoming case, respectively for the ECAC2 dataset and for April 9th of the US_April dataset. Results are similar for the scheduled outgoing case and for the actual network. The difference between the two metrics is to be expected due to the different weights given to walks. In particular, an airport which is well connected to an airport with large degree will have a large Trip centrality, but its TripRank centrality will be much smaller.

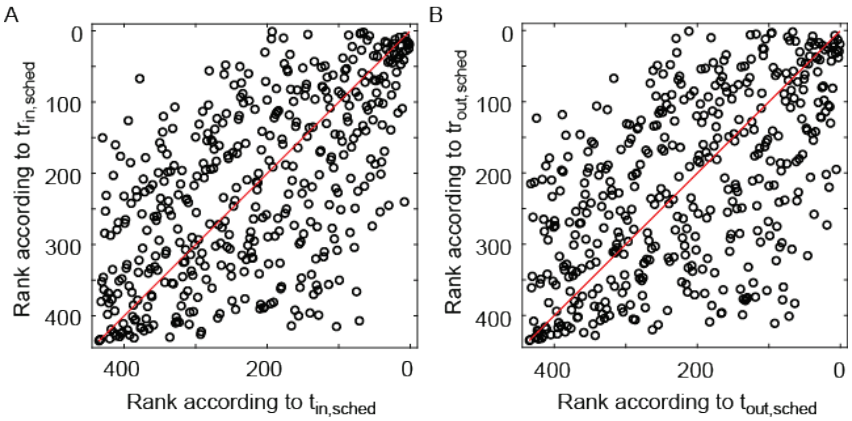

Figure 22. Comparison between rankings for Trip centrality and TripRank in ECAC2. A) According to incoming, for the scheduled network with ε=0. B) According to outgoing, for the scheduled network with ε=0. Red lines are 1:1 lines.

Table 13. Top-ten airports in the ECAC2 dataset according to incoming Trip centrality and TripRank for the scheduled network with ε=0.

Trip centrality	TripRank
Atatürk International Airport	London Gatwick Airport
Copenhagen Kastrup Airport	Amsterdam Airport Schiphol
Stockholm-Arlanda Airport	Barcelona International Airport
Henri Coandă International Airport	Adolfo Suárez Madrid–Barajas Airport
Munich Airport	Sabiha Gökçen International Airport
Geneva Cointrin International Airport	Dublin Airport
Oslo Gardermoen Airport	Düsseldorf Airport
Graz Airport[9]	Oslo Gardermoen Airport
Brussels Airport	London Heathrow Airport
Vienna International Airport	Palma De Mallorca Airport

[9] The high centrality of Graz Airport is potentially explained by incoming flights from large airports (Munich, Vienna, Dusseldorf and Stuttgart) in the last part of the day, bringing large contributions to centrality.

Table 14. Top-ten airports for the day April 9th of the US_April dataset according to incoming Trip centrality and TripRank for the scheduled network with ε=0.

Trip centrality	TripRank
McCarran International Airport	Hartsfield-Jackson Atlanta International Airport
Phoenix Sky Harbor International Airport	Chicago O'Hare International Airport
Oakland International Airport	Dallas/Fort Worth International Airport
Los Angeles International Airport	Denver International Airport
San Diego International Airport (Lindbergh Field)	George Bush Intercontinental Airport
Sacramento International Airport	Seattle-Tacoma International Airport
Ontario International Airport	Los Angeles International Airport
Norman Y. Mineta San José International Airport	Minneapolis-Saint Paul International Airport
Denver International Airport	Phoenix Sky Harbor International Airport
San Francisco International Airport	Detroit Metropolitan Airport

3.3.2.4 Comparison between standard metrics and Trip centrality metrics

Figure 23, panels A to D, shows a comparison of the rankings produced by the standard Katz and PageRank centralities presented in Section 3.3.1 and their 'Trip' counterparts introduced here, for the ECAC2 dataset. As expected, the rankings are very different. When $= 0$, on the scheduled network, the correlation coefficient between the rankings according to incoming Katz and Trip centrality is $\tau = 0.59$, while for incoming PageRank and TripRank it is $\tau = 0.77$. A comparison between the top ten airports according to the standard and Trip metrics can be done comparing Table 10 and Table 13. Similar results hold for the outgoing case and on the actual network. When $\varepsilon = e^{-1/2}$, the rankings according to incoming Katz and Trip centrality become more similar ($\tau = 0.67$), while those according to incoming PageRank and TripRank become less similar ($\tau = 0.67$).

This difference between the standard and the Trip metrics is to be expected, as the standard metrics consider many more walks, most of which are not time-ordered and use flights of different alliances or airlines.

3.3.2.5 Comparison between Trip centrality and 2-legs Trip centrality

Here we compare the ranking produced by Trip centrality with $\varepsilon = 0$ and 2-legs Trip centrality where only connections between flights of the same alliance (or airline, in case the airline does not belong to an alliance) are counted and where no threshold on connecting time is imposed. The only difference between the two compared metrics is that the former considers also walks of more than two-legs, while the latter does not. A notable difference is found in the rankings, as shown in Figure 23, panels E and F, for the ECAC2 dataset. The correlation coefficients are $\tau = 0.62$ in the incoming case and $\tau = 0.59$ in the outgoing case. Table 15 compares the top ten airports according to the two centralities, in the scheduled incoming case. This difference proves that walks longer than two legs have an important role in determining the ranking according to Trip centrality. Reducing the value of

α makes the two metrics more and more similar, since longer walks are counted less and less by Trip centrality.

Table 15. Top-ten airports for the ECAC dataset according to incoming Trip centrality with ε=0 and incoming 2-legs centrality for the scheduled network.

Trip centrality	TripRank
Atatürk International Airport	London Heathrow Airport
Copenhagen Kastrup Airport	Amsterdam Airport Schiphol
Stockholm-Arlanda Airport	Munich Airport
Henri Coandă International Airport	Barcelona International Airport
Munich Airport	Frankfurt am Main Airport
Geneva Cointrin International Airport	Adolfo Suárez Madrid–Barajas Airport
Oslo Gardermoen Airport	Charles de Gaulle International Airport
Graz Airport[9]	Copenhagen Kastrup Airport
Brussels Airport	Leonardo da Vinci–Fiumicino Airport
Vienna International Airport	London Gatwick Airport

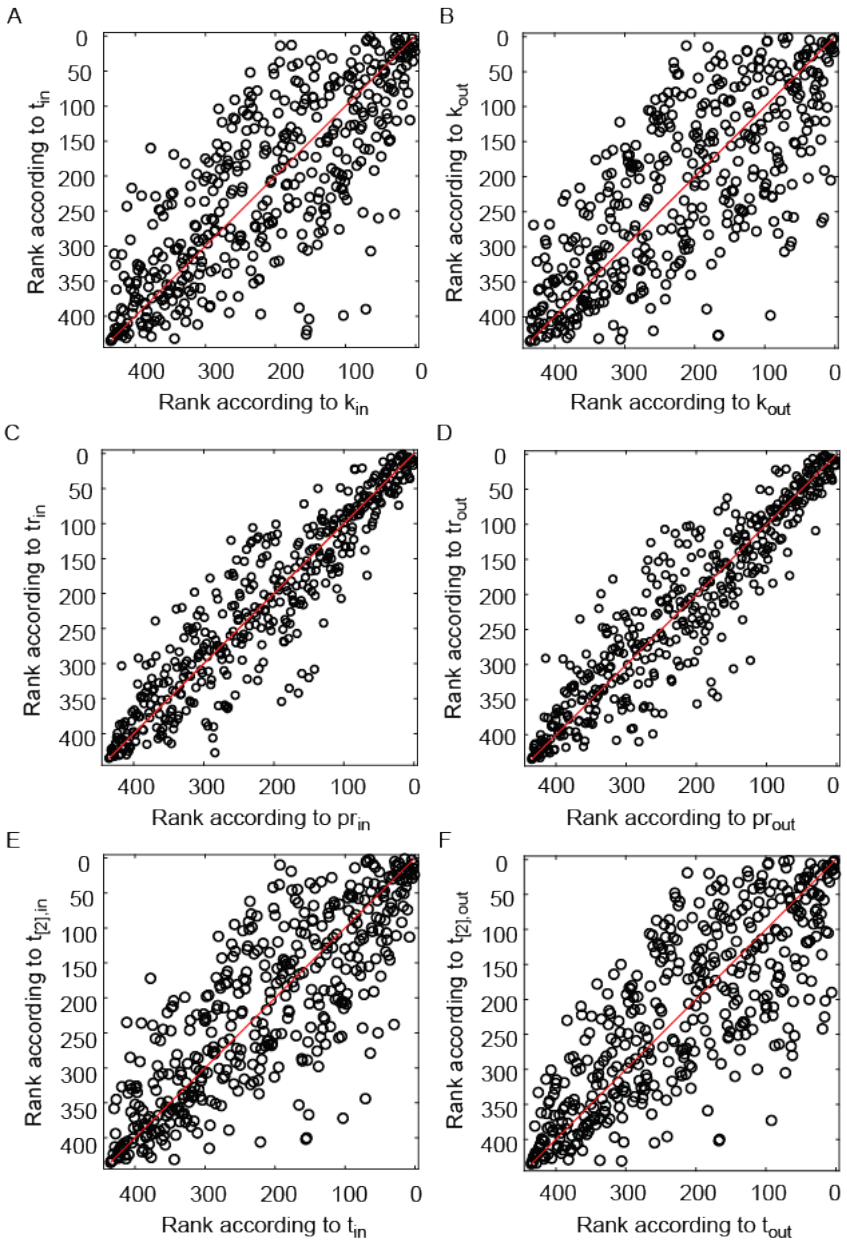

Figure 23. Airport ranking and centrality for ECAC2. Panels A-D: Comparison of the airports' ranking according to standard and "Trip" metrics for the scheduled network in the ECAC2 dataset, ε=0, A) incoming Trip centrality; B) outgoing Trip centrality; C) incoming TripRank; D) outgoing TripRank; Panels E and F: comparison between Trip centrality with ε=0 and 2-legs Trip centrality for the scheduled network in the ECAC2 dataset, E) incoming, F) outgoing.

3.3.2.6 Comparison between centralities on the scheduled and on the actual network

Centralities are always lower in the actual network with respect to the scheduled network, due to cancelled flights and disrupted connections. A decrease of centrality, however, does not always imply a change of rank of the airport. In fact, overall the ranking in the scheduled and in the actual network are very similar, according to all metrics.

Let us first consider the ECAC2 dataset. Figure 24 compares the rankings obtained with each metric in the actual and scheduled network. The red dots represent distressed airports. Specifically, airports where the fraction of departing flights with departure delay larger than the average is, itself, larger than the average fraction of delayed flights. In Figure 25, instead, the percentage centrality loss for each airport between the scheduled and the actual network is plotted against the centrality of the airport in the scheduled network. In each plot, red and gray lines are curves obtained by local smoothing (i.e., LOWESS: locally weighted scatterplot smoothing) of, respectively, the red and black dots, and show the trend of the loss of centrality for distressed and not distressed airports.

From the figures, we see that not all distressed airports lose rank or a large percentage of centrality, not all airports lose rank nor is a large percentage of centrality associated with distress. This means that comparing centralities in the scheduled and actual networks provides different information than that provided by delay statistics. In fact, the incoming centrality of an airport is not directly linked to the delays of flight departing from (or arriving to) that airport, but rather to those of flights coming from other airports, which are part of incoming walks. Therefore, the loss of incoming centrality (or of the corresponding rank) of an airport informs us of the impact of delays in the entire network on that particular airport. In the case of outgoing centrality, departing delays of flight departing from an airport do have a primary role in determining the disruption of outgoing connections, and therefore the decrease of outgoing centrality. In fact, in the outgoing case, we remark that airports losing rank are more often distressed ones, i.e., dots under the diagonal are often red, and that distressed airports tend to lose a larger percentage of outgoing centrality than non-distressed one, i.e., the red curves in Figure 25 B), D) and F) are above the grey curves. In particular, the average percentage of lost outgoing centrality for distressed airport is 25% according to Trip Centrality, 11% according to 2-legs Trip centrality and 5% according to TripRank, while for non-distressed airports the percentages are, respectively, 16%, 4% and 2%. This effect is maintained also if we consider the absolute loss of centrality, instead of the percentage loss. However, small delays can also cause the disruption of connections, which explains why non-distressed airports can have a rank loss or a large percentage centrality loss. Similarly, large delays can have no impact on centrality if they happen on walks with a large connecting time, which explain why not all distressed airports lose rank.

Although the rankings remain very similar in the scheduled and in the actual network, some airports lose large percentages of their centrality, meaning that many connections are lost. The percentage of lost centrality, averaged on all airports, is a measure of the impact of delays on the overall connectivity of the network. It can be used to compare the performances of different scenarios at a whole network level. The percentage of lost centrality of a particular airport can also be used to compare the effects of different scenarios on that airport. If the layer-specific centrality is used for the comparison, it informs on the effects of delay on that layer (i.e., that airline or alliance).

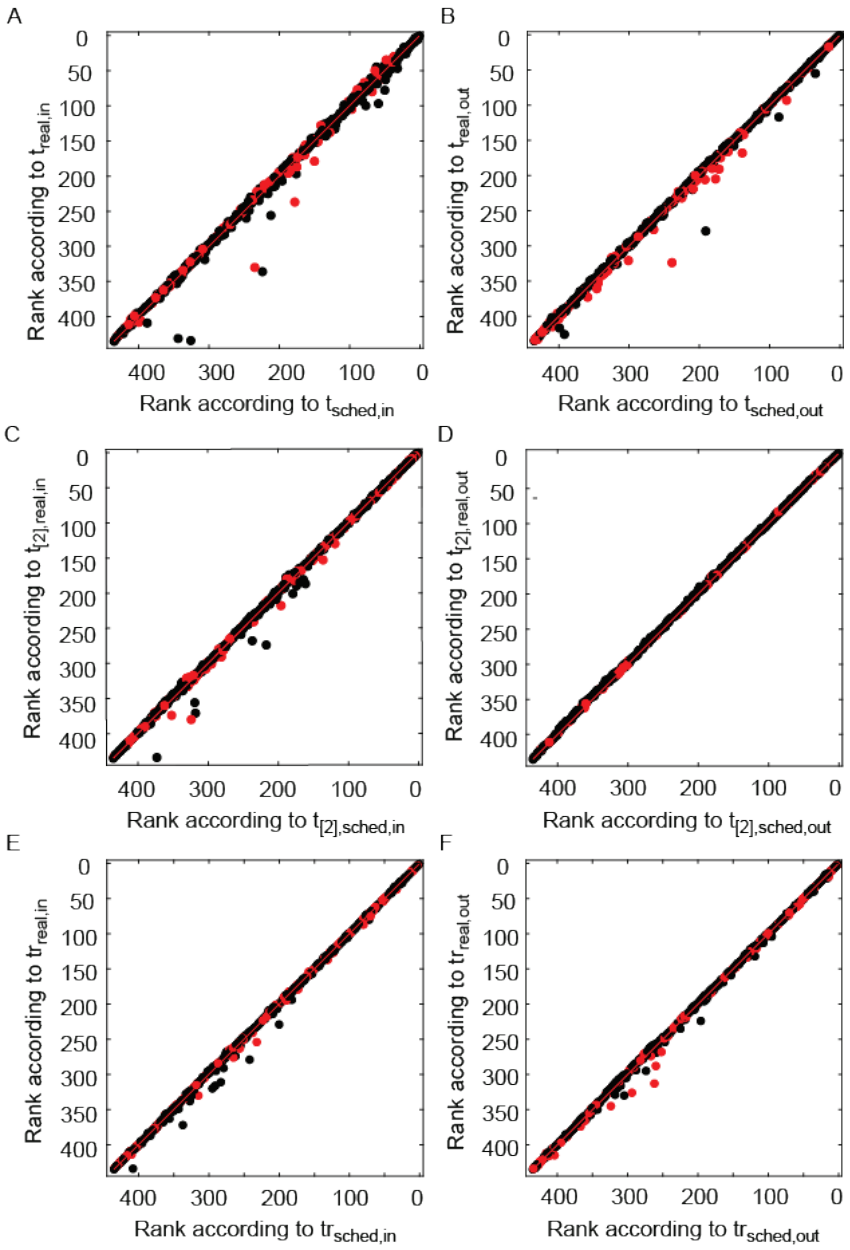

Figure 24. Comparison airport rankings on the scheduled and on the actual networks, for ECAC2 dataset.A) Incoming Trip centrality; B) Outgoing Trip centrality; C) Incoming 2-legs Trip centrality; D) Outgoing 2-legs Trip centrality; E) Incoming TripRank; F) Outgoing TripRank. Red dots represent distressed airports (see text for definition of distress). Red lines are 1:1 lines.

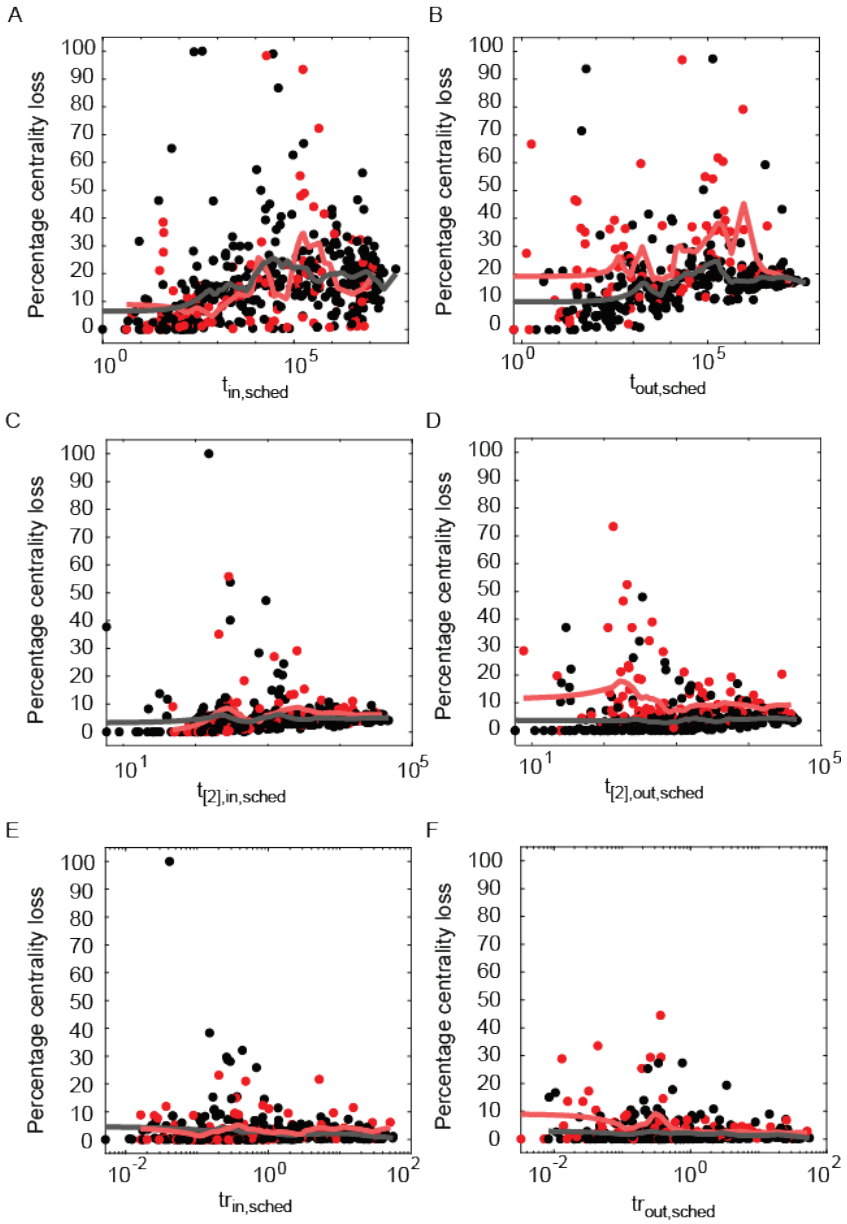

Figure 25. Centrality in the scheduled network versus percentage loss of centrality between the scheduled and the actual network, computed as Δc=(c_sched-c_real)/c_sched, for the ECAC2 dataset. A) Incoming Trip centrality; B) Outgoing Trip centrality; C) Incoming 2-legs Trip centrality; D) Outgoing 2-legs Trip centrality; E) Incoming TripRank; F) Outgoing TripRank. Red dots represent distressed airports (see text for definition of distress). Red and gray lines are obtained by a locally weighted smoothing (LOWESS) of, respectively, the red and black dots.

The same analysis is performed on each day of the US_April dataset. In this case, to relate centrality losses to the delay situation of each day, we characterize days by the average delay of flights on that day, by the average fraction of delayed flights in an airport on that day, and by the number of airports that were congested on that day. For each day, we compute the correlation coefficient between the ranking of airports on the scheduled network and that on the actual network according

to incoming and outgoing Trip centrality and TripRank, the average percentage centrality loss of all airports, that of distressed airports and that of non-distressed ones.

According to all metrics considered, the average percentage centrality loss has a clear tendency to increase with all three delay statistics considered, as shown in Figure 26, while the correlation coefficient tends to decrease, as shown in Figure 27. Figure 28, instead, shows separately the trends for distressed and non-distressed airports, confirming the observation made in the ECAC dataset that the loss of outgoing centrality is noticeably larger in distressed airports, while a small difference is seen for incoming centrality. This confirms that the delays of flights in an airport have little power in explaining its loss incoming centrality.

These results prove that the generalized metrics which we have introduced are able to tell apart different delay conditions, differently from the existing centrality metrics, and therefore they represent a suitable tool to compare the different scenarios simulated by the ABM in terms of disruptions caused by delays. Additionally, as explained above, the loss of centrality provides more specific information with respect to delay statistics, as it measures the real impact of delays in terms of missed connections, a central point to assess because it entails the highest costs for airlines and disutility for passengers.

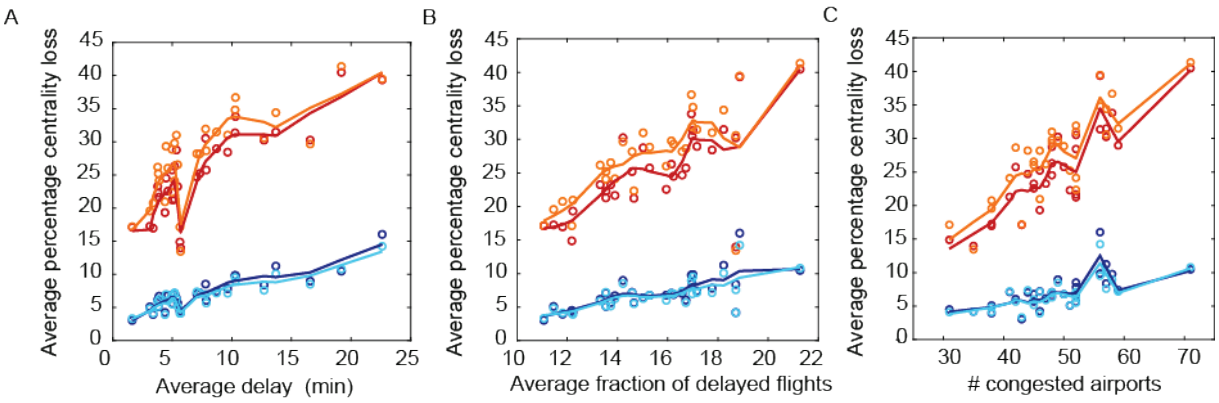

Figure 26. Average percentage centrality loss in the days of the US_April dataset, computed as Δc=(c_sched-c_real)/c_sched, according to incoming Trip centrality (red), outgoing Trip centrality (orange), incoming TripRank (blue) and outgoing TripRank (light blue), plotted against three delay statistics: A) average departure delay; B) average fraction of flights with departure delay in one airport; C) number of distressed airport (see text for definition of distress). Each point corresponds to one day of the dataset. Lines are obtained by a locally weighted smoothing (LOWESS) of the dots of the correspondent color.

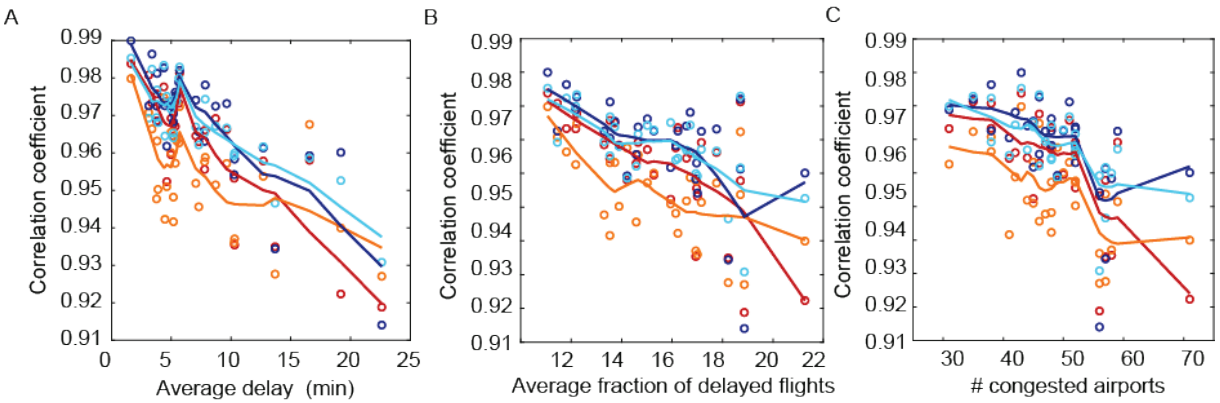

Figure 27. Correlation coefficient between the ranking of airports on the scheduled network and that on the actual network in the days of the US_April dataset according to incoming Trip centrality (red), outgoing Trip centrality (orange), incoming TripRank (blue) and outgoing TripRank (light blue), plotted against three delay statistics: A) average departure delay; B) average fraction of flights with departure delay in one airport; C) number of distressed airport (see text for definition of distress). Each point represents a day of the dataset. Lines are obtained by a locally weighted smoothing (LOWESS) of the dots of the correspondent color.

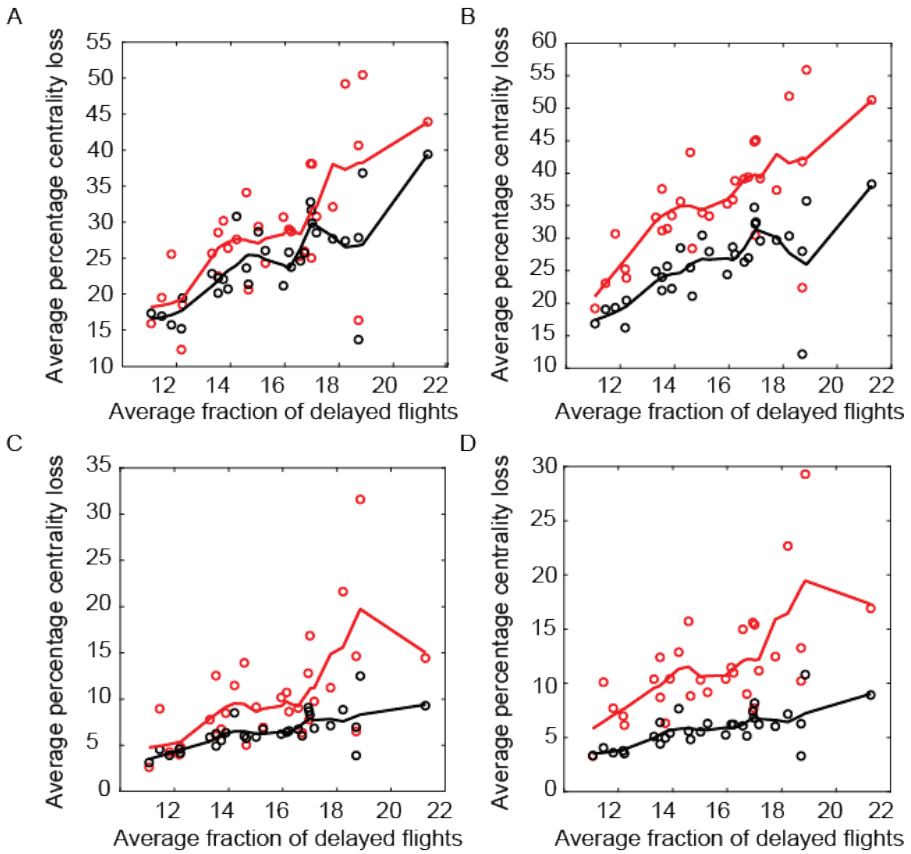

Figure 28. Average percentage centrality loss of airports for the days of the US_April dataset plotted against the average fraction of flights with departure delay in an airport on that day. Distressed airports (red dots) and non-distressed ones (black dots). A) Incoming Trip centrality; B) Outgoing Trip centrality; C) Incoming TripRank; D) Outgoing TripRank. Each point represents a day of the dataset. Lines are obtained by a locally weighted smoothing (LOWESS) of the dots of the correspondent color.

3.4 Causality metrics analysis

In this Section we consider an application of the causality metrics to the network of airports and flights, and in particular we study the problem of delay propagation from one airport to another. We quantify the state of delay of an airport as the average delay of flights taking off from that airport. If we consider the average over all flights, independently from the airlines operating them, we study the airport network at the maximum level of aggregation. However, the network has a multi-layer structure, where each layer contains the flights of a different airline. Hence, by defining an airline-specific state of delay, obtained by averaging only over its flights, we can assess the causality structure of delay propagation among different airlines at different airports. In the following, we consider both aggregated and multi-layers structures of the network of airports:

1. The causality network obtained from the detection of causal relationships between the time series of the state of delay of airports, where no distinction is made between different airlines (i.e., the state of delay is obtained as the average departure delay of all flights which

take off from the airport). In this network, we consider the aggregated information on flights and a direct link represents a directional channel for the propagation of delays;

2. The multi-layer causality network obtained from the detection of causal relationships between the airline-specific time series of the state of delay of airports. In this case, the network has as many layers as the number of airlines, and each airport is present in each layer. An airport is characterized by a different state of delay in each different layer, obtained averaging the departure delays of flights operated by the corresponding airline. Causality links can then be either intra-layer or inter-layers. Even if the process of delay propagation among airports cannot be decoupled per airline, studying the causality links within a single layer tends to capture how flights of a particular airline interact each other at different airports, thus revealing the channels of delay propagation within a single airline. On the other hand, the analysis of the inter-layer connections between two layers describes how the delayed flights of an airline affect the flights of another airline, thus quantifying the level of interaction between the two companies.

This application is considered for validation purposes and to give an example of how causality metrics can be successfully applied to the ATM system.

As an empirical case, we study the US dataset for the period from January 1st 2015 to March 31st 2015 and we consider the Granger causality tests, both 'in mean' and 'in tail', to build the causality network of airports. Specifically, the nodes of the causality network are the US airports and a directional link $i \rightarrow j$ represents a causal relationship, i.e., airport i 'Granger-causes' airport j. Each airport is described by its state of delay and observed states of delay at different times define the time series of the airport state, as described in the next section.

3.4.1 State of delay of the airport

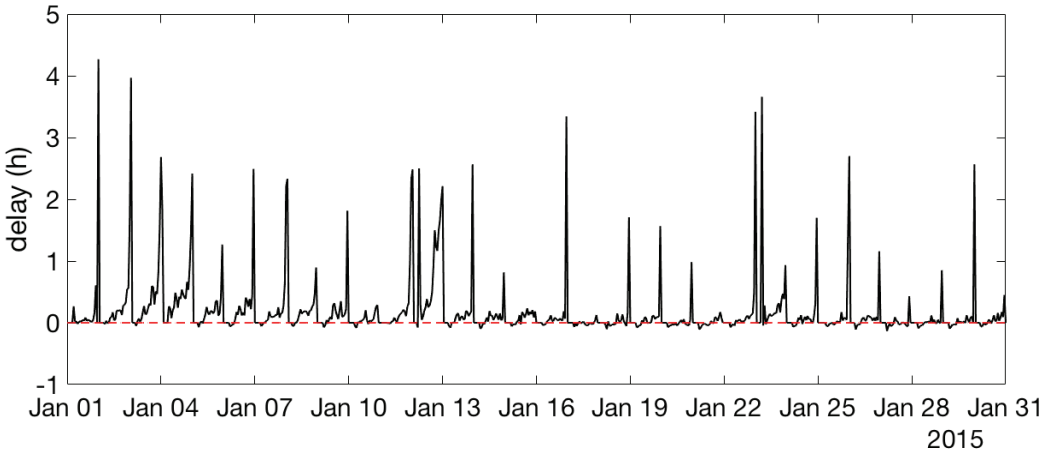

Figure 29. State of delay for Atlanta (ATL) airport in the month of January 2015.

The state of delay at time t for each airport is given by the average departure delay of the flights taking off from that airport during time window of one hour duration starting at t. An example of time series of states of delay for the US airport of Atlanta (ATL) is shown for the month of January 2015 in Figure 29. Note that the time series is characterised by peaks appearing with a daily

frequency. This is mainly due to the non-stationarity of delays, whose expected value is higher at the end of the afternoon, and lower during the first hours of the day. As suggested in [19], in order to reduce the non-stationarity of the time series caused by daily seasonality, which may result in a biased evaluation of the Granger causality metric, we can apply a Z-Score detrending procedure. The standardized time series of airport i is calculated as

$$\underline{x}_{i,t} = \frac{x_{i,t} - < x_{i,t} >}{{\sigma_i}^t}$$

where $< x_{i,t} >$ and ${\sigma_i}^t$ are the mean and the standard deviation of the delay states of airport i recorded at hour t across all available days. Thus, the resulting time series has zero mean by construction and it is standardised to be considered as input for the statistical test in order to prevent the detection of spurious causal relationships as a consequence of daily patterns.

Finally, especially for small airports, periods of the day with zero activity are observed frequently in the data when no flights take off from the airport. In this case, we define the state of delay of the airport as equal to zero. However, zeros may correlate because they occur in the same periods of the day, e.g., during night, and result in spurious causality detections. In order to quantify the importance of this effect, we compare the causality detections obtained on the original dataset with those obtained on a randomized dataset where each time series is shuffled keeping the positions of the zeros fixed. If a non-negligible number of causality relationships are detected in the randomized dataset, they are due to the zeros. We apply this procedure on the US dataset for the period from January 1st 2015 to March 31st 2015 (see below) and we conclude that the presence of zeros does not affect crucially the causality analysis. In fact, the link density[10] of the Granger causality network built with the original dataset is 0.045, whereas for the randomized data is 0.003. In conclusion, our choice of defining the state of delay as equal to zero when zero activity is observed at the airport in a given hour does not introduce spurious correlations.

3.4.2 Causality metrics: results

In this Section we present the results of the analysis of causal relationships in the US ATM system obtained applying the methods presented in Section 2.2.2 to the US_JM dataset, containing the days from the three months of January, February and March 2015. In the following, we first show the results of Granger causality networks, focused on the detection of channels of delay propagation in the whole ATM system by using aggregated information on flights, then we present some insights into the interdependence structure among airlines which use the same airspace, obtained by looking at the multi-layer causality networks.

3.4.2.1 Granger causality networks for the US ATM system

The Granger causality in mean network is created by applying the pairwise Granger causality tests to the detrended time series which describe the states of delay of the N=315 airports in the period from January 1st 2015 to March 31st 2015. For each pair of time series, we assume a vector autoregressive model VAR(p) to describe their evolution. The Granger causality test is performed for different values of the parameter *p*, ranging from 1 to 6. The maximum lag is chosen equal to 6

[10] Link density is defined as the ratio between the number of observed links and the number of all possible links, i.e., M=N(N-1) in the case of directed networks without self-loops.

because the empirical partial autocorrelation function becomes statistically zero after the sixth lag for the time series of any airport. The partial autocorrelation function of a time series identifies the appropriate lag of an autoregressive model, since the partial autocorrelation of an AR(p) process is zero for lag *p+1* and greater. Then, in case of rejection of ${H_0}^{mean}$, the best *p* is selected according to the Bayesian Information Criterion. Best *p* values are distributed mainly around 1 and 2. Since the observed states of delay are built within time windows of one hour duration, these small values of *p* highlight that delay propagation captured by the Granger causality in mean test occurs for short timescales. The distribution of inferred *p* is shown in Figure 30.

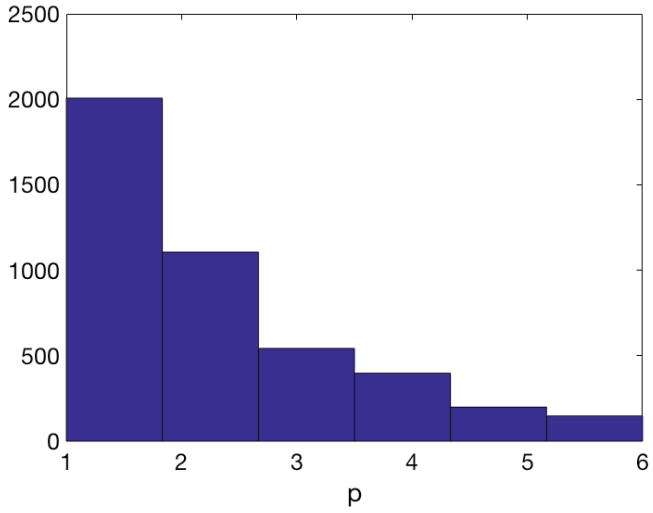

Figure 30. Histogram of inferred p for the VAR(p) model in the Granger causality in mean test applied to the US dataset from January 1st 2015 to March 31st 2015.

The Granger causality in tail network is created by applying the Granger causality in tail tests to the same time series. Similarly to Granger causality in mean, an autoregressive process AR(p) is assumed to describe the time evolution of the state of delay with p=6. Thus, the forecasting of density distribution uses the *p* past steps and extreme events are defined as the observations which fall on the right tail above the 95th percentile[11]. Hence, the memory of the autoregressive process describing the state of delay of an airport is six hours, as before. Then, we move from the description in terms of states of delay to the definition of the state of distress of an airport, namely the binary variable which takes value equal to one if an extreme event for the delay occurs, zero otherwise.

However, differently from Granger causality in mean, the Granger causality in tail test is performed by checking for non-zero correlations for the states of congestion of airports at any lag, because of the definition of the null hypothesis to be tested. This means that delay propagation for extreme events is checked at any time scale, from the shortest to the largest. Describing the process of propagation of distress at any time scale may be useful in capturing effects of two- (or more) legs, especially in an ATM system like the US where some flight have very long durations.

[11] Once the density distribution is estimated, the *q*-quantile of the probability distribution specifies the value of the random variable, i.e., the state of delay, such that the probability of the variable is less than or equal to *q*. Thus, the 95th percentile is the quantile associated with a probability equal to 0.95.

For both the in mean and in tail cases, we set the significance level for the overall tests as equal to $\xi = 5\%$, and, as a consequence, the significance level of each test is $\xi' = \frac{0.05}{N(N-1)}$ where N=315 is the number of airports and N(N-1) represents the number of all possible pairs of airports tested for causality. It is important to comment on the role of the multiple hypothesis correction in the construction of the Granger causality network. To the best of our knowledge multiple hypothesis test correction has not been applied before to Granger networks in ATM, but it has a large impact on the inferred causality networks. For example, in the case of Granger causality (in mean) network built for the US dataset for the period from January 1st 2015 to March 31st 2015, the link density for the Bonferroni corrected network is 0.05, whereas without the correction we obtain 0.45, a much larger value. Therefore, neglecting to introduce a correction means considering a large number of non-significant causal links.

The two Granger causality networks built with the US traffic data from January 1st 2015 to March 31st 2015 differ greatly. The Jaccard index is a statistic used for comparing the similarity of two sample sets, in this case the two graphs described by their own adjacency matrix. It is defined as the size of the intersection divided by the size of the union, i.e., the number of links in common divided by the total number of links, and varies between zero and one. In this case, the Jaccard index is 0.12, a very small value suggesting that the two networks considerably differ. In the following, we consider several network statistics which explain this difference and describe what kinds of phenomena the two Granger causality tests tend to capture.

A network system can be studied according the usual methods adopted in complex network analysis to highlight the main characteristics and to obtain a global description of the system. The standard topological network metrics considered here (and defined below) are: link density, diameter, average path length, clustering coefficient, reciprocity, one particular kind of motifs, namely feedback triplets, and (PageRank) centrality of nodes. These standard statistics for the two causality networks are reported in Table 16.

Table 16. Network statistics for the Granger causality networks compared with both Erdos-Renyi and fitness models.

	Granger in mean	Erdos-Renyi (w.r.t. GC in mean)	Fitness model (w.r.t. GC in mean)	Granger in tail	Erdos-Renyi for (w.r.t. GC in tail)	Fitness model (w.r.t. GC in tail)
Link density	0.045	0.045 ±0.001	0.045 ±0.001	0.152	0.152 ±0.001	0.152 ±0.001
Diameter	8	4.0 ±0.1	5.4 ±0.5	5	3.0 ±0.1	4.0 ±0.1
Average path length	3.07	2.46 ±0.01	2.47 ±0.02	1.95	1.84 ±0.01	1.90 ±0.01
Clustering coefficient	0.27	0.077 ±0.001	0.280 ±0.006	0.26	0.161 ±0.001	0.251 ±0.008
Feedback triplets	14 856	908 ±46	7 656 ±352	71 127	3 631 ±871	64 136 ±1415
Reciprocity	0.20	0.022 ±0.002	0.094 ±0.004	0.14	0.076 ±0.002	0.112 ±0.002

We compare these statistics with the corresponding values obtained for two random graph models: (i) the Erdos-Renyi model which describes a random graph with the same number of links of the causality network but uniform link probability for each couple of nodes and (ii) the fitness model [34]. The latter model describes an ensemble of random graphs with the property that the average (over the ensemble) in- or out-degree of a node is equal to the one of the real data. The node's in-(out-) fitness represents the likelihood that the node has an incoming (outgoing) link and it is correlated with the in (out) degree. These models represent a method alternative to randomization algorithms and have the advantage of a better control on the network statistics we aim to preserve, i.e., link density for the former and degree sequence for the latter.

The obtained Granger causality in mean network has L=4 401 Granger causal links, i.e., link density equal to 0.045, while for the Granger causality in tail we observe L=15 027, i.e., link density equal to 0.152. Since the two statistical tests are applied to the same data with the same confidence level, this result suggests that the airport system is more interconnected by the extreme (delay) events than by the average (delay) behavior. This is because the Granger causality in mean test weights equally the forecasting errors of both extreme and average events. The latter may represent random fluctuations which have no impact on delay propagation. In fact, small delays at the origin airport can be easily absorbed in the en-route phase, having no influence on the state of delay of the destination airport. On the contrary, it is unlikely that rare events corresponding to large delays have no common causes and interdependencies. Hence, selecting information about extreme events tends to keep random fluctuations out of the analysis. Moreover, our result shows that extreme events yield

predictive signals on the occurrence of extreme events in other airports, much more strongly than when considering average events.

The diameters of the Granger networks, i.e., the longest path connecting two nodes, are equal to 8 and 5, respectively for in mean and in tail cases, while for the corresponding Erdos-Renyi networks are 4 and 3, thus suggesting the presence of outlying nodes less connected with the central core. This is (partially) confirmed by the average path length, equal to 3.05 for the Granger in mean network and 1.95 for the Granger in tail network. These values are slightly larger than 2.46 and 1.84 when compared with the Erdos-Renyi case. However, the difference becomes smaller in the case of the fitness model, thus highlighting that the average path length can be explained in terms of degree distributions of the nodes and the existence of few peripheral nodes may be responsible of the larger diameter.

The clustering coefficient of a graph is a measure of the likelihood that nodes cluster together, specifically it is the number of closed triangles, i.e., subgraphs of three nodes connected each other by links having any direction, divided by the number of potential triangles. For the causality networks, the clustering coefficient is much larger than the corresponding Erdos-Renyi networks, a difference explained by the different degree of nodes. In fact, the fitness model has a clustering coefficient in line with the Granger networks. Among all possible triplets, the clockwise or anticlockwise triangles are of great importance in the case of the ATM system. Indeed, they represent potentially unstable feedback subsystems for delay propagation. When we restrict to the feedback triplets, we can notice that the Granger networks are characterized by much larger values than the corresponding random cases. Hence, in the case of ATM systems, an interesting clustering measure is the one which considers feedback and any innovation which aims to increase the robustness and resilience of the system, should reduce it.

Finally, reciprocity is a measure of the likelihood that nodes in a directed network are mutually linked, i.e., the ratio of the number of links pointing in both directions to the total number of links. In ATM systems, we expect that reciprocity in the Granger network is higher than in the random case, since return trips, common in the flight network, are expected to induce a mutual causality link. The results confirm this behaviour, especially for Granger causality in mean. Mutual causality links are more often detected for short flights because there are likely more return trips during the day. This is confirmed in Figure 31, where we show the fraction of mutual links in the Granger causality networks classified according to the mean duration of flights connecting the two node-airports incident to the link.

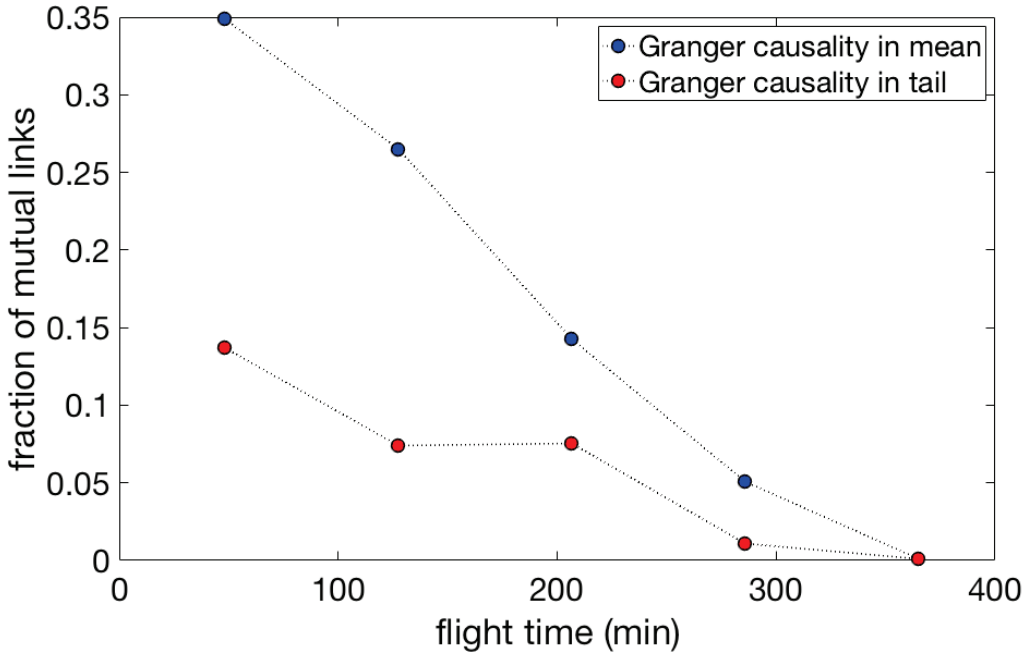

Figure 31. Fraction of mutual links in the Granger causality networks, both in mean (blue dots) and in tail (red dots), as a function of the flight duration between the two node-airports incident to the link in the Granger network. At each bin, each point is the number of detected mutual links normalized by the number of airports whose mean flight distance falls inside the time window. Each bin is identified in the figure by the value corresponding to its upper bound.

To summarize the conclusions of our causality analysis, two specific features of the causality network emerged: the overexpression of feedback triangles and of mutual linkages with respect to the randomized graphs. The interpretation of this result is that delay tends to propagate in both directions between a couple of airports and in triangle loops. Both these patterns tend to amplify delays in the system, and therefore innovations should aim at their reduction. Hence, we conclude that these two metrics, i.e., feedback triangles and mutual linkages in the Granger causality networks, are two important metrics to assess the impact of innovations on the system's performance.

For completeness, in Figure 32, we show the degree distribution of both Granger causality networks. Networks shows no evidence in favour of scale-free distribution for the degree according to the test for power-law distributions [35].

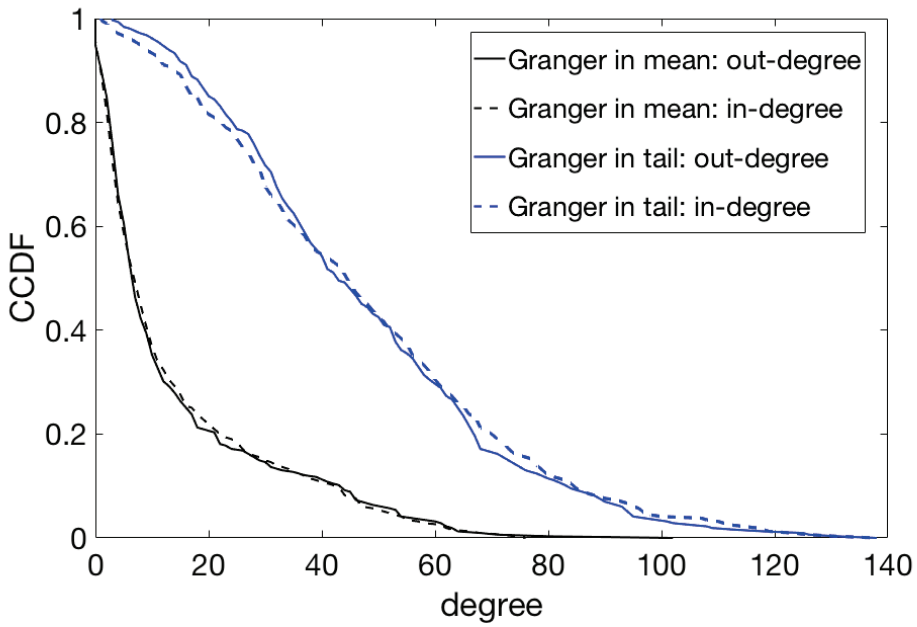

Figure 32. Complementary cumulative distribution function for the degree of the Granger causality networks built with the US dataset from January 1st 2015 to March 31st 2015, both for outgoing and incoming links.

We now compare the Granger causality networks with the physical networks of airports and flights used to build them. We investigate whether the properties of the former networks can be explained by some characteristics of the latter, for example whether high degree airports in the causality networks are typically the major hubs or the smaller/regional airports (as suggested in the empirical analysis of the Chinese airspace in [19]).

We find a positive Kendall rank correlation (0.47 for out-degree and 0.46 for in-degree, see the left panel of Figure 33) between airport traffic size, measured as the average number of flights per day, and node degree in the Granger causality in mean network, meaning that airports having many flights tend to have also many causal relations, both incoming and outgoing. For Granger causality in tail, instead, the Kendall rank correlation is much smaller (0.14 for out-degree and 0.16 for in-degree), see the right panel in Figure 33, i.e., the number of causal relationships is much less correlated to the traffic size. Therefore, small and medium airports may be more central for the process of propagation of extreme delay events and information on their states more important for the prediction of congestions in the system.

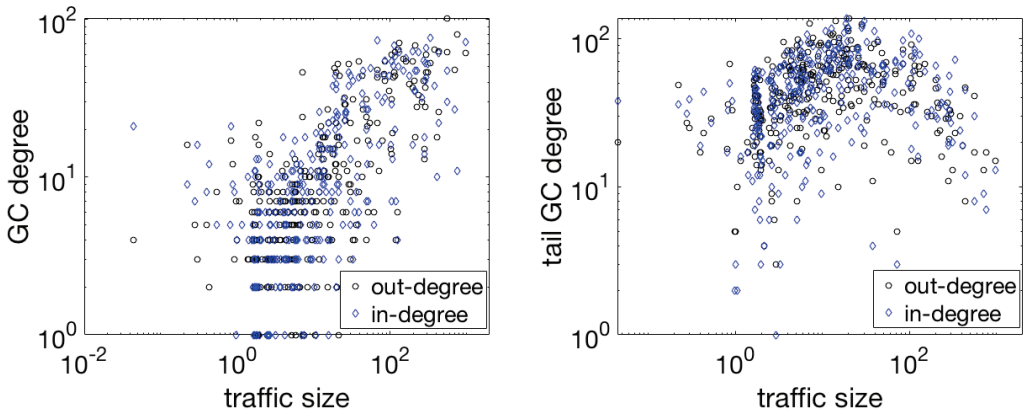

Figure 33. Node degree of both the Granger causality as a function of the traffic size of airports, in mean (left) and in tail (right), i.e., the average number of flights per day. Black dots represent the out-degree of nodes while blue dots the in-degree.

A more specific tool to compare the Granger causality network with the network of airports and flights is degree overlap. Let G and A be, respectively, the adjacency matrices of the two networks. *A* has an entry equal to one if there exists at least one flight connecting the two corresponding airports in the entire period considered, zero otherwise. The degree overlap is defined as

$$o_{out} = \frac{\Sigma_j G_{ij} A_{ij}}{\Sigma_j G_{ij}}, \qquad o_{in} = \frac{\Sigma_i G_{ij} A_{ij}}{\Sigma_i G_{ij}}$$

for the out-degree and in-degree, respectively. The degree overlap between the Granger causality network and the aggregated network of airports and flights measures how often two airports that are linked in the causality network are also linked in the other. A large degree overlap, therefore, means that a causality link between two airports is often present when the two airports are linked by direct flights. On the contrary, a small degree overlap means that causal relationships are often present even in the absence of a direct flight. Hence, the degree overlap can be interpreted as an indirect measure of the fraction of one-leg effects as channels of delay propagation.

It is interesting to note, in Figure 34, that the degree overlap increases with the airport size. In both Granger in mean and Granger in tail cases, we find a positive rank (Kendall) correlation, 0.62 and 0.56 respectively. In particular, note that the overlap for large airports is very high and tends to be close to one for increasing sizes. This is a signal that the primary channels of delay propagation for large airports are the one-leg effects, i.e., direct flights connecting the airport to other nodes of the network. On the contrary, especially for the Granger causality in tail, the overlap for small and medium airports is small or very close to zero, suggesting that the mechanisms of delay propagation are represented by two or more legs effects. In other words, a channel of delay propagation from a

small airport to another node of the network occurs by means of two or more flights which create a path connecting them by involving other airports in between[12].

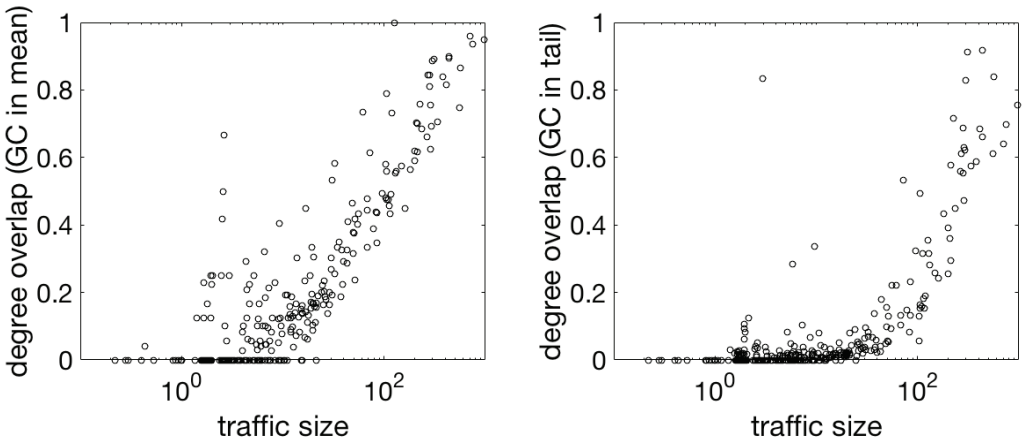

Figure 34. Degree overlap between the Granger causality networks, for both causality in mean (left) and in tail (right), and the network of airports and flights described by the adjacency matrix having enty equal to one if there exist flights connecting two airports, zero otherwise. Each point represents the degree overlap averaged over the out-degree and the in-degree of the node

The centrality of a node in a Granger causality network is a measure of its importance in the process of delay propagation. Here, we adopt the standard PageRank centrality measure to classify the airports. A graphical visualization of the result is shown in Figure 35 for both Granger networks. PageRank centrality reveals a clear split between the two macro geographical regions of the US, i.e., East and West. Figure 35 shows the ranking of nodes according to PageRank centrality for the Granger causality network. The geographical disequilibrium is related to the fact that flights depart earlier (in the EST reference frame) in the East with respect to the West, thus it is more likely that a delay starts propagating in the system from the East, making the eastern airports more central in the Granger network. Table 17 shows the top ten US airports according to PageRank centrality in both cases.

[12] In principle, other exogenous sources may be responsible for the presence of a causal relationship, e.g. weather may create a correlation between the states of delay of two airports that are geographically close. Thus, a dependence between two states of delay might also not be due to flights.

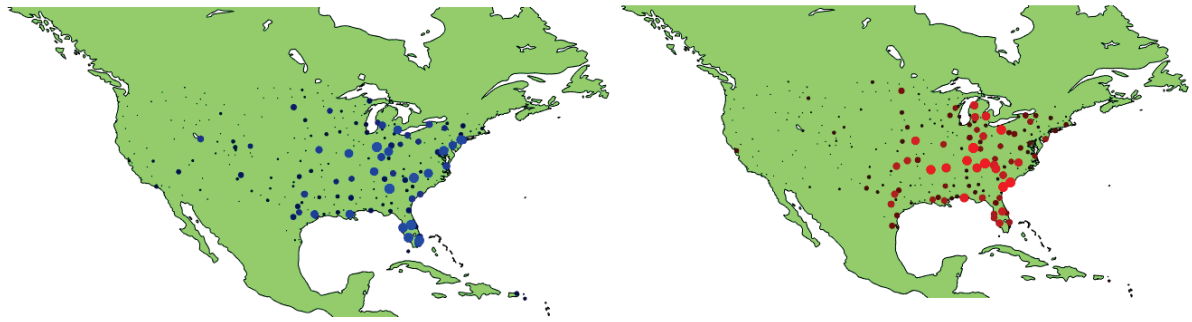

Figure 35. PageRank node centrality for the Granger causality in mean network (blue dots) and the Granger causality in tail network (red dots). Increasing dot size and brighter colour represent a rank increase.

Table 17. Top-ten of US airports according to PageRank centrality for Granger causality networks, both in mean and in tail. We show the average number of flights per day as measure of the airport's traffic size.

PageRank: Granger causality in mean network	Page Rank: Granger causality in tail network
MCO (Orlando) - traffic size = 340	SDF (Louisville) - traffic size = 32
ATL (Atlanta) - traffic size = 996	TYS (Knoxville) - traffic size = 20
IND (Indianapolis) - traffic size =71	CHS (Charleston) - traffic size = 31
CLT (Charlotte) - traffic size = 300	SAV (Savannah) - traffic size = 21
FLL (Miami) - traffic size = 241	CAK (Akron-Canton) - traffic size = 20
RSW (Southwest Florida) - traffic size = 114	BNA (Nashville) - traffic size = 136
TPA (Tampa) - traffic size = 203	LIT (Little Rock) - traffic size = 31
BWI (Washington) - traffic size = 232	MEM (Memphis) - traffic size = 41
EWR (New Jersey) - traffic size = 293	CHA (Chattanooga) - traffic size = 12
RDU (Raleigh-Durham) - traffic size = 90	PNS (Pensacola) - traffic size = 21

We remark that the more central airports in the first case are characterized by high traffic, whereas, in the second case, the top-ten ranking is formed by small and medium airports. This is supported by the values of linear correlation between airport sizes and PageRank centralities: 0.63 for Granger causality in mean and 0.14, a much smaller value, for Granger causality in tail. The results obtained here for the US ATM system by Granger causality in mean analysis contrast with those in [19] for the Chinese ATM system. In fact, here we find that larger airports tend to have higher degree in the Granger in mean network (Figure 33), and larger degree overlap (Figure 34), resulting in higher

centrality, while in [19] small airports have the largest degrees, indicating that delays are mostly propagated from small and regional airports. According to our results, large airports are more informative regarding the prediction of the state of delay of the whole system and more central for the process of delay propagation. However, when the focus is on extreme events, i.e., the states of distress of airports, information on small and medium airports is central for predictions and the characteristics of the corresponding causality network are not strongly correlated with traffic.

Finally, we want to stress the fact that Granger causality networks help to capture the dependence structure of delay propagation, which is related to how the ATM system performed in a given period, in particular when macroscopic variables, e.g., total traffic, do not change significantly. Hence, we repeat the pairwise Granger causality in mean analysis for a time window of one month, starting from January, and rolling the window week-by-week, up to the end of March, see Figure 36. The result suggests that link density, i.e., a measure of how much interconnected the system is, does not depend trivially on both the total traffic and the active air routes connecting airports. In fact, these global variables are quite constant in the considered time windows. For example, we observe the largest number of links (January) when traffic is smaller than its maximum (March), thus suggesting a complex dynamics of delay propagation. This analysis reveals that the dependence structure of causal relationships arises in a non-trivial manner from the dynamics of delay propagation. This quantitative argument supports the relevance of causality metrics in analyzing the outcomes of both the toy model and the full ABM model, where we aim to assess the performance of the modelled ATM system as a function of the Domino mechanisms once all the other conditions have been fixed.

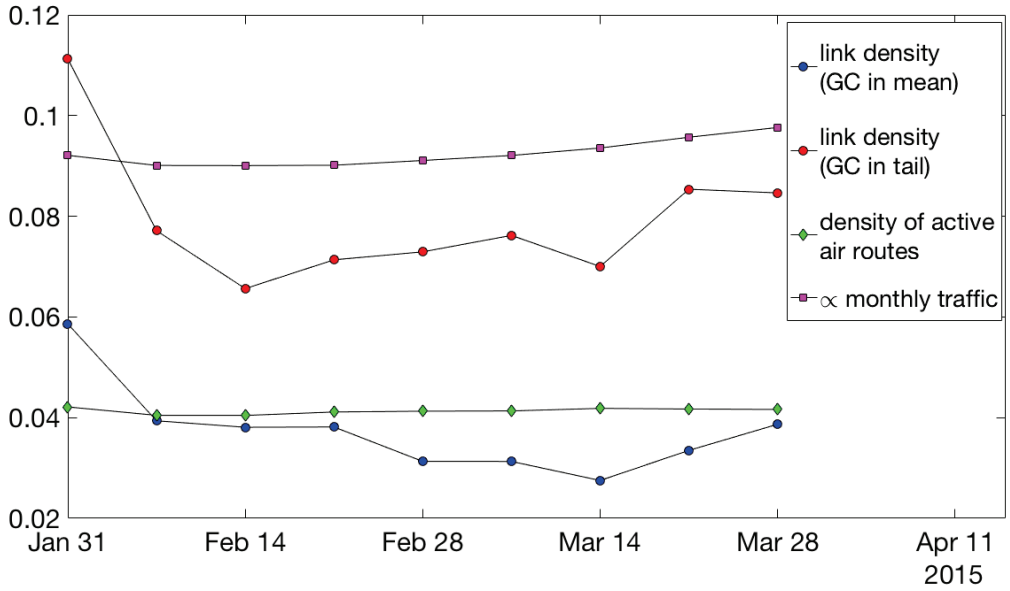

Figure 36. Link density of the Granger causality networks, both in mean (blue dots) and in tail (red dots),for different 30-days periods (indicated by the last day) and compared with traffic (purple dots) measured as the total number of flights within the considered 30 days (rescaled by a factor 4.5×10^6) and with density of active air routes in the aggregated network of airports and flights (green dots).

3.4.2.2 Multi-layers causality networks for the US ATM system

The main goal in Domino is to characterize how different subsystems interact and to quantify the dependence structure of the interactions, and especially to understand whether they become stronger or weaker with the implementation of ATM innovations. In order to study the dependence structure among airlines, considered as interacting subsystems, the causality analysis presented before can be generalized to multi-layers Granger causality networks, where each layer identifies an airline (or alliance). In each layer, nodes represent the airports served by that airline. Different copies of the same airport are present in different layers, if that airport is served by several airlines. Then, each node-airport on each layer is described by its state of delay obtained by averaging over the departure delays of flights operated by the corresponding airline. A directional link between any two nodes of the multi-layer Granger causality network, either in mean or in tail, represents the rejection of the null hypothesis of no causality from one node to the other, as before. Finally, the single-layer causality network is described by a unipartite graph, with a square adjacency matrix, as in the aggregated network considered before. However, the inter-layers links from one layer to another determine a bipartite graph whose adjacency matrix is rectangular because, in general, there is a different number of nodes in each layer. Thus, depending on the direction of links, we obtain two different bipartite causality networks for each pair of layers, where, on a case-by-case basis, one airline functions as the 'causer', and *vice versa*.

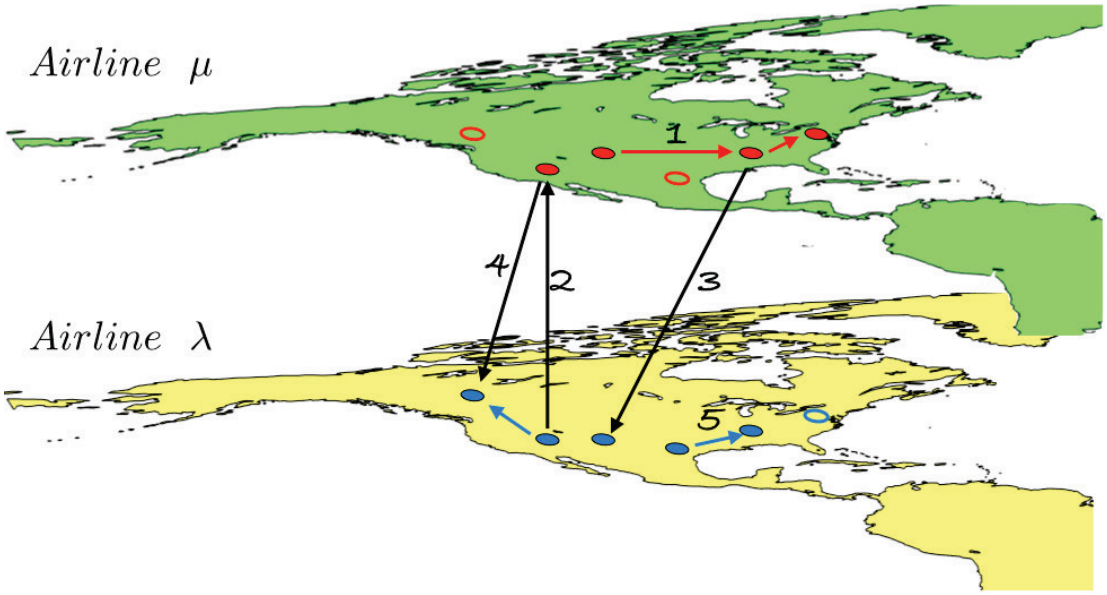

Figure 37. Scheme of interactions in the multi-layers Granger causality network. Red and blue nodes in the green and yellow layers, respectively, represent airports served by two different airlines. Empty circles represent airports served by one airline but not by the other one. Red links represent causal relationships from airline μ to itself (intra-layer links, e.g., link 1), blue links represent causal relationships from airline λ to itself (intra-layer links, e.g., link 5), and black links represent causal inter-layer links between the two airlines, including the case of different airports (e.g., links 3 and 4) and copies of the same airport in different layers (e.g., link 2).

The kind of causal interaction depends on whether the link lies on a single layer or connects two different layers. In Figure 37 we show all possible interactions in the multi-layers Granger causality networks, i.e.

1. A link lying on a single layer captures a causal relationship of delay propagation between two airports served by the airline: see links 1 and 5 in Figure 37;
2. A link between two copies of the same airport on different layers describes a local causality interaction between two airlines, meaning that the delays of one airline at that airport cause delays in the flights of the other airline at the same airport. This might be due, e.g, to non-rotational effects of reactionary delays at the airport[13]: see link 2 in Figure 37;
3. An inter-layer link connecting an airport on the first layer to another airport, on the second layer, which is served by the first airline, represents a channel of delay propagation mediated by flights of the first airline: see link 3 in Figure 37;
4. An inter-layer link pointing to an airport not served by the first airline represents a second-order effect of delay propagation, which can be mediated by two- or more- legs effects because no flights of that airline may propagate delays in that airport: see link 4 in Figure 37.

[13] The primary delay of a flight tends to create reactionary delays for other flights departing from the same airport.

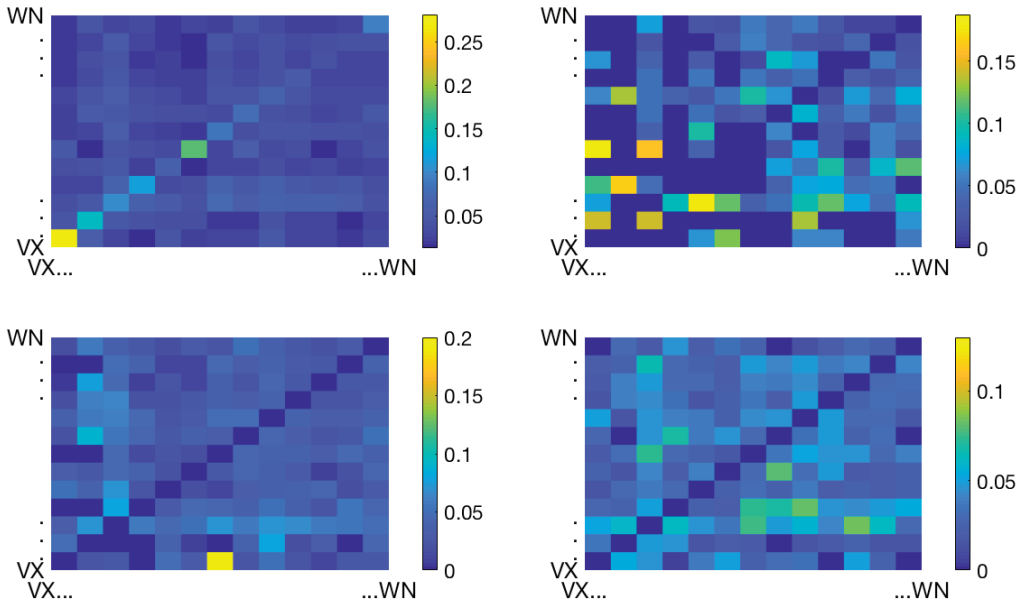

Figure 38. Link density in the multi-layers Granger Causality in mean network by considering all kinds of causal links (top left) or distinguishing according to interactions of type 2 (top right), type 3 (bottom left) and type 4 (bottom right). In the top left panel, diagonal elements represent link density in each layer, whereas off-diagonal elements describe density of inter-layers links.

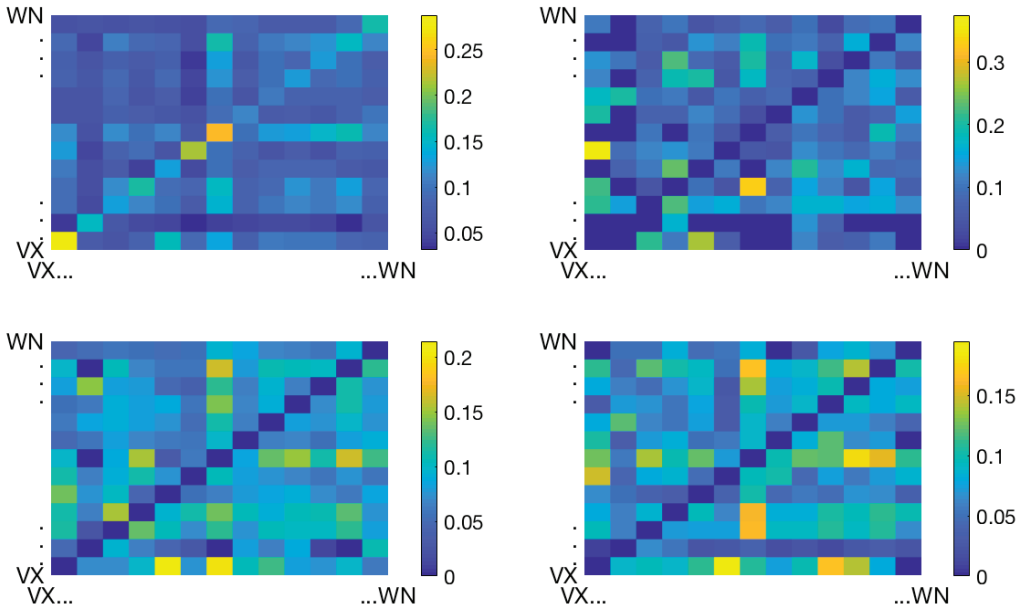

Figure 39. Link density in the multi-layers Granger Causality in tail network by considering all kinds of causal links (top left) or distinguishing according to interactions of type 2 (top right), type 3 (bottom left) and type 4 (bottom right). In the top left panel, diagonal elements represent link density in each layer, whereas off-diagonal elements describe density of inter-layers links.

We apply the multi-layers causality analysis to the US_JM dataset, containing information on flights of the 14 airlines. As preliminary result, we show in Figure 38 and Figure 39 the link densities of the subgraphs obtained according to the four types of interactions described above for both Granger causality in mean (Figure 38) and in tail (Figure 39). In each panel of the figures, airlines are ordered according to ascending size, measured as the total number of flights, thus the smallest airline (Virgin America (VX)) is found at the bottom left of the panel while the largest (Southwest Airlines (WN)) at the top right. The color of each pixel represents the density of causal links where the airline on the x-axis is the 'causer' and the one on the y-axis is caused. For instance, in the bottom left panel of Figure 38 the yellow pixel is the density of causal links in the bipartite graph which describes how Envoy Air (MQ) 'causes' Virgin America (VX). Finally, in both figures: (i) the top left panel shows link densities for both single-layer causality networks (on the diagonal) and bipartite causality networks (off-diagonal), by averaging over all three types of inter-layers interactions; (ii) the top right panel shows link density for the interaction of type 2, i.e., the number of links between copies of the same airports divided by the number of airports in common between the two airlines; (iii) the bottom left panel shows link density for the two-layers causality network by accounting only for interactions of type 3, i.e., causal links among different airports but in common to the two airlines[14], while (iv) the

[14] This is a subgraph whose square adjacency matrix has dimension equal to the number *n* of airports in common to the two airlines. Thus, link density is the ratio between the number of causal links and the number of all possible links, i.e., *n(n-1)*.

bottom right panel displays the case of second-order effects (interactions of type 4), i.e., when the 'causer' has no flights at the airport in the other layer[15].

According to Granger causality in mean, we notice that each airline tends to interact more with itself than with other airlines, see top left panel of Figure 38. This behaviour is more evident for small size airlines. Even if at the moment we do not have a definite explanation between the size of the airline and the strength of the causal interactions, we conjecture that this can be due to the fact that in an airline with few flights during the day, reactionary delays associated with the same aircraft connecting different airports might have a larger impact on the states of delay of those airports. Moreover, the organization of the flights in a hub-and-spoke or point-to-point scheme might explain this correlation. Further analyses are needed to better elucidate this point.

However, in the bottom right region of the top left panel of Figure 38, we can notice that medium and large airlines influence the small companies slightly more than the average level of causality. By looking at the other panels of Figure 38, this effect can be explained in terms of both local interactions at the same airport (interactions of type 2) and second-order effects (interactions of type 4), thus delay propagation occurs mainly because airlines share the same resources and/or the high traffic, due to the large number of flights, impacts the system by means of two-legs effects as well.

When we consider delay propagation of extreme events by means of Granger causality in tail test, we note that the results of the multi-layers analysis display an average level of causality larger than the previous case (in agreement with the results for the aggregated network). However, distribution of causal links is more heterogeneous, and no clear patterns appear in terms of airlines' size, see Figure 39. Further investigations and a better characterization of multi-layers Granger causality networks are left for future outlooks.

3.5 Conclusions on metrics

In this Section we proposed several new network metrics and methods to assess centrality and causality in the ATM system. We plan to use these metrics (together with basic, classical measures) to compare simulations of the ATM system obtained from the Domino ABM under different innovation scenarios. Summarizing:

Considering centrality

- The difference between scheduled and actual centrality using Trip Centrality and TripRank gives an indication of the loss of centrality of airports and of flights. When aggregated across the different elements of the system (e.g., airports), the difference measures the loss of performance of the whole ATM. When considering single airports or flights, this study could highlight the effect of innovation implementation on the airlines/airports adopting them and on those sharing the same resources (airports, airspace, etc.).

[15] This is a bipartite subgraph whose rectangular adjacency matrix has dimension $q \times p$. Thus, link density is the ratio between the number of causal links and the number of all possible links, i.e., $q \times p$.

- In the case of the simulations of the model with 4D Trajectory Adjustment[16], we could verify whether actual and scheduled centrality of the airline implementing it (i.e., of its flights and of the copy of the airports where it is active) become more similar. Moreover, the change in centrality of airlines sharing the same resources of the one implementing DCI could give indications of positive or negative externality of the innovation.

Considering causality:

- The density of the Granger causality network is a measure of interconnection and tightness of the system (or of a part of it). Therefore, it is interesting to investigate whether the introduction of innovations modifies the causality network making it less dense and/or more modular.
- Similarly, the number of reciprocated causal links and feedback triangles is a simple measure of causal feedbacks in the system, and its monitoring under different innovations can provide insights into their effects on the stability of the ATM system.
- The multilayer causality network, and specifically the density and level of reciprocity of its components, give indications of the tightness of its elements and their mutual interconnection.

Whilst in this Section we have considered the empirical application of the proposed metrics to the network of flights and airports, they can be applied much more generally, especially in other parts of ATM that can be mapped by a network or monitored with multivariate time series. Domino's ABM will provide many different types of data (often not available for real systems) that we will analyze with the considered metrics. For example, we could consider as nodes the DMAN and AMAN at the different airports and their queue size as the variables describing their state. Finally, we plan to extend the existing metrics by weighting the links by considering (i) the number of passengers traveling in the flight represented by the link or in the considered walk and (ii) the cost of delay to study how the network structure affects the propagation of cost and to identify the most central nodes (flights or airports) in terms of airline cost.

[16] See Section 4 for an explanation of how this mechanism is implemented in the toy model.

4 Toy model

This Section of the deliverable is dedicated to the detailed description of a simple, yet controllable model of the interactions between airlines, the agents of the model, and other elements of the ATM system, such as Departure and Arrival managers and passengers. This model, referred to as the 'toy model', is much simpler with respect to the ABM model, as it contains less details on the processes and considers only one type of agent, the airline. While it is certainly less realistic than the ABM model, the small number of parameters and of mechanisms in the toy model makes it easier to interpret its results. Therefore, the outcomes of the toy model in different scenarios, analysed by means of the metrics presented in the previous sections, will provide us with a first understanding of what changes each may innovation bring and what are the mechanisms that cause these changes. Additionally, this model, which takes less computing effort to simulate with respect to the ABM, will be used to explore the effect of local implementations of innovation as well as of disturbances (e.g., delays) to the system.

As in the ABM case, we model the pre-tactical and tactical phase. Each airline takes decisions on its flight based on a cost function, possibly exploiting the innovations implemented in Domino. In Section 4.1 we present the general modeling framework, while in Section 4.2 we describe a baseline implementation of the model. Finally, in Section 4.3 we show some preliminary results of this baseline version of the toy model for validation purposes.

4.1 Toy model design

4.1.1 Introduction

We model a day of ATM operations for the ECAC airspace as a chain of events, each event representing a flight, which take place on the network of airports and air routes connecting the airports. The airline decision making process is only modeled for flights departing from ECAC airports. However, flights departing outside the ECAC space but landing in an ECAC airport are considered in the slot assignment, in order to correctly account for the amount of traffic at each airport. Both commercial and non-commercial flights are taken into account to reproduce as closely as possible the occupation of the airports' slots. However, particular attention is reserved to commercial flights, with a detailed study of the cost function of airline companies.

In the implementation of the toy model, we consider three levels in the description of mechanisms that the Domino project aims to model:

- level 0 aims to represent current operations;
- level 1 implies some further capabilities from technological and operational improvements;

- level 2 is the most advanced case which might require further research and it is, by definition, more prospective.

For a detailed description of the three levels of implementation of the Domino mechanisms see Deliverable 3.1 (Architecture Definition). In the context of the toy model, in particular, we aim to model 4D trajectory adjustment (limited to tactical dynamic cost indexing and wait for passengers for simplicity) and flight prioritisation mechanisms.

The baseline level of the toy model aims to describe the current state of the ATM system, with no (further) innovation, and interactions among airlines arise because of the competition for the same resources, specifically, tactical slots. The implementation of the Domino mechanisms at both level 1 and level 2 will give rise to a strong interdependence due to the increasing complexity of interactions (see Section 4.1.6).

Since mechanisms studied in Domino work in the tactical phase, the strategic phase is assumed to be given as initial input for the toy model. Hence, we describe how the ATM system evolves during the day of operations.

The subsections are organised as follows: (i) we introduce the considered subsystems and who are the agents, i.e., the airlines, of the toy model; (ii) we describe briefly the strategic phase that represents the input of the toy model; (iii) we describe the framework and the chain of events; (iv) we describe how airlines take decisions in terms of cost functions and (v) how Domino innovations work at the three different levels of implementation; (vi) finally, we present the scenarios we will analyse and the proposed case studies for investigation.

4.1.2 Agents and subsystems

Domino is interested in the macro-effects arising in the ATM system during the tactical phase from the interactions of the agents active in the different subsystems.

Agents in the toy model represent airlines which take decisions about their own flights during the tactical phase. According to the implemented scenario, the optimisation problem to be solved may regard a single flight or, in the case of connections, more than one. This last aspect is relevant, in particular, when applying 4D Trajectory adjustments and in particular dynamic cost indexing at the network level. In the decisional process, airlines interact with other subsystems, such as the DMAN, AMAN, and passengers.

In the toy model, each aircraft is identified by its tail number, and each flight by a flight number. Flights are also characterized by their departure and arrival times and by their departure and arrival airports. Decisions regarding a flight come from a complex process involving the flight operation centre (FOC), i.e., the airline, which supplies different information and directives, and the possible use of dynamic cost indexing (DCI), i.e., adjusting the cost index based on updated delay and cost information. As a consequence, according to the considered scenario decisions can be more or less scaled up to the network level, and rich in detail.

An airport is a node of the studied network and is identified by its ICAO code. An airport is a complex entity which includes terminal(s), departure (DMAN) and arrival (AMAN) managers and runways. In the context of the toy model, we reduce all these entities to a single unit, i.e., the airport, which interacts with flights according to pre-ordered schemes and rules. We assume that an airport has

time slots for departures and for arrivals, which represent the airport's resources used by airlines for the take-off and the landing of flights.

Other relevant subsystems are the Network Manager (NM) and the flight prioritisation processor, if implemented. In the real world, each flight interacts with NM to get the ATFM slot if required, by providing a flight plan (including departure time, route and possibly a trajectory). In the presence of the flight prioritisation processor, airlines can further request the swap of the departure of two flights, in order to recover delay for the flight with higher priority (c.f. UDPP). Thus, the NM and flight prioritisation processor handle the request and accept or reject it after some checks about airspace traffic, sector capacities and security. However, the modelling of sectors and 4D air routes is out of the scope of the toy model. Hence, the Network Manager is taken into account by using information about the last accepted plan (m1 files) of flights in the pre-tactical phase, while the flight prioritisation processor is modelled as further rules that work as constraints for the optimisation problem faced by airlines in allocating their own flights during the tactical phase.

Passengers are passively aggregated subsystems which take flights and move from one flight to another in the case of connecting flights. Passengers are relevant to establish the costs due to missed connections, *inter alia*. Consequently, they have to be considered by the airlines in the decision process during the tactical phase.

4.1.3 Strategic phase: input of the toy model

The Domino model aims to assess several characteristics of the ATM system during the tactical phase, such as the propagation of delay and the cost of the elements in the system, the identification of critical nodes in terms of delay and cost propagation, and to quantify what is the impact of innovations on the considered characteristics. The toy model is inspired by the same aim, but in a simpler and controllable framework. Then, in both cases, the strategic phase, such as flight planning, number and itineraries of passengers, are assumed to be known and this information is taken from data. More specifically, an example of input data for the toy model is summarised in Table 18 and Table 19.

Table 18. Example of input information for toy model implementation (1)

Flight number	Tail number	Airline	Origin	Destination	Scheduled departure	Scheduled arrival
BAW605	GEUXM	BAW	LIRP	EGLL	12-Sep-2014 11:15:00	12-Sep-2014 13:30:00

Table 19. Example of input information for toy model implementation (2)

Last off block time (m1 file)	Last landing time (m1 file)	Taxi-out	Taxi-in	Number of passengers	Inbound connecting passengers	Outbound connecting passengers
12-Sep-2014 11:17:00	12-Sep-2014 13:28:00	10 min	10 min	104	(0,0,…0)	(0,0,2,3,…,0)

Scheduled times (the time shown on the ticket of passengers) describe the strategic phase planned by airlines months before the day of operations. Scheduled times will be used to estimate the cost of compensation and care of passengers in case of delay. We are also interested in the last-filed flight plan, contained in the m1 files, as they determine the slots initially assigned to each flight at the departure airport. Finally, the planned landing time in the last-filed flight plan is used for simulating the time elapsed from the departure to the landing (see Baseline implementation Section).

Information about the number and the fraction of passengers on connecting flights are of fundamental importance for the implementation of dynamic cost indexing and Flight Prioritization in order to generalise the cost function by accounting for real time-lag effects arising from connections. This information is contained, for each flight, in the last two columns of Table 19. They have the form of two vectors, where entry *i* is the number of passengers connecting from or to the present flight and flight *i*.

The number of available time slots at an airport is estimated by considering the maximum capacity of the airport by looking at the traffic data for the day of operations. For instance, if we observe one departing flight per minute in the moment of highest traffic, we set the time slot for the departure equal to 1 minute for that airport. In the case of small airports, we fix the lower bound for the duration of the time slot as equal to 15 minutes. This is a simplifying assumption which makes the toy model more tractable and, at the same time, it permits us to reproduce properly what we observe in reality.

Finally, we also use taxi-in and taxi-out durations for each airport. These durations, contained in the dataset, are used to properly assess the different kinds of delay costs, e.g., the en-route delay has great impact on the cost of fuel, whereas the delay at-gate concerns mostly passengers and crew costs.

4.1.4 Chain of events

Each flight is associated with an event, represented by an 18-tuple

$$(f, \tau, a, i, j, {t_d}^0, {t_d}^1, r, \delta_g, P, \pi_i, \pi_o, t_d, v, {t_l}^0, {t_l}^1, t_l, x)$$

where:

1. flight number f
2. tail number τ

3. airline ICAO code a
4. origin airport i
5. destination airport j
6. scheduled departure time ${t_d}^0$
7. off block time ${t_d}^1$ of the last-filed flight plan
8. reactionary delay r
9. delays at-gate δ_g, different from the reactionary delay (e.g., administrative delay, airport congestion, etc.)
10. number of passengers P
11. a vector variable π_i describing the number of incoming passengers from connecting flights
12. a vector variable π_o describing the number of outgoing passengers to other connecting flights
13. departing time t_d
14. average (cruise) speed $v$from origin to destination
15. scheduled arrival time ${t_l}^0$
16. landing time ${t_l}^1$of the last-filed flight plan
17. arrival time t_l
18. cancellation index x.

Some entries of the 18-tuple are fixed by the input data described before, specifically both flight and tail numbers, the airline ICAO code, both origin and destination airports, passenger information and, more importantly, both scheduled departure and arrival times and last-filed flight plan with departing and landing times.

We differentiate reactionary delay from other possible delays at gate:

1. reactionary delay is caused by the late arrival of aircraft from a previous journey. It occurs when the arrival delay cannot be recovered by the buffer planned in the strategic phase. In the toy model, we define the reactionary delay as the difference between the actual off-block time and the scheduled departure time. At that time, the airline will request a departing slot starting from the moment the aircraft will be ready;
2. $\delta_g$accounts for other possible delays at-gate, such as aircraft technical delay and state of congestion of the airport (e.g., as a consequence of weather). Airlines can consider UDPP innovation to recover $\delta_g$for flights with high priority by requesting the swapping of flights. We do not consider delays at-gate other than reactionary in the baseline version of the toy model, while we aim to consider stress test scenarios with non-zero δ_g in future developments.

The variables of the toy model which might be affected by the agents' decisions are the departure time, the speed of flight, and the arrival time:

1. airlines take decisions about the departure time of their own flights by trying to minimise their own cost function (see below). Late departure time of a flight, e.g., because of wait-for-passengers strategy, is determined depending on airport traffic capacity and time slots availability;

2. the process of making decisions concerns also the average speed of flights, e.g., because airlines could decide to recover en-route some delay. Here, we consider the average cruise speed from origin to destination, without modelling explicitly both climb and descent phases;
3. arrival time is a stochastic variable which is the sum of the departing time and the flight time. The modelling of the flight time from origin to destination is explained below in the Baseline implementation Section.

The cancellation index is a binary variable describing whether the flight is cancelled.

Once we have defined a single event as described above, we can form the chain of events which describes the day of operations. The main constraint in forming the chain is the consistency in the succession of events, i.e., we can admit only time sequences which respect aircraft itineraries. For this purpose, the simplest option is to sort the events according to scheduled departure times, adjusted according to the observed reactionary delays. In the process of requesting time slots to NM and DMAN, at each run of the model we locally randomise the chain within short time windows. This randomisation takes place over disjoint time windows shorter than the minimum flight time, preventing in this way impossible occurrences, e.g., an aircraft which departs before arriving at the airport.

In the running of the toy model, we simulate events following their order in the chain. When simulating an event, departure time and speed of the flight are obtained by solving an optimization problem which accounts for the cost of delay, detailed in the following section. In absence of any delay previously assigned to the corresponding flight and when no wait-for-passengers strategy is applied, the departure time coincides with the scheduled one. When the departure time is delayed, the solution of the optimization problem must take in consideration the availability of time slots at the departure airport. Then, the arrival time is obtained by summing to the departure time the flight time, simulated as explained in Section 4.2.2. If the flight has an arrival delay which prevents the aircraft from being on time for the next departure, a reactionary delay is assigned to the next flight operated by that aircraft. Hence, the next event of the chain is considered.

4.1.5 Cost functions for airlines

Delays cause airlines to incur (high) costs. Although delays cause costs to other subsystems, such as airport authorities, air traffic control and, not least, passengers, the focus here is on airlines because the interest is on the process of decision-making in the tactical allocation of flights. The cost of delay can be differentiated in strategic delays (those accounted for in advance in the strategic phase) and tactical delays (those incurred on the day of operations and not accounted for in advance). Airlines mitigate the occurrence of the strategic costs by adding buffers to the flight schedules. However, in the toy model the strategic phase is an input, thus, we consider only the impact of tactical delays. In assessing these costs, we refer mainly to the report in [3] where authors have estimated on historical data the different components which sum to the total cost of delay.

Tactical costs of delay can be differentiated in four contributions: at-gate (engine off), taxi, en-route, and arrival management. The cost of fuel mainly affects the last three phases. In the context of the toy model, we model a composite gate-to-gate delay, without accounting separately for the four different contributions. We separate primary costs from reactionary costs.

Primary costs

The different primary components which sum to the total cost of primary delays are:

1. the cost of fuel is related to the fuel consumption which depends on the mean speed. The flight (or the airline FOC) may decide to speed up in order to recover some delay. Hence, we consider the excess cost of fuel arising from the choice of increasing the speed of the flight. By assuming a possible range of values for the speed, ranging from the velocity characterized by the lowest cost to the maximum velocity permitted by physical constraints[17], the excess cost of fuel can be described as
2. $C_{excess-fuel} = \varphi T_{ij}(v-1)$
    a. where φ is a normalisation parameter, T_{ij} is the expected flight time and the unit value represents the adopted mean speed by the typical aircraft. Based on EUROCONTROL's 'Base of Aircraft Data' ('BADA') [36]. In [37] it is shown that a maximal (positive) speed variance is equal to 7.5% (in the following, the curves of cost are obtained by considering the mean speed for each aircraft); Note that when the Cost Index is changed, not only the cruising speed is modified but the whole trajectory, particularly the climb and descent phases which can account for significant amount of delay recovery. However, for simplicity, in the toy model, only the flight speed is considered.
3. maintenance costs incurred by delayed aircraft relate to factors such as the (mechanical) attrition of aircraft waiting at gates;
4. crew costs are based on the cost of crewing for additional minutes over and above those planned at the strategic phase;
5. passenger costs are the costs related to any possible consequence of delay regarding passengers. Costs of passenger delay may be classified as either a 'hard' or 'soft' cost (see [3] for further details):
    - hard costs arise from passenger rebooking, compensation and care;
    - soft costs represent the cost paid in the future by an airline as a results of passenger dissatisfaction; these kinds of costs are difficult to estimate because of unknown utility function of passengers, although they have been estimated from the literature and surveys.

For a single flight in a normal day of operations, the cost function for primary delays, both at-gate and en-route, aggregated for all considered components is reported in Figure 40 where the data are taken from [3].

[17] In principle, the theoretical range of values of possible speeds is characterized by a left bound smaller than the observed one. In practice, indeed, very low values of speed are avoided, due to negligible differences in fuel consumption for non-negligible flight time increases [41], [42].

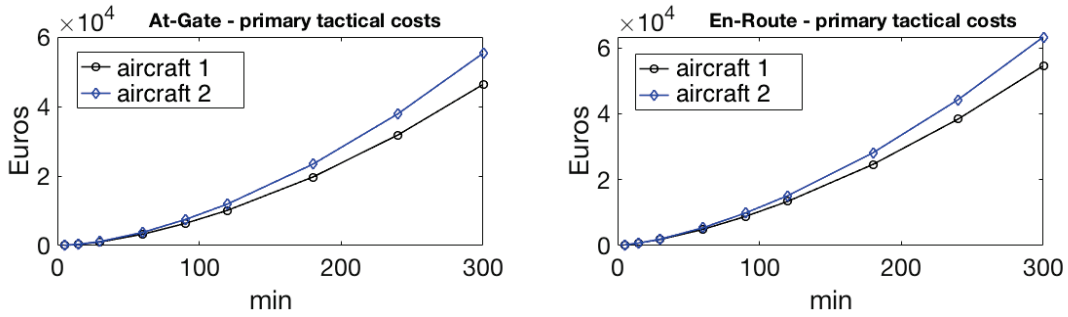

Figure 40. Costs of primary delay for both delays at-gate (left) and delays en-route (right), by considering two commonly used twin engines, mid-range twin aircraft 1 (black curve) and mid-range twin aircraft 2 (blue curve).

Black and blue curves represent the cost function for two commonly used twin engines. In this case, as suggested in [3], the curves can be fitted by a power function to make calculations more tractable. The smoothed cost profiles describe the cost functions adopted in the implementation at level 0 of the toy model.

Reactionary costs

Tactical costs discussed before described the cost of the delay of a single flight, assuming that such delay has no impact on other flights. However, on the day of operations, primary delays caused by one aircraft cause 'knock-on' effects in the rest of the network, known as 'secondary' or 'reactionary' effects. Primary delays do not only affect the initially delayed aircraft on subsequent legs, i.e., rotational reactionary effect, but also other aircraft, i.e., non-rotational reactionary effect. Hence, primary tactical costs need to be scaled up to the network level.

First, the 'knock-on' effects have impact on the cost of subsequent legs because of maintenance and crew costs. Since the average impact of these components is known per single flight (see left panel of Figure 40), reactionary costs can be scaled up to the network level once: (i) we have estimated the expected reactionary delays, considering both rotational and non-rotational effects, on subsequent flights[18]; (ii) we have complete information on the network of the scheduled flights of each airline[19].

Second and more importantly, the 'knock-on' effects have impact on passenger costs because of compensation, care and rebooking of delayed passengers. The cost of rebooking is considered as reactionary because it follows from a missed connection caused by a primary delay.

[18] See, e.g., [3] where the curve for the reactionary delay as a function of the primary delay is obtained by statistical methods applied to historical observations.

[19] See Data Section.

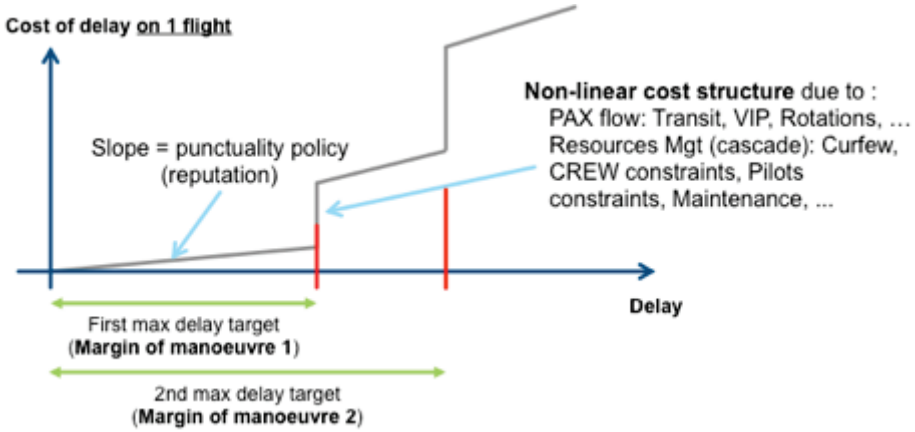

Figure 41. A sketch of the cost model to be implemented in the toy model.

Figure 41 shows the cost model we will use in the implementation of the dynamic cost indexing. The cost structure of a flight is typically not linear due to the presence of different milestones and time constraints for each flight, such as crew out-of-hours constraints, maintenance slot requirements, passenger missed-connection costs, or a missed airport curfew, etc. For instance, when arrival delay is larger than 180 minutes passengers are entitled to compensation, which results in a jump in the cost profile at 180 minutes [38].

In the context of the toy model, we focus on the estimation of passenger cost of reactionary delay at two different levels, representing level 1 and level 2 of dynamic cost indexing: (i) a heuristic estimation which is not specific for the single flight and its connections; (ii) a precise flight-specific estimation which takes into account real-time network effects of connecting flights as well as the current number of passengers. More specifically,

- at level 1, the cost profile is calibrated on historical observations of the cost impact of reactionary delays. Specifically, it does not model explicitly passengers' itineraries and connections, but it considers the average number of passengers per flight and itinerary. The presence of jumps in the cost function refers to margins of manoeuvre which are known by airlines, e.g., information on buffers introduced in the strategic phase to preserve connections and delay threshold above which airlines face costs of compensation;
- at level 2, we aim to describe the cost model in Figure 42Figure 41 by accounting explicitly for passengers' itineraries as well as the current states of connecting flights, e.g., by considering the presence of delays during the day of operations. As in level 1, the cost of compensation is introduced as the presence of a jump at the delay threshold for compensations, but it is calibrated over the current number of passengers on the aircraft. We model the reactionary costs of missed connections as the sum of rebooking, compensation and care of passengers ('hard' components), expressed by the following equation

$$C^{f}_{connections} = \sum_{b} (\gamma_i{}^{f}\ \pi_i{}^{bf} I_{t_l{}^{b}+\Delta_j - t_d{}^{f}>0} + \gamma_o{}^{f}\ \pi_o{}^{fb} I_{t_l{}^{f}+\Delta_k - t_d{}^{b}>0})$$

where the first term accounts for incoming connecting passengers and the second term for the outgoing ones. Specifically, $\gamma_i{}^{f}$ and $\gamma_o{}^{f}$ represent respectively the cost of missed connections per passenger for inbound and outbound connections $\pi_i{}^{bf}$ and $\pi_o{}^{fb}$ are

respectively the number of inbound and outbound connecting passengers for the flight f, $I_{(..)}$ is the indicator function which takes value equal to one in the case of missed connection, zero otherwise. We consider a simple rule for identifying missed connections, namely when the arrival time of the incoming flight, plus the time Δ_j necessary to move passengers from one aircraft to the other one at airport j, is later than the departure time of the outgoing flight. The sum is over all flights connected to flight f, considering both inbound and outbound connecting flights. The advantage of the explicit description of 'hard' reactionary costs of passengers in connecting flights is that we evaluate when the wait-for-passengers strategy is appropriate for airlines by considering the network effects in real time and not only by means of historical evaluations.

4.1.6 Domino mechanisms

In the toy model, we consider two Domino mechanisms for implementation and assessment of their impact on the ATM system, namely 4D trajectory adjustment and flight prioritisation. Regarding 4D trajectory adjustment, we model the delays management strategies from the airline point of view by including dynamic cost indexing and hub management (e.g., waiting for connecting passengers). flight prioritisation refers instead to centralised decisions by airlines in reordering and swapping their own flights in order to minimise their expected costs. Both mechanisms are implemented at the three levels considered in Domino. For further information, see Deliverable D3.1 [39].

In the toy model, we specify the three different levels of implementation as follows. For the 4D trajectory adjustment mechanism:

- **level 0** tries to replicate the current practices used by airlines to manage their own flights. This consists in adopting the standard Cost Index (CI), which takes into account both the cost of delay and the cost of fuel deriving from the excess expense of using more fuel to increase the speed of the flight. We use the primary cost function (which does not include reactionary costs (Figure 40);

- **level 1** implementation considers the use of dynamic cost indexing for flights, but with a cost estimation based on heuristics. In this case we consider a cost model which includes reactionary costs (estimated statistically) and implement a naive wait-for-passengers strategy which may be adopted by airlines to reduce the impact of reactionary costs;

- **level 2** implementation considers an advanced estimation of expected costs due to delayed flights by including the network effects of passengers in connecting flights. In particular, the cost function is scaled up to the network level as in level 1, but the estimation of costs takes into account the current number of passengers as well as the current state of the system. Wait-for-passengers strategy is thus implemented with detailed information about the connecting flights.

Regarding the flight prioritisation mechanism:

- **level 0** does not consider any input of prioritisation from the airlines. Slots are assigned to flights following a First Planned First Served (FPFS) protocol;

- **level 1** implementation considers the application of UDPP principles, i.e., when there is a mismatch between capacity and demand, UDPP could be set, allowing airlines to reorder and protect their delayed flights. We consider two protocols in the implementation of UDPP, i.e., FDR (Fleet Delay Reordering) which allows the swapping of slots in order to reduce the delay of flights with high priority, and SFP (Selective Flight Protection) which suspends the planned departure of a flight with low priority to push the other flights one slot forward in the queue. The implementation of UDPP principles requires centralised decisions by the airline FOC;

- **level 2** extends level 1 with the possibility to exchange slots between airlines. In this case, a credit system needs to be implemented and the cost function for airlines needs to be generalised to account for the benefit of exchanging slots with other airlines. However, it is out of scope of Domino to implement a market-like platform for slot exchange. The toy model aims to explore the potential consequences of possibility of exchanging slots between airlines.

4.1.7 Decision-making process

Depending on the level of implementation, airlines face a different optimization problem in the process of decisions making about their own flights during the tactical phase. In the simplest setting we do not take into account current information about the network and the decision process only concerns the single flight. For instance, let us assume a cost model as in Figure 41 but with historically based margins of manoeuvre. When a flight is delayed at departure with a delay slightly larger than the first max delay target (see Figure 41), the airline could decide to increase the speed of flight to recover en-route the delay difference (w.r.t. the target), not paying for passenger compensation at the expense of an extra cost of fuel. In the context of the toy model, this represents a possible event at level 1 of implementation of dynamic cost indexing.

When we introduce explicit estimations of costs based on current information about the network effects, we move to describing airline decisions as a centralised process which consider more than one flight at once. Explicit modelling of costs, in particular regarding the passenger cost of delay for the airline, leads to a more complex decision problem, especially in the case of airline hubs where passengers and flights traffic is more concentrated. However, informed decision-making scaled up to the network level with explicit information about passengers and flights could reduce considerably the costs faced by airlines. In mathematical terms, the solution of the *global* minimisation problem for airline decisions is always better than finding *local minima* for the decision variables. In practice, if one incoming flight has been delayed, the airline might decide to actively delay outbound flights to wait for inbound connecting passengers. In some cases (depending on explicit estimation of costs), e.g., for the last flights of the day where passengers missing their connection need to be rebooked on next day flights, leading to significant costs of care, the wait-for-passengers strategy may reduce the passengers cost for the airline.

On a case-by-case basis, we aim to consider the different levels of implementation for the cost function, ranging from level 0 to level 2. Furthermore, we aim to introduce also the UDPP principles in the decision-making process, allowing airlines to re-order their own flights. Thus, the considered optimization problem for airline decisions describes, from time to time, the Domino mechanisms of which we aim to assess the macro effects on the ATM system.

4.1.8 Scenarios and case studies

Several research questions arise when we consider implementation at both level 1 and level 2 of the innovations studied in Domino. Among others, we aim to answer the following:

1. implementation of 4D trajectory adjustment through dynamic cost indexing and other advanced estimations of the impact of delays require the creation of complex structures to support the decision process of airlines. In reality, this process of innovation is usually very expensive. Hence, at least at the beginning, only the largest companies will implement. Then, it is important to assess the impact of the 4D trajectory adjustment mechanism only when implemented by a subset of airlines, in order to evaluate their performances and those of the other airlines. Finally, this scenario will be compared to the one when all airlines implement the innovation;
2. the most important distinction among airports refers to the traffic size. Congestion events concern mainly the major airports where a large number of flights compete for the same resources, and slot allocation enhanced by the flight prioritisation mechanism aim to reduce, among other aspects, the probability of congestion. Hence, it is important to assess the impact of this innovation on the whole system by considering from time to time a different number of airports where the innovation is implemented. By ordering the airports according to traffic sizes and implementing the innovation starting from the major airports, we expect that there should exist a threshold for the number of airports above which any index of system performance would not change by increasing further this number. This should select those airports where the flight prioritisation mechanism would have an effective impact on the whole system.

4.2 Baseline implementation

In this Section, we detail the baseline implementation of the toy model representing level 0 of innovation, as described above. For this purpose, we need to specify: (i) the decision-making process faced by airlines, (ii) how simulations of flight times are performed, (iii) what happens to the chain of events when a primary delay prevents the aircraft to be on time for the next flight, and finally (iv) model calibration.

4.2.1 Optimization problem for airline decisions

In making decisions, airlines try to minimise their own cost functions. In the baseline toy model, we assume decentralised decisions by airlines, meaning that decision on one flight is independent from the presence of other flights, either arriving at or departing from the airport. Furthermore, we do not consider any wait-for-passenger strategy. Hence, the flight-specific cost function in the process of decision making is

$$C_f(t_d, v) = C_{delay}(t_d) + C_{excess-fuel}(v)$$

where the first component is the cost of primary delay as described by the cost profile in the left plot in Figure 40, while the second component is the excess cost of fuel which depends on the speed of flight v. The cost of delay depends on the departure time t_d of flight and the flight delay is computed with respect to the scheduled departure time. In Figure 40 we have shown the cost functions for two

commonly used twin engines. However, for the sake of simplicity, we assume that all airlines use the same type of aircraft. The cost curve for this typical aircraft is thus obtained by averaging the cost functions of Figure 40 at each time point. Then, we smooth the cost curve by using polynomial kernels in order to achieve flexibility in computing the cost of delay for a generic delay.

Hence, the problem for the airline is minimising the cost function of flight f

$$min_{(t_d,v)}\ C_f(t_d, v) \quad s.t. \quad (v-1)^2 < 7.5\%, (t_d - {t_d}^0) < 300\ minutes$$

where some constraints are imposed, specifically the variance of speed has to be smaller than 7.5% according to [37] and we assume a maximum delay of 300 minutes, otherwise the flight is cancelled.

In the context of Flight Management System (FMS), the Cost Index (CI) is defined as the ratio between the cost of delay and the cost of fuel and it is used by airlines to optimize the aircraft's trajectory including their speed. The solution of the described optimization problem defines the optimal value for CI, thus linking the adopted approach to the standard practice in FMS.

At the baseline level of implementation, the decision process takes place before the departure of flights, without considering dynamic decisions about speed during the en-route phase, as the cost impact of reactionary delays. Hence, the cost of delay refers to the primary delay at-gate.

4.2.2 Implementation details

The simulation of the flight times in the toy model is stochastic. The expected gate-to-gate duration is taken from m1 files. The simulated gate-to-gate duration is obtained by summing a stochastic flight time estimated by looking at the actual flight time in m3 files and a term that depends on the availability of an arrival slot.

To compute the first term, we proceed as follows. Since the variance for the actual gate-to-gate duration depends on the duration of the specific flight trajectory, we consider the coefficient of variation of the gate-to-gate duration, i.e., the ratio between the standard deviation and the mean of the gate-to-gate durations.

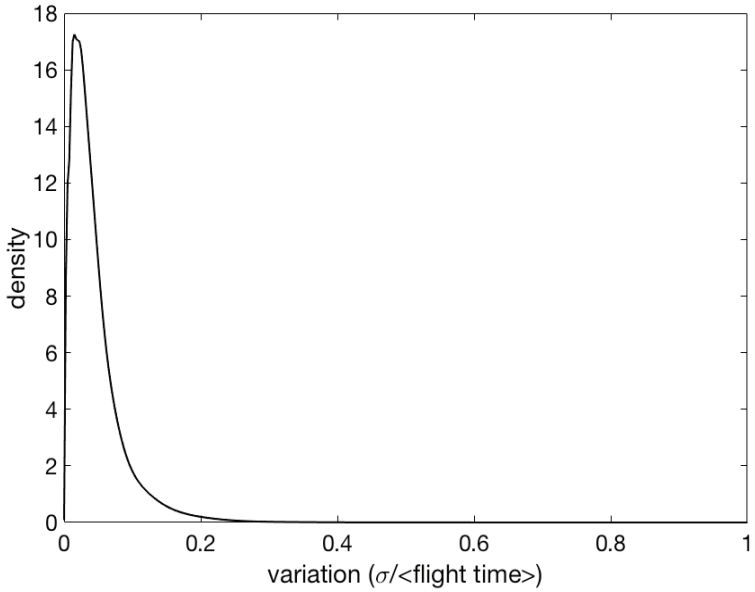

Figure 42. Distribution of the coefficient of variation for observed flight times (m3 files) at the day of operations September 12nd 2014.

In Figure 42, we show the empirical distribution of the coefficient of variation by considering all flights crossing the European ECAC airspace during 12 September 2014 (see Data Section below). Since this distribution is very peaked and the correlation between observed flight times and the corresponding variations is statistically equal to zero, we assume the coefficient of variation as independent from the distance between the two airports. Thus, the first term of the gate-to-gate delay is obtained by sampling the coefficient of variation for that flight from the empirical distribution. Then a normally distributed random variable with standard deviation equal to the sampled coefficient times the expected flight time is sampled. Finally, the second term, corresponding to the arrival phase is added.

In the case of a primary delay for a flight which prevents the aircraft to be on time for the next flight, the previously booked departure slot is cancelled, and the airline has to take new decisions about the departing time according to the available departure slots. In the baseline implementation, the new departure time is the solution of the optimisation problem which minimises the cost function for the flight but considering available time slots for departures.

In the simulation of the model, reactionary delays are determined by using the tail number of flights. The chain of flights for an aircraft is obtained by concatenating the events with the same tail number and reactionary delays are determined step by step by accounting for arrival delays and the turnaround time at the airport.

The turnaround time is assumed to be equal to 30 minutes for each airport. This homogeneity assumption is taken to reduce the complexity of the toy model. The maximum departure delay of a flight is considered equal to 300 minutes, otherwise the flight is cancelled in simulations. In the case of cancellation, all subsequent flights with the same tail number are cancelled.

Regarding the costs faced by airlines, we use the information present in [3] and the cost functions are obtained as explained before. At the baseline level, the challenging issue refers to calibrate the parameter φ of the cost of fuel. At the present moment, this is out of the scope of this Deliverable. In fact, the implementation of the baseline (level 0) does not involve the parameter φ. Indeed, it can be proved that the solution of the optimization problem for level 0 has the unit value as optimal choice for the speed of the flight, whatever the value of φ is. This is due to the fact that the function describing the cost of delay is convex, see the left plot of Figure 40. Hence, the calibration of φtogether with the parameters involved in the explicit estimation of passenger cost of delay are left for future implementations of the model.

4.2.3 Data

For the implementation and calibration of the toy model, we use two datasets describing air traffic and passengers' itineraries for 12 September 2014, namely EUROCONTROL's DDR_1409 which contains all flight information on the day of operations, and pax_itineraries which is based on previous datasets developed by the University of Westminster and allows us to align passenger information with flight data and, more importantly for the current implementation of the model, it contains the scheduled departure and arrival times. The available information is summarized in Table 20.

Table 20. Information available for each flight by crossing information of DDR_14_09 and pax_itineraries databases.

Field	What it contains
Date	Day, month, year of the scheduled departure time
Flight number	Unique flight id
Tail number	Registration number of the aircraft
Origin airport	ICAO code of origin airport from DDR
Destination airport	ICAO code of destination airport from DDR
taxi-out	average duration of the taxi-out phase for origin airport
taxi-in	average estimated duration of the taxi-in phase for the destination airport
Scheduled departure time	the expected departure time on the tickets of passengers
Actual departure time	Actual off block time from m3
Actual take-off time	Actual take-off time (Actual departure time + taxi-out)
Expected departure time	Expected off block time from m1

Scheduled arrival time	the expected arrival time on the tickets of passengers
Actual arrival time	Actual in block time from m3
Expected landing time	Expected landing time from m1
Airline	ICAO code of airline operating the flight
Cancellation	1 if the flight was cancelled, 0 otherwise
ddr number	id to cross data with pax_itineraries dataset

4.3 Preliminary results

In this Subsection we describe some preliminary results from the baseline implementation of the toy model for validation purposes. The goal is to show that the proposed toy model is able to describe the main characteristics of the ECAC ATM system on the day (September 12nd 2014) of operations when no Domino innovations were implemented. Starting from this benchmark, the future developments of this work will consider the implementation of the full mechanisms in order to evaluate their impact on air traffic dynamics.

For this purpose, we apply the baseline metrics introduced in Section 3.2 to the outputs of the toy model for a comparison with the observations of the day of operations for 12 September 2014.

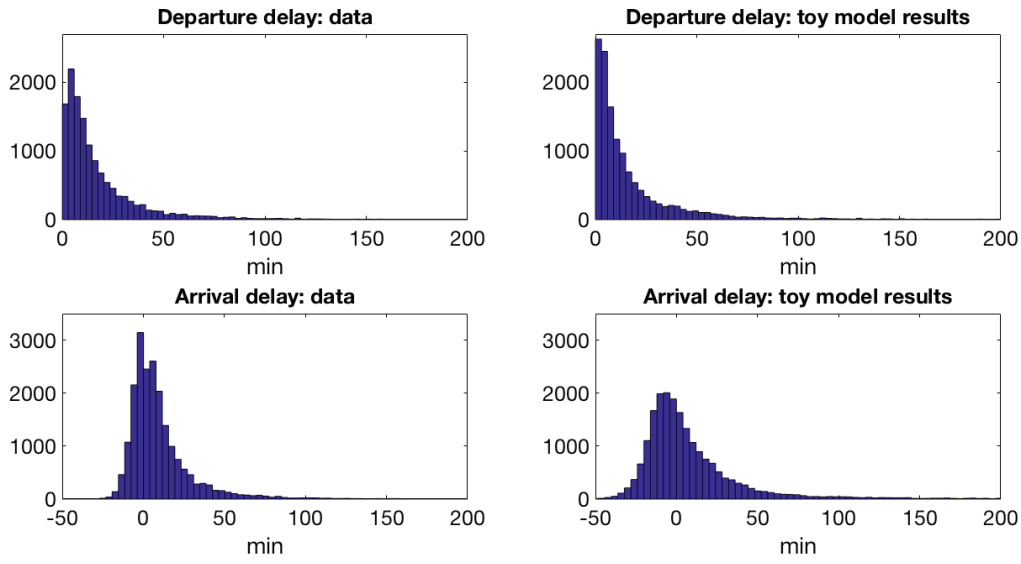

Figure 43. Delay on data and toy model comparison. Top panels: distribution of departure delays for delayed flights for the day of operations (12 September 2014) (left) and toy model simulations (right). Bottom panels: distribution of arrival delays of flights for the day of operations (12 September 2014) (left) and toy model simulations (right).

In Figure 43 we show the distributions of both departure and arrival delays of flights by comparing observed data (left) with the result of simulations. Note that for departure delays we observe a larger density at the first bin for the toy model outcomes with respect to data. This is due to the fact that we consider as different from zero also delays smaller than one minute. In Table 21 standard statistics of these distributions are shown to verify the consistency of the toy model outputs with respect to the real observations.

Table 21. Standard statistics of flight delay for the baseline implementation of the toy model

	Departure delay of delayed flights (data) – (min)	Departure delay of delayed flights (toy model) – (min)	Arrival delay (data) – (min)	Arrival delay (toy model) – (min)
Mean	19.6	20.9	11.1	12.9
Standard deviation	30.2	36.3	27.4	38.1
Interquartile range	18	17.9	17	23
Max	398	299	294	298
Min	0.5	1.0	-45.7	-47.5

In Table 21, we compare the departure delays of delayed flights because, as highlighted in Section 3.2, some negative departure delays may occur in reality when the aircraft boarding phase has been completed in advance with respect to the scheduled time. In these cases, the departure time may be anticipated, whereas the toy model does not consider this possibility, by design. The statistics obtained with the outputs of the toy model are consistent with the real data, thus indicating that the toy model is able to reproduce the characteristics of the system at the largest level of aggregation.

On 12 September 2014, there were no cancelled flights that can be identified from the available data. However, the simulations of the toy model give as output an average percentage of cancellations equal to 0.7%, that is a value close to the percentages observed in the datasets used in Section 3.2, i.e., 1.75% for ECAC_1 and 0.9% for US_april.

Let us consider also the intraday profile of mean departure delay for delayed flights (see the left panel of Figure 42). The toy model reproduces the well-known behaviour of increasing delays during the day, because of reactionary delay propagation by means of rotational and non-rotational effects. The model output reproduces quite well the real observations, except in the last part of the day when the mean delay is further increasing according to the toy model, while it remains quite constant when we look at real data. There are different reasons to explain this discrepancy. We exclude that it is due to the flight buffers. In fact, buffers concern the strategic phase that is used as input for the toy model. It could, instead, be related to the mechanism of delay recovery during the en-route phase by means of deviations from the planned trajectory which shorten the distance from the destination airport (sometimes called direct). The relevance of this effect in the ECAC airspace

and its frequency as a function of the time of the day has been elucidated, for example, in [40], see the right panel of Figure 44. It might also be due to other actions that airlines undertake to maintain delays low, which are not included in the model, especially as they get nearer towards curfew times at the end of the operational day. However, more research is called for the reconstruction of the intra-day patterns in the context of the toy model, when complete data for different days will be available. In Figure 45, we show the fraction of delayed flights at each hour of the day and we compare the results of the toy model with real observations for both departure and arrival delays. The two patterns are in good agreement.

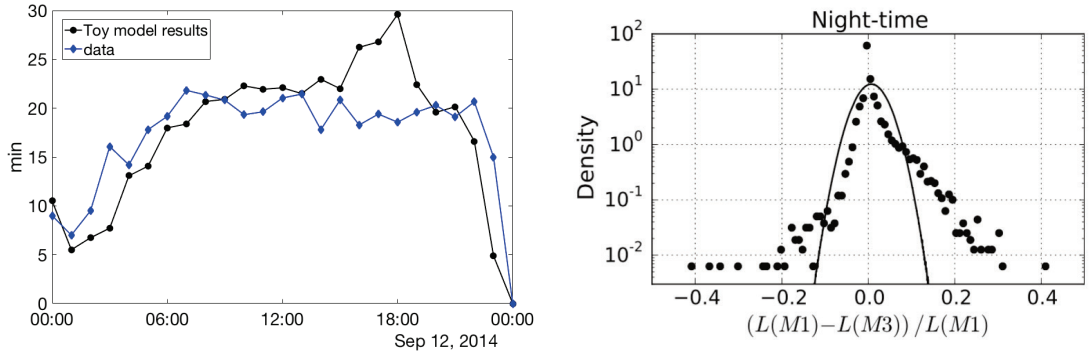

Figure 44. Departure delay and flight plan length variation as a function of time of the day. Left panel: mean departure delay for delayed flights at each hour for 12 September 2014. Right panel: probability density function of the relative difference between the length of the planned and actual trajectory of flights in the German airspace during the 334 AIRAC from 9pm to 6am, taken from [40].

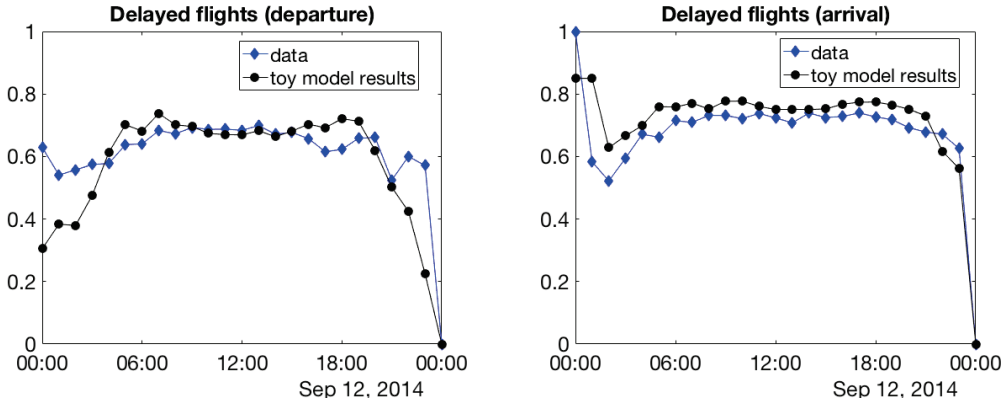

Figure 45. Fraction of delayed flights at each hour for 12 September 2014 by considering both departure and arrival delays.

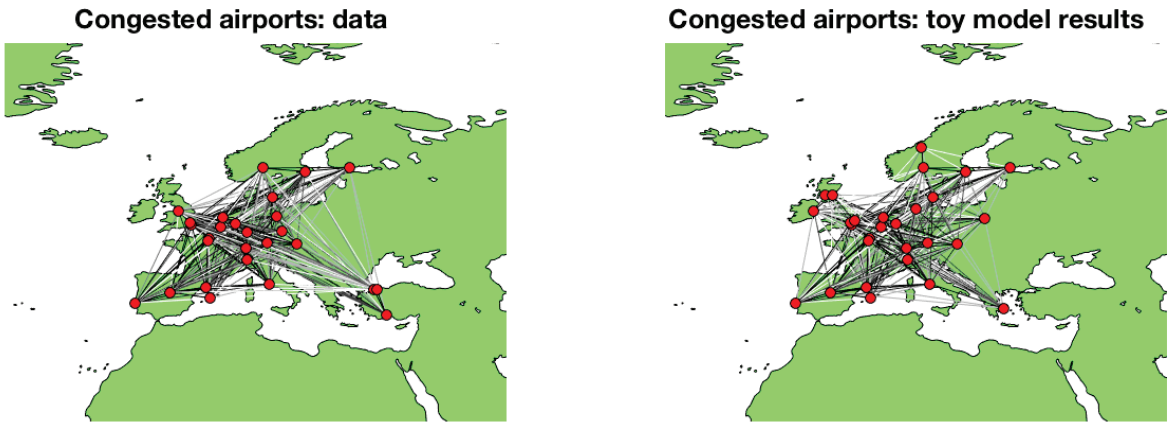

Figure 46. Geographical characterization of delays for 12 September 2014 according to the definition of distressed airports, i.e., red dots represent airports where more than 25% of flights have a departure delay larger than the average departure delay of delayed flights. Two node-airports are connected by a weighted link according to the number of flights connecting them at the day of operations: the darker is the colour, the larger is the number of flights. We compare the real case (left plot) with the outputs of the toy model (right panel). In both cases, 25% is the overall average fraction of delayed flights with a departure delay larger than the mean.

Finally, according to the definition of distress given in Section 3.2, we show in Figure 46 the European network of distressed airports for 12 September 2014 and compare it with the same network built with simulated data. In order to quantify the degree of overlap between the two networks, we adopt the error matrix which allows us to classify how much we learn from simulations about the distress state of the system by reporting the number of false positives, false negatives, true positives, and true negatives, see Table 22. The diagonal terms are quite high, showing a good classification ability. However, the number of false positive, i.e., the airports that the model wrongly classifies as distressed, is significant, indicating the need of refining further the toy model.

Table 22. Error matrix for the measure of distress applied to the outcome of the baseline implementation of the toy model with respect to real observations

True positive (%) 0.7407	False positive (%) 0.2963
False negative (%) 0.0155	True negative (%) 0.9823

5 Next steps and look ahead

In parallel to the analysis of metrics presented in this deliverable, the ABM model of Domino is being developed. The outputs of this model applied to different scenarios considering different levels of implementation of the mechanisms will be analysed using both classical and network metrics, as described in this deliverable. The results obtained from these analyses will be presented to stakeholders in a dedicated workshop. The feedback from the workshop will help to validate the utility of the metrics and help us defining the adaptive case studies. The results of these first results on the investigative case studies and of the workshop will be reported in D5.2 – Investigative case studies results and in D6.3 – Workshop results summary respectively. Both deliverables are planned for April 2019.

The toy model will be further developed to continue with the testing of the different metrics under different simplified scenarios.

Note that when using the full Domino ABM model, explicit passengers' itineraries and connections will be modelled. As described previously, in Domino, we expect to see how passengers and their behaviour in the system (e.g., being rebooked when missing connections) contribute to the networks metrics, affecting for example centrality metrics.

Finally, as the ABM model will consider the different systems in the ATM besides airports, the metrics will be applied/modified to capture the interaction between the ATM systems and not remain only at flights and airports level.

6 References

[1] M. Zanin and F. Lillo, “Modelling the air transport with complex networks: A short review,” *The European Physical Journal Special Topics,* vol. 215, no. 1, pp. 5-21, 2013.

[2] P. Mazzarisi, S. Zaoli, F. Lillo, L. Delgado and G. Gurtner, “Toward new metrics assessing air traffic interaction,” in *SESAR Innovation Days*, Salzburg (Austria), 2018.

[3] A. Cook and G. Tanner, “European airline delay cost reference values - updated and extended values (Version 4.1),” 2015.

[4] SESAR Joint Undertaking, “European ATM Master Plan, Ed. 2015,” 2015.

[5] DCI-4HD2D Project Consortium, “D3.2 Final technical report,” 2016.

[6] Vista Project Consortium, “D5.2 Final Report,” 2018.

[7] L. Katz, “A new status index derived from sociometric analysis,” *Psychometrika,* pp. 39-43, 1953.

[8] M. Newman, Networks: an introduction, Oxford university press, 2010.

[9] R. K. Pan and J. Saramäki, “Path lengths, correlations, and centrality in temporal networks,” *Physical Review E,* vol. 84, no. 1, p. 016105, 2011.

[10] P. Grindrod, M. C. Parsons, D. J. Higham and E. Estrada, “Communicability across evolving networks,” *Physical Review E,* vol. 83, no. 4, p. 046120, 2011.

[11] S. Boccaletti, G. Bianconi, R. Criado, C. Del Genio, J. Gomez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang and M. Zanin, “The structure and dynamics of multilayer networks,” *Physics Reports,* vol. 544, pp. 1-122, 2014.

[12] M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno and M. A. Porter, “Multilayer networks,” *Journal of complex networks,* vol. 2, no. 3, pp. 203-271, 2014.

[13] P. Holme and J. Saramäki, “Temporal networks,” *Physics reports,* vol. 519, no. 3, pp. 97-125, 2012.

[14] D. Kempe, J. Kleinberg and A. Kumar, "Connectivity and inference problems for temporal networks," *Journal of Computer and System Sciences,* vol. 64, no. 4, pp. 820-842, 2002.

[15] M. Zanin, L. Lacasa and M. Cea, "Dynamics in scheduled networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science,* vol. 19, no. 2, p. 023111, 2009.

[16] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society,* pp. 424-438, 1969.

[17] M. Billio, M. Getmansky, A. W. Lo and L. Pelizzon, "Econometric measures of connectedness and systemic risk in the finance and insurance sectors," *Journal of financial economics,* vol. 104, no. 3, pp. 535-559, 2012.

[18] F. Corsi, F. Lillo, D. Pirino and L. Trapin, "Measuring the propagation of financial distress with Granger-causality tail risk networks," *Journal of Financial Stability,* vol. 38, pp. 18-36, 2018.

[19] M. Zanin, S. Belkoura and Y. Zhu, "Network analysis of Chinese air transport delay propagation," *Chinese Journal of Aeronautics,* vol. 30, no. 2, pp. 491-499, 2017.

[20] A. J. Cook, G. Tanner, S. Cristóbal and M. Zanin, *New perspectives for air transport performance,* 2013.

[21] A. J. Cook, G. Tanner, S. Cristóbal and M. Zanin, "Delay propagation–new metrics, new insights," in *Eleventh USA/Europe air traffic management research and development seminar*, Lisbon, Portugal, 2015.

[22] Y. Hong, Y. Liu and S. Wang, "Granger causality in risk and detection of extreme risk spillover between financial markets," *Journal of Econometrics,* vol. 150, no. 2, pp. 271-287, 2009.

[23] M. Zanin, "On causality of extreme events," *PeerJ,* vol. 4, p. e2111, 2016.

[24] S. Belkoura and M. Zanin, *Phase changes in delay propagation networks. arXiv preprint arXiv:1611.00639.,* 2016.

[25] R. Tsay, Analysis of financial time series (Vol. 543), John Wiley & Sons, 2005.

[26] J. Johnston, Econometric Methods (Second ed.), New York: McGraw-Hill., 1972.

[27] B. Hansen, "Autoregressive conditional density estimation," *International Economic Review,* pp. 705-730, 1994.

[28] R. F. Engle and S. Manganelli, "CAViaR: Conditional autoregressive value at risk by regression quantiles," *Journal of Business & Economic Statistics,* vol. 22, no. 4, pp. 367-381, 2004.

[29] C. Bonferroni, "Teoria statistica delle classi e calcolo delle probabilità," *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze,* vol. 8, pp. 3-62, 1936.

[30] M. Tumminello, F. Lillo, J. Piilo and R. Mantegna, "Identification of clusters of investors from their real trading activity in a financial market," *New Journal of Physics,* vol. 14, no. 1, p. 013041, 2012.

[31] EUROCONTROL, "DDR2-Webportal," 8 June 2018a. [Online]. Available: http://www.eurocontrol.int/articles/ddr2-web-portal..

[32] EUROCONTROL, FAA, "Comparison of Air Traffic Management - Related Operational Performance: U.S./Europe," 2016.

[33] P. Fleurquin, J. J. Ramasco and V. M. Eguiluz, "Systemic delay propagation in the US airport network," *Scientific reports,* vol. 3, p. 1159, 2013.

[34] G. Caldarelli, A. Capocci, P. De Los Rios and M. A. Munoz, "Scale-free networks from varying vertex intrinsic fitness," *Physical review letters,* vol. 89, no. 25, p. 258702, 2002.

[35] A. Clauset, C. R. Shalizi and M. E. Newman, "Power-law distributions in empirical data," *SIAM review,* vol. 51, no. 4, pp. 661-703, 2009.

[36] EUROCONTROL, "Base of Aircraft Data (BADA)," 8 June 2018b. [Online]. Available: http://www.eurocontrol.int/services/bada .

[37] R. Ehrmanntraut and F. Jelinek, *Performance parameters of speed control and fuel,* 100(150), 200, 2010.

[38] European Parliament, "Regulation No 261/2004 of the European Parliament and of the Council. Establishing common rules on compensation and assistance to passengers in the event of denied boarding and of cancellation or long delay of flights, and repealing Regulation No 295/91," 2004.

[39] Domino Project Consortium, "D3.1 Architecture definition," 2018.

[40] C. Bongiorno, G. Gurtner, F. Lillo, R. N. Mantegna and S. Micciché, "Statistical characterization of deviations from planned flight trajectories in air traffic management," *Journal of Air Transport Management,* vol. 58, pp. 152-163, 2017.

[41] A. Cook, G. Tanner, G. Williams and G. Meise, "Dynamic Cost Indexing," in *6th EUROCONTROL Innovative Research Workshops & Exhibition*, EUROCONTROL Experimental Centre, Bretigny sur Orge, France, 4-6, 2007.

[42] A. Cook, G. Tanner, V. Williams and G. Meise, "Dynamic cost indexing–Managing airline delay costs," *Journal of air transport management,* vol. 15, no. 1, pp. 26-35, 2009.

7 Acronyms

ABM: Agent-based model

ACI EUROPE: Airport Council International Europe

AIRAC: Aeronautical Information Regulation and Control

ANSP: Air Navigation Service Provider

ATC: Air Traffic Control

ATFM: Air Traffic Flow Management

ATM: Air traffic management

AU: Airspace user

BADA: Base of Aircraft Data

DCI: Dynamic cost indexing

DDR2: Demand Data Repository

IFPS: Integrated Initial Flight Plan Processing System

NM: Network Manager

UDPP: User Driven Prioritisation Process

1 Annex – List of airlines in ECAC2 dataset

All airlines in the SkyTeam Alliance
All airlines in the One World Alliance
All airlines in the Star Alliance
Air Baltic
Air Berlin
Aer Lingus
Air Nostrum
Blue Air
EasyJet (DS)
easyJet
European Air Transport
Eurowings
Flybe
Germanwings
Hop
Jet2.com
Monarch Airlines
Niki
Norwegian Air International (D8)
Norwegian Air Shuttle
Olympic Airlines
Pegasus Airlines
Ryanair
SunExpress
Thomas Cook Airlines
Thomsonfly
Transavia Holland
Ukraine International Airlines
Vueling Airlines
VOLOTEA Airways
Widerøe
Wizz Air

2 Annex – Walks appearing in actual network

In this annex, we show how to partially solve the issue of new walks appearing in the actual network due to delays, which should not be counted by our centrality measures. Let us consider Trip centrality, and let $Q = (\mathbb{I} + \alpha A^{[1]}K)\dots(\mathbb{I} + \alpha A^{[T]}K)$. With this definition,

$$\vec{k}_{tm}^{\,in} = [Q - \mathbb{I}]K^{-1}\,\mathbb{1}$$

$$\vec{k}_{tm}^{\,out} = \mathbb{1}^{\,T}[Q - \mathbb{I}]K^{-1}.$$

The element Q_{ik} contains the contribution to the outgoing centrality of i given by walks from *i* to *j* (or equivalently the contribution to the incoming centrality of *j* given by walks from *i* to *j*). We call Q_{sched} the matrix computed with A_{sched}, and Q_{real} the matrix computed with A_{real} . Now, let us compute the matrix Q one time frame at a time. $Q_1 = (\mathbb{I} + \alpha A^{[1]}K)$ only counts the walks during the first time frame, $Q_2 = (\mathbb{I} + \alpha A^{[1]}K)(\mathbb{I} + \alpha A^{[2]}K)$ counts the walks up to the second time frame, and so on. At each time step, if an element of Q_{real} is larger than the corresponding element of Q_{sched}, it is because of a new walk opened up by a delay. In fact, in the real network departure and landings take place either at the same time or later than in the scheduled network (having put all negative delays to zero), therefore all new acceptable contributions to centrality are added to Q_{real} either at the same time as in Q_{sched}, or later. Therefore, at each step we pose $Q_{real} = min\{Q_{sched}, Q_{real}\}$. This eliminates most spurious walks. The correction procedure is analogous for TripRank, with the appropriate definition of Q. We remark that there are still some spurious walk that remain, even after this correction

-END OF DOCUMENT-