Towards New Metrics assessing Air Traffic Network Interactions

Piero Mazzarisi, Silvia Zaoli*, Fabrizio Lillo
Dipartimento di Matematica
University of Bologna
Bologna, Italy

Luis Delgado, Gérald Gurtner
School of Architecture and Cities
University of Westminster
London, United Kingdom

***Abstract*—In ATM systems, the massive number of interacting entities makes it difficult to predict the system-wide effects that innovations might have. Here, we present the approach proposed by the project Domino to assess such effects and identify the impact that innovations might bring for the different stakeholders, based on agent-based modelling and complex network science. Domino will model scenarios mirroring different system innovations which change the agents' actions and behaviour. Suitable network metrics are needed to evaluate the effect of innovations on the network functioning. We review existing centrality and causality metrics and show their limitations in characterising the network by applying them to a dataset of US flights. We finally suggest improvements that should be introduced to obtain new metrics answering to Domino's needs.**

I. Introduction

The introduction of changes in the ATM system is often difficult due to the tight interdependencies that exists across the different systems, subsystems and institutional frameworks. The full implications of changes on parts of the system are difficult to predict at system level.

This vision of how the system's elements are connected and of their criticality to propagate delay and cost might be different from the perspective of different stakeholders [1], [2].

At a time of increased traffic, the ATM system can improve its performance by being better tuned for flexibility to exploit the margins laying in operations to the best for stakeholders. For example, understanding the coupling between flights helps understand the margins embedded into the flight schedules designed by the airlines and can lead to better understanding of the coupling between stakeholders and processes.

This papers describes the approach taken in Domino (Section II), giving an overview of the three pillars of the project: the *methodology*, the *platform*, and the *network analysis toolbox*. Section III presents some work already performed on the toolbox,, consisting in an analysis of the limitations of current metrics to capture centrality and causality by analysing data from the Department of Transportation's (DOT) Bureau of Transportation Statistics from the US and in the identifications of directions of improvement of these metrics. In Section IV we draw some conclusions.

* The first two authors contributed equally to this work.

This project has received funding from the SESAR Joint Undertaking under grant agreement No 783206 under European Unions Horizon 2020 research and innovation programme. The opinions expressed herein reflect the authors views only. Under no circumstances shall the SESAR Joint Undertaking be responsible for any use that may be made of the information contained herein.

II. Domino's approach

The three main objectives of Domino can be summarised as: (i) provide a *methodology* to analyse the impact of changes in the system at network level; (ii) create a *platform* in the form of a detailed agent-based model where technological innovations and behavioural functions can be incorporated; (iii) create a *toolbox* based on complexity science techniques to analyse at network level the coupling of the ATM elements from a flight and passengers perspective.

A. Methodology

The delivered methodology will consist of a 'how-to' guideline to analyse the system at a network level. Recommendations on how to implement changes in the platform to study other test cases and on metrics to analyse the results will be laid out once enough knowledge is drawn from the analysis of the model's results.

Domino's methodology, summarised in Figure 1, will be developed and tested in different steps:

1) Definition of mechanisms that will be assessed by Domino and identification of *investigative case studies*.
2) Identification of changes in the ATM system needed to implement the mechanisms of each case study. Changes might concern the technical system, the stakeholder behaviour or the communication and information exchange.
3) If needed, modification of the elements in the model to adapt to 2.
4) Execution of Domino's model to generate realisations of the system under different scenarios.
5) Analyse the outcome of the model executions with the complexity science *network analysis toolbox* to characterize the system in each case study.
6) With the help of experts/stakeholders, modify the mechanism defined in 1 to create the *adaptive case studies* which are considered to mitigate any negative network effect identified in 5. Iterate the steps until an understanding of the implications of the changes introduced is achieved.

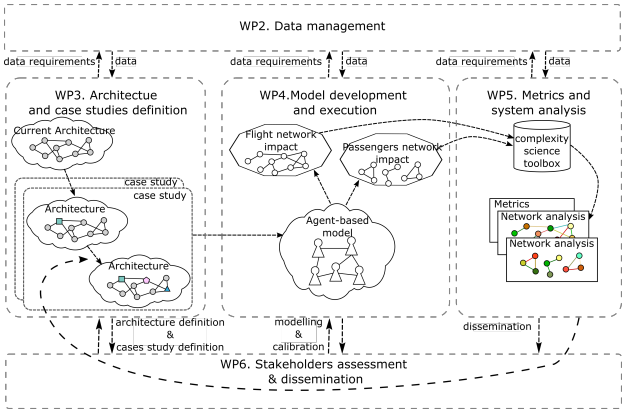

Figure 1. Workpackage structure and flow of activities

B. Platform

The platform developed by Domino, which is under development, will consist in a detailed agent-based model (ABM) able to execute pre-tactical and tactical phases ECAC-wide down to the passenger level. ABM are able to capture highly non-linear feedback by simulating massively interacting entities.

The model acts as a detailed numerical experiment where the modeller is in control of all the parameters and has access to all the intermediate states of all the agents. The interactions between entities drive the system and allow to capture high-level emergent phenomena.

C. Toolbox

The outcome of this model will be analysed by a complexity science toolbox including classical metrics but also network level indicators. This will allow modellers to identify potential bottlenecks and provide solutions to them, gaining understanding on how changes in elements of the system have system-wide implications. The toolbox is first validated on historical data and then applied to the outcome of the model, as shown in section III. This toolbox will identify, test, and validate metrics based on complexity science. The approach will be tested by developing relevant case studies.

III. Network Analysis Toolbox

To capture the effects of the considered mechanisms on the functioning of the ATM system, Domino needs to use a holistic approach not only in the system modelling but also in the interpretation of its results, moving from a microscopic view concerning the single flight to a macroscopic perspective considering the whole system. To this aim, Domino will establish a toolbox of complex network metrics able to characterise the ATM system and tell apart the consequences of the different innovations studied in Domino from the point of view of regulators, airlines and passengers. Regulators are mostly concerned by the system robustness and resilience, *i.e.*, the capability of the system to remain close to optimal state or to return quickly to it in the presence of perturbations like massive delays. Airlines, instead, measure delays especially in terms of cost. Finally, from the passengers' point of view, delays affect the network connectivity, *i.e.*, their possibilities to move through the network. Domino network metrics aim to capture all these different aspects.

The agent-based model will produce as an output the set of realised flights for the analysed day. This output can be seen as a directed network where the airports are the nodes and the flights are the links. The most general representation of such network is dynamic in time, as links appear and disappear according to the schedule. Moreover, it has a multi-layer (or multiplex) structure, where each layer contains the network of flights and airports of a different company, and inter-layer links connect nodes corresponding to the same airport. In the following analysis, we will neglect the temporal and multiplex structure, showing the limits of the associated metrics.

The network obtained as an output of the model will differ from the network of the scheduled flights due to delays and to cancellations. Delays and cancellations might change the network connectivity, the cost of delay paid by airline, and the probability of congestion with respect to the scheduled network. To assess the impact of the innovations introduced in the different scenarios on the realised flight network, we need a set of network metrics able to identify and quantify such changes.

Here, we review existing metrics and apply them to the dataset of US flights of 2015, with the aim of pinpointing their limits. In particular, we focus on two categories of metrics: centrality metrics and causality 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. When airports are ranked according to an appropriate centrality measure, the airports with the highest ranks are the ones providing to the passengers the highest potential of moving through the network. The loss of centrality of an airport, between the scheduled and the realised network, signals a diminished potential of moving through the network passing through that node, which means, from the passenger's point of view, a diminished performance of the network. In the light of Domino's scope, this loss of centrality should reflect both the missing links due to cancellations and the disrupted paths due to delays. Provided a centrality metrics satisfying these requirements, comparing the loss of centrality or the rank change between the realised and the scheduled flight network among case studies implementing different mechanisms would allow to assess the impact of innovations on the network performance. In particular, an innovation minimising the centrality losses and globally preserving the ranks between the scheduled and realised network represents an improvement from the passengers' point of view. In section III-B1 we review some of the most commonly used centrality metrics and in section III-B2 we show their limitations in describing the loss of connectivity of the network due to delays. Finally, in section III-B3 we suggest what improvements could be

introduced to make the existing metrics suited to Domino's purposes.

In the ATM system, delays and congestion states propagate through the system due to the entangled interactions between the flights and the environment, *e.g.*, the network manager, the airports or the arrival coordinators. As innovations aim to reduce the propagation of delays, the complex network toolbox should include a metric able to detect the extent to which the congested state of an airport causes congestion in other nodes of the network. In Statistics, a (directional) causal relation between two systems is detected when the information on the state of one system helps in predicting the future state of the other. The presence of a causal relation is assessed by means of statistical tests whose most famous example is the Granger causality metrics [3]. Indeed, it has been recently applied to airport networks [4], [5].

Here, a data driven approach is adopted to describe the ex-post dependence structure of delay propagation, identifying the channels of the spreading process and establishing a network of causal relations. This analysis is applied to the network of airports, where the average flight delay measures the state of congestion of an airport and a link between two airports represents a channel of delay propagation. In view of Domino's goals, the study of causality networks, whose topology may change depending on implemented scenarios, relates the presence of innovations at the micro level to its impact on delay dynamics and propagation at some macro level of aggregation, such as airports or airline companies. For example, a smaller number of causal links and less causal feedbacks can be seen as an improvement of the system, as they signal a diminished coupling of the systems' elements. In section III-C1 we review Granger causality metrics and its recent application to ATM systems. Then, in section III-C2 we show some limitations in describing non-linear aspects of delay propagation and possible spurious causal relations as a consequence of the autocorrelation structure of the delay states. Finally, in section III-C3 we suggest the improvements that could be introduced to make the existing metrics suited to Domino's purposes.

A. Dataset

To show the limitations of existing metrics, in the following we apply them to the network of flights operated in 2015 by 14 major US airlines. The dataset was obtained from the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics. For each flight, the dataset reports the date, the airline operating it, the departure and arrival airport, the scheduled departure and arrival times and the realised ones, the aircraft tail number, whether it was cancelled or diverted. All schedules were converted from local time to Eastern Standard Time (EST). For the centrality analysis, performed on one day, the day was considered to start at 4AM EST. This choice reflects the fact that, as shown already in [6], very few flights depart between 0AM an 4AM local time, therefore 4AM EST is a time of minimum activity across all the country. Causality analysis was instead performed on hourly time series ranging from one to three months.

B. Centrality metrics

1) State of the art: Commonly used centrality metrics apply to single-layer static networks. Let us therefore start by considering the network of flights and airports aggregated across layers, *i.e.*, across airlines, and across time frames, *i.e.*, where all flights are present at the same time regardless of their schedule. Let A be the weighted adjacency matrix of the network, such that $A_{ij} = k$ if there are k flights going from i to j. Here, we consider three among the most common and well known centrality metrics: degree centrality, Katz centrality and Page Rank. Since the network of flights and airports is directed, a distinction should be made, in each case, between incoming and outgoing centrality.
The incoming (outgoing) degree centrality of a node i is given by the number of incoming (outgoing) edges (each flight is considered as an edge),

$$d_i^{IN} = \sum_j A_{ji}, \tag{1}$$

$$d_i^{OUT} = \sum_j A_{ij}, \tag{2}$$

where the index j runs on all nodes. This centrality measure with how many flights node i can be reached (respectively, how many flights depart from node i). However, an important feature of the flight network are connections, which make use of two or more flights. 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]. The incoming Katz centrality of node i is

$$k_i^{IN} = \sum_j (\mathbb{I} - \alpha A)^{-1}_{ji} = \sum_j \sum_{n=0}^{\infty} \alpha^n (A^n)_{ji}, \tag{3}$$

that is, each walk of length n from any node j of the network to i contributes α^n to the centrality of i. Since $\alpha < 1$, longer walks contribute less and its value determines what is the contribution of long walks to centrality. The weight α must be smaller than the inverse of the largest eigenvalue of A for the expression to converge [7]. Correspondingly, the outgoing Katz centrality of node i is

$$k_i^{OUT} = \sum_j (\mathbb{I} - \alpha A)^{-1}_{ij} = \sum_j \sum_{n=0}^{\infty} \alpha^n (A^n)_{ij}. \tag{4}$$

Page Rank 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,

$$p_i^{IN} = \sum_j (\mathbb{I} - \alpha D^{-1} A)^{-1}_{ji}, \tag{5}$$

where $D_{ij} = \delta_{ij} d_j^{OUT}$, so that a link from j to k is weighted by the inverse of the out-degree of j, $1/d_j^{OUT}$.

2) Application of the existing metrics to the US flights dataset: To apply centrality metrics, we selected two days of the dataset differing in the amount of delay realised on the network. We considered 4 global parameters characterising delay: the fraction of delayed flights, the total delay, the average delay and the average delay of delayed flights. On first selected day, April 3rd 2015, all these parameters are below or close to the average (computed on all days), while on the second considered day, April 9th 2015, all parameters are above average. Additionally, April 3rd had 87 cancelled flights while April 9th had 246. In the following, we refer to these two days respectively as "day 1" and "day 2". For each day, we computed the airports' ranking according to each of the three centrality metrics reviewed in section III-B, incoming and outgoing, for the scheduled and the realised network. The obtained ranking are compared using 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.

For Katz centrality, we chose $\alpha = 0.003$, assuring convergence of the metric for both chosen days. Note that this small value of α penalises strongly long walks, therefore we do not expect the ranking to differ much from the degree ranking. For Page Rank centrality, instead, larger αs still allow convergence, therefore we chose $\alpha = exp(-1/2)$, so that walks of length $n \leq 2$ are given a non negligible weight.

The rankings according to incoming and outgoing centralities result are very similar according to all three metrics, displaying, for day 1, respectively τ =0.97, 0.97 and 0.93 on the scheduled network and τ = 0.97, 0.97 and 0.93 for the realised one. Also the rankings according to the centrality computed on the scheduled network and on the realised one are quite similar for both days. For day 1, the rankings display correlations, respectively for the three metrics, τ =0.996, 0.995 and 0.995 in the incoming case and τ =0.996, 0.991 and 0.991 in the outgoing case. For day 2, we have τ =0.990, 0.985 and 0.995 in the incoming case and τ =0.980, 0.976 and 0.992 in the outgoing case. The slightly smaller rank correlations coefficients for day 2 are due to the larger number of cancelled flights with respect to day 1. However, none of the considered centrality measures is able to reflect the fact that, on day 2, the much larger and abundant delays certainly caused more disruption of the network connectivity.

While the rankings according to degree and Katz centrality are similar (for the scheduled network, incoming case, τ =0.90 for day 1 and τ = 0.88 for day 2), Page Rank introduces stronger ranking differences with respect to Katz (for the scheduled network, incoming case, τ =0.77 for day 1 and τ = 0.68 for day 2)[1]. Figure 2 shows a comparison of the two rankings, highlighting that most of the difference is due to a group of airports having a low ranking according to Katz centrality and getting a strong ranking boost with

[1]This difference is not due to the different values of α in the two cases.

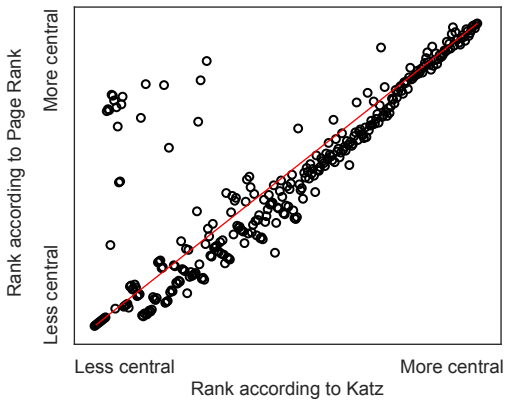

Figure 2. Comparison of airport ranks according to incoming Katz centrality and incoming Page Rank centrality for the scheduled network on day 1. The red line is the 1:1 line. Points above the red line represent airports having gained importance with Page Rank.

Page Rank (in the upper left part of the figure). T hese are mostly small airports in Alaska having direct flights t o the airport of Anchorage. As Anchorage has itself a strong rank increase due to having several directed flights f rom airports with low out-degree, all the airports connected to it by a direct flight a lso i ncrease t heir r anking. T his o utcome, with a set of peripheral airports climbing the ranking, questions the suitability of Page Rank centrality to characterise node importance in ATM networks. In genreal, these differences between different centrality metrics highlight the fact that each metric describes a different aspect of the network structure, and care should be taken in their comparison. For example, degree considers only direct links, therefore it is appropriate if we are interested in assessing the potentiality of an airport to provide direct connections to other airports of the network, but it is not able to evaluate the role of flight c onnections. Katz centrality and Page Rank, instead, take into account also walks of any length on the network. While walks on the aggregated, static network considered here 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 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 way of weighting walks in the two metrics, Katz centrality favours airports linked to large airports (with many link), as they will have many walks departing or arriving, while Page Rank rather tends to favour airports with more links to smaller sized airport.

3) Limitations of existing metrics and suggested improvements: To evaluate the effect of the innovations addressed by Domino on the network performance, a centrality metric should be able to tell apart a situation where delays disrupt connections to one where they do not. Specifically, an airport's centrality should reflect i ts p articipation t o w alks t hat can actually be travelled, *i.e.*, respecting the schedule, so that disrupted connections imply a centrality drop. We showed in

section III-B2 that this is not the case for existing centrality metrics. In fact, all three metrics presented here do not account for the temporal structure of the network. Katz and Page Rank centrality, in particular, count walks on the network which are not time ordered and therefore have no relation with the trajectories that passengers could travel. As a consequence, these metrics cannot reflect the effect of delays on the network's connectivity. Additionally, the weight assigned to each walk does not consider which airline each flight composing the walk belongs to, 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, as it is travelled with a smaller probability. Accounting for this requires considering the multiplex structure of the network.

Generalisations of the existing metrics should therefore be devised to overcome these limitations. A version of Katz centrality for temporal network has been proposed in [8]. It considers adjacency matrices $A^{[t]}$ containing only the links present in a time frame around time t and counts walks which are ordered in time, However, it does not account for the links' schedule. A solution to account for schedule by introducing secondary nodes is introduced in [9]. Therefore, Katz centrality and Page Rank centrality could be generalised by joining the approaches of references [8] and [9]. Furthermore, to differentiate between within-airline and across-airlines walks, the multiplex nature of the network should be considered. Centrality measures for multiplexes are reviewed in [10], however they either consist in computing the centrality of an airport separately on each layer and then aggregating the single-layer centralities to obtain a global centrality (*e.g.*, by summing or averaging the single-layer centralities) or in computing the centrality on an aggregated network, which adjacency matrix is the sum of the adjacency matrices of all layers. The first approach only counts within-airline walks, neglecting inter-layer ones. The second one, which corresponds to what we have presented in section III-B1, counts instead both intra- and inter-layer walks without distinction in weights. An intermediate approach should weight with a parameter $\epsilon \in [0, 1]$ each change of layer, so that walks using links on several layers are included in the centrality computation but contribute less than an intra-layer walk of the same length. Such a solution could be implemented, for Katz and Page Rank centrality, by considering one copy of each airport on each layer and having all copies connected at all times by a link of weight ϵ. Centrality could therefore be computed for each copy of an airport and then suitably aggregated.

C. Causality metrics

1) State of the art: A method to test whether there is a causal relation between two time series was first proposed by Granger [3] and is based on the idea that, if the knowledge of past observations of one time series allows us to estimate future observations of the other time series better than without considering them, then there exists a directional causal relation. Here, we review the application of the Granger causality metrics to the ATM network system. We quantify an airport's congestion by a stochastic variable X whose realisation x_t at time t is given by the average delay of flights taking off from that airport in the time interval centred in $[t, t + \Delta t]$. Flight delay is defined as the difference between the take-off time and the scheduled departing time. We considered Δt =1 hour and when no departing flights are present in the interval we set $x_t = 0$.

a) Granger causality in mean [3]: $Y \equiv \{y_t\}_{t=1,\dots,T}$ is said to Granger-cause $X \equiv \{x_t\}_{t=1,\dots,T}$ 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 [11]. Let us consider X and Y described by

$$\begin{cases} x_t &= \phi_0^1 + \sum_{j=1}^{p} \phi_j^{11} x_{t-j} + \sum_{i=1}^{p} \phi_i^{12} y_{t-i} + \epsilon_t^1 \\ y_t &= \phi_2^1 + \sum_{j=1}^{p} \phi_j^{21} x_{t-j} + \sum_{i=1}^{p} \phi_i^{22} y_{t-i} + \epsilon_t^2 \end{cases} \quad (6)$$

where $\epsilon_t^1, \epsilon_t^2$ are taken to be two uncorrelated white-noise series. The goal of the test [3] is to assess the statistical significance of $\{\phi_i^{12}\}_{i=1,\dots,p}$ by considering as null hypothesis that they are zero, *i.e.*, $H_0 : \{\phi_i^{12} = 0\}_{i=1,\dots,p}$. The null hypothesis H_0 is equivalent to considering that $\{x_t\}$ evolves according to a AR(p) process. After estimating both VAR(p) and AR(p) models, an F-test [11] is applied in order to test if VAR(p) outperforms statistically AR(p) in fitting the observations $\{x_t\}$. If it does, H_0 is rejected, meaning that Y 'Granger-causes' X.

b) Granger causality network: Having established how to detect a causal relation, we can consider the network of airports where a link $i \to j$ is present if i 'Granger causes' j. This approach has already been considered in some recent works in Econometrics [12], [13] and in a recent analysis of the Chinese air transportation network [5]. Given N time series, representing the state of delay of the N airports in the network, Granger causality test is performed on all the possible $M = N(N-1)$ pairs. When performing multiple hypothesis testing, a correction to the significance level of each single test should be applied to obtain the desired overall level γ, *i.e.*, if we test M hypotheses simultaneously with a desired γ, then a significance level $\gamma' < \gamma$ should be applied to each single test to correct for the increased chance of rare events, and therefore, the increased probability of false rejections [14]. This has typically not been considered in the literature. However, it can have a huge impact on the number of detected causal links, as we show in the following. Here, we apply the Bonferroni correction which compensates for this effect in the most conservative way by setting $\gamma' = \gamma/M$. Standard topological network metrics can then be extracted from the network of causal relations, *e.g.*, link density, clustering, assortativity, efficiency, diameter, centrality rankings of nodes. Each of these metrics describes some specific structural characteristic of the Granger causality network. For example, link density is a measure of the coupling of airports, since a larger number of links means more delay propagation, while measures of node

centrality indicate which airports are participating more often to delay propagation.

2) Application of Granger causality metrics to the US flights dataset: Time series of the state of delay for each airport are built for the period from January 1st 2015 to March 31st 2015. As suggested in [5], a Z-Score standardization procedure is applied 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. The standardized time series of airport i is calculated as $\tilde{x}_{i,t} = (x_{i,t} - \bar{x}_i^t)/\sigma_i^t$ where $\bar{x}_i^t$ and σ_i^t are the mean and the standard deviation of the delay states of airport i recorded at hour t across all available days. Hence, pairwise Granger causality tests are applied to the new standardized time series according to Eq. 6 for different p, ranging from 1 to 6 hours. The maximum lag is chosen equal to 6 because the empirical partial autocorrelation function becomes statistically zero after the sixth lag for the time series of any airport. In case of rejection of H_0, the best p is selected according to the Bayesian Information Criterion. Best p values are distributed around 1 and 2 hours, meaning that delay propagation happens on short timescales. Finally, we set $\gamma = 5\%$ and, as a consequence, the significance level of each test is $\gamma' = \frac{0.05}{N(N-1)}$ where $N = 315$.

The obtained Granger causality network has $L = 4401$ Granger causal links. Note that the link density for the Bonferroni corrected network is ~ 0.04, whereas without the correction we obtain ~ 0.45, much larger. Therefore, neglecting to introduce a correction means considering a large number of non-significant causal links. We find a positive linear correlation (0.62) between airport size, measured as the average number of flights per hour, and node (in- or out-) degree in the Granger network. The diameter of the Granger network, *i.e.*, the longest path connecting two nodes, is equal to 8 while for an Erdos-Renyi network with the same number of links (on average) is 4, thus suggesting the presence of outlying nodes less connected with the central core. This is confirmed also by the average path length, equal to 3.05 in the Granger network and to 2.4 in the corresponding Erdos-Renyi. 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 any open and closed triangle. It is 0.28 in the Granger network. This number is much larger than the one of the corresponding Erdos-Renyi network (0.08 ± 0.01), a difference explained by the different degree of nodes. In fact, the fitness model [15], which preserves on average the degree sequence, has a global clustering coefficient of 0.29 ± 0.01, in line with the Granger network. However, when we consider only feedback triangles, i.e. triangles with all links directed clockwise (or anti-clockwise), among all possible triplets, we count 14856 such triangles, a number much larger than the corresponding random cases, 908 ± 46 for the Erdos-Renyi network and 7656 ± 352 for the fitness model, suggesting that these feedback loops are over-expressed in the ATM system. In fact, a feedback triangle represents a positive feedback subsystem which tends to amplify delay propagation, thus making the system more unstable. Hence, in the case of ATM systems, an interesting clustering measure is the one which considers feedback loops and any innovation which aims to increase the resilience of the system, should tend to reduce it.

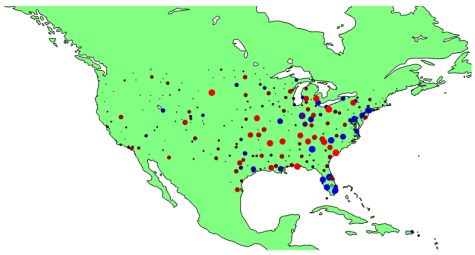

Figure 3. Outgoing PageRank node centrality for the Granger causality in mean network (blue dots) and the Granger causality in tail network (red dots). Increasing dot size and brighter colour represent a rank increase.

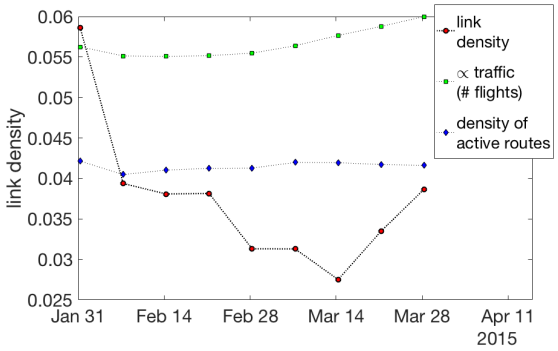

Figure 4. Link density (red dots) of the Granger causality in mean network for different 30-days periods (indicated by the last day) and compared with traffic (green dots) measured as the total number of flights (rescaled by a factor 8×10^6) and with density of active air routes in the aggregated network of airports and flights (blue dots).

Moving to node-specific topological metrics allows us to better characterize the US ATM system. In particular, PageRank centrality reveals a bipartite structure corresponding to the two macro geographical regions of US, *i.e.*, East and West. Figure 3 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.

Finally, we repeat the pairwise causality 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 4. The result suggests that link density, *i.e.*, a measure of how much the system is interconnected, does not depend trivially from both total traffic and active air routes connecting airports, that 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 result

highlights the need of further studies and improvements of network causality metrics.

3) Limitations of existing metrics and suggested improvements: The results presented in the previous section are based on linear models. However, the complex nature of the delay propagation dynamics might not be fully captured by linear models. For example, departing delays which are small with respect to flight time are probably not relevant for delay propagation, as they are easily absorbed during the flight or by buffers. These small delays are nevertheless considered by the Granger causality test and might produce spurious causality relationships. For this reason, we propose to use an extension of the Granger causality test, namely *Granger causality in tail* [16], which considers only extreme events, defined as states of delay falling in the right tail of the distribution, *i.e.*, large delays. With the same spirit of [3], Granger causality in tail aims to evaluate whether extreme events in an airport cause extreme events in another airport. Airports are now described by a binary variable $\tilde{X}$, the state of congestion, which is 1 if its (detrended) state of delay is extreme and zero otherwise. The Granger causality in tail test works as follows. Assume to know at each step the probability density function of $\tilde{X}$ conditional on past values[2] and let us define $V_t \equiv V(\tilde{x}_1, ..., \tilde{x}_{t-1}, \beta)$ as the $(1-\beta)$-quantile of the conditional probability distribution of the time series $\tilde{X}$, *i.e.*, $\mathbb{P}(\tilde{X} > V_t | \tilde{x}_1, ..., \tilde{x}_{t-1}) = 1-\beta$ almost surely with $\beta \in (0,1)$ defines V_t implicitly. The null hypothesis H_0^{tail} of [16] is:

$$\mathbb{P}(\tilde{X} > V_t | \{\tilde{x}_s\}_{s=1}^{t-1}) = \mathbb{P}(\tilde{X} > V_t | \{\tilde{x}_s\}_{s=1}^{t-1}, \{\tilde{y}_s\}_{s=1}^{t-1}) \quad \text{a.s.} \tag{7}$$

meaning that predicting an extreme event for $\tilde{X}$ with or without the past information on $\tilde{Y}$ is statistically equivalent. A rejection of the null hypothesis H_0^{tail} means that Y 'Granger causes in tail' X at level β. For further information on how to make testable the definition in Eq. 7, see [16]. Preliminary result were obtained with this method using the autoregressive conditional density model [17] and by assuming an AR(p) model for $\tilde{X}$ with i.i.d. Gaussian innovations and $\beta = 0.05$. Figure 3 shows PageRank centralities of nodes of the Granger causality in tail network. As before the structure is bipartite, but the most central airports are now different from the ones selected by Granger causality in mean and, more specifically, PageRank centralities obtained with the two metrics have low correlation (Kendall correlation coefficient 0.21). Further investigations are required to better understand the cause of this different outcome and therefore if Granger causality in tail captures relevant information about the ATM system.

One of the goals of Domino is to characterise the coupling of different subsystems in the different scenarios. In this direction, causality metrics could be applied to the multiplex of airlines to detect causalities among layers, measuring how delays propagates to one airline to another.

Finally, to assess the system's performance from the point of view of airlines and passengers, the same analyses can be performed considering as a variable the cost of delay instead of the delay itself, which determines the importance of a delay for these stakeholders.

[2]Conditional density for a time series can be estimated, *e.g.*, by historical simulation methods or autoregressive conditional density model [17].

IV. Conclusions and Further Work

In this paper we presented the approach of the Domino project to evaluate the effects of technical and behavioural innovations introduced in the ATM system. The approach is based on an agent-based model describing the complex interaction taking place among a large number of entities in different innovation scenarios and on a complex network toolbox to analyse the modelling results. The toolbox should include metrics to apply to the network of airports and flights able to evaluate the improvement (or worsening) of the network functioning in the different scenarios form the point of view of different stakeholders. In particular, centrality and causality metrics have been considered, owing to their capacity to measure the network connectivity and the propagation of delays and congestion in the network. However, we have shown here that existing centrality and causality metrics are not sufficient for the scopes of Domino. Specifically, existing centrality metrics are not able to tell apart a situation where delays disrupt important connections to one where they do not and do not account in a satisfactory way for the multiplex nature of the network. On the other hand, commonly used causality metrics assume linearity in the delay propagation, which might not be realistic. We therefore suggested directions in which these metrics should be improved to serve Domino's purposes and, possibly, to analyse other types of transportation networks as well. Further work is now required to implement the new metrics suggested here and prove their success in characterising the system under study.

References

[1] A. Cook, G. Tanner, S. Cristóbal, and M. Zanin, "Passenger-oriented enhanced metrics," in 2nd *SESAR Innovation Days*, 2012.

[2] A. Montlaur and L. Delgado, "Flight and passenger delay assignment optimization strategies," *Transportation Research Part C: Emerging Technologies*, vol. 81, pp. 99 – 117, 2017.

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

[4] A. Cook, H. A. Blom, F. Lillo, R. N. Mantegna, S. Miccich, D. Rivas, R. Vzquez, and M. Zanin, "Applying complexity science to air traffic management," *Journal of Air Transport Management*, vol. 42, pp. 149 – 158, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0969699714001331

[5] 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.

[6] P. Fleurquin, J. J. Ramasco, and V. M. Eguiluz, "Systemic delay propagation in the US airport network," *Sci. Rep.*, vol. 3, p. 1159, jan 2013.

[7] M. Newman, *Networks: An Introduction*. Oxford University Press, 2010.

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

[9] 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.

[10] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-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.
[11] J. Johnston and J. DiNardo, *Econometric methods*. New York, 1972, vol. 2.
[12] 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.
[13] 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.
[14] M. Tumminello, S. Micciche, F. Lillo, J. Piilo, and R. N. Mantegna, "Statistically validated networks in bipartite complex systems," *PloS one*, vol. 6, no. 3, p. e17994, 2011.
[15] 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.
[16] 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.
[17] B. E. Hansen, "Autoregressive conditional density estimation," *International Economic Review*, pp. 705–730, 1994.