Network-wide Assessment of 4D Trajectory Adjustments using an Agent-based Model

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

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

Damir Valput
Innaxis
Madrid, Spain

***Abstract*—This paper presents results from the SESAR ER3 Domino project. It focuses on an ECAC-wide assessment of two 4D-adjustment mechanisms, implemented separately and conjointly. These reflect flight behaviour en-route and at-gate, optimising given (cost) objective functions. New metrics designed to capture network effects are used to analyse the results of a microscopic, agent-based model. The results show that some implementations of the mechanisms allow the protection of the network from 'domino' effects. Airlines focusing on costs may trigger additional side-effects on passengers, displaying, in some instances, clear trade-offs between passenger- and flight-centric metrics.**

I. Introduction

A major challenge facing ATM architects is understanding the complex interdependencies and coupling of various components of the system. Predicting how the introduction of change in one (sub)system will impact others, not only locally, but downstream and more widely across the system, is a particular challenge. The 'Airspace Architecture Study' [1] flags how the ATM system is comprised of nodes, which are often operating close to maximum capacity and without appropriately connected resources (including for data). Problems due to stresses in the system thus often propagate and "knock it out of optimal flow". This 'domino' effect is offset, for example, by buffers in schedules, although these are often insufficient to absorb all of the disturbance(s). The Study flags the need for "stronger linking between airspace, operations and technical evolution and measurement of the impact through simulations factoring in known deployments ...". It also recommends targeted incentives for "early movers". Acceleration of market uptake of the next generation SESAR technologies and services, to support defragmentation, is, *inter alia*, further endorsed in the corresponding 'Transition Plan' [2], which identifies three such key operational and technical measures that need to be implemented in the very short term (2020 to 2025), to initiate the changes outlined in the Study. This particular measure will allow different parts of the system to be implemented at different speeds, but with awareness of local needs and coherence at the network level.

This project has received funding from the SESAR Joint Undertaking under grant agreement No. 783206 under the European Union's 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.

The Domino project has developed a platform to assess the coupling of ATM systems from a flight and passenger perspective, allowing the ATM system designer to better understand the relationships between (sub)systems and the nature of such relationships, which emerge in a given technological and operational context. It uses metrics from network science and classical approaches – see [3], for an early review of the value of non-classical metrics and the need to differentiate between flight- and passenger-centric indicators. Exploring several issues outlined in the Airspace Architecture Study in the context of different (future) operational and stressed environments, we present in this paper an assessment of two 4D flight trajectory adjustment mechanisms. The first is dynamic cost indexing (DCI), whereby a flight is able to adjust its cruise speed to manage expected delay. The second is wait for passengers (WFP), whereby a flight can wait at-gate for late connecting passengers. The conjoint use of DCI and WFP was studied in operations at a specific hub in previous research [4]. In this paper, the use of these mechanisms is considered at the network level. Moreover, different implementations modes of these mechanisms are explored, at current and advanced 'levels'. Section II presents the model and its scenarios; Section III focuses on the novel indicators developed to capture the network effects of the mechanisms. Section IV brings these together in the presentation of the results, whence the paper closes with conclusions and proposals for future work.

II. Model and Scenarios

A. Summary of the model

The Domino model is a new, expanded version of a simulator called 'Mercury'. Its main features have been presented in [5]. Mercury is an ECAC-wide microscopic agent-based model (ABM) comprising the following main agents:

- 'Flight': compute iteratively, segment by segment, the real flight trajectory based on the planned one.
- 'Airline': the airline manages its operations by taking care of passengers (connections, compensation, etc.) and flights (dispatching, cancellations, slot swapping, etc.); decisions are based on a cost of delay function, calibrated with [6]; airlines are part of alliances, used for rebooking passengers when needed.
- 'AMAN': the arrival manager takes care of the sequencing of flights close to the destination airport.

- 'Radar': follows the flight trajectory and relays important information to interested parties (AMAN, airline).
- 'Ground airport': takes care of the flight before and after departure; used to sample taxi, turnaround, and pax connecting times.

Mercury is executed as an event-driven simulator. In addition to events (e.g., flight pushback), agents also react to messages from other agents (e.g., a request for rebooking). Random variables are used for uncertainties such as taxi times, wind, pax connecting times, turnaround times, and cancellations, as well as for some non-optimal decision- making processes by agents (e.g. selecting the flight plan to be operated as part of the dispatch process). All distributions are built on various empirical data. Due to the stochasticity, several runs of the model are performed for each scenario.

Each airport has an instance of the 'ground airport' agent and the 'AMAN'. There is one instance of the 'airline' agent per individual airline and one 'flight' agent per flight. Passengers typically have only few decisions to make and thus are not modelled as agents, even though their preferences are taken into account by the airlines indirectly through a soft cost based on a logit utility function (calibrated again with [6]). The results in this paper have been obtained with an early version of the model, presented in [7]; it will be subsequently refined.

B. Scenarios

The entire ECAC space is modelled considering commercial flights on 12 September 2014. This includes approximately 27k flights and 3.4M passengers (considering premium and non-premium ticketed passengers), between 800 airports. This date was carefully selected to be representative of a high-traffic, non-disrupted day in 2014, with demand thus similar to an average day in 2023 (STATFOR baseline forecast). The traffic is based on historical DDR data, schedules and generated passenger itineraries (see [8] for a detailed description of the scenario generation), and calibrated using historical data from CODA and DDR. The main scenarios simulated in Domino have been presented extensively in [8]. The main variables considered to generate the scenarios are:

- the system delay: nominal and higher delays (baseline and stressed scenarios). The nominal delay is based on historical data on uncertainty and delay (e.g. number of ATFM regulations) for average days. The stressed scenario considers degraded days of operations with high levels of delay by selecting a high number of ATFM regulations, lower airport capacities and longer en-route and taxiing operations.
- the technological environment, by defining three mechanisms: flight prioritisation (allowing ATFM slot swapping); flight arrival coordination (with different implementations of E-AMAN); and, 4D trajectory adjustments (4DTA) (which considers DCI and WFP).

Results from several combinations of these factors have been obtained and are reported in [7]. In this article, we focus on the results obtained for the 4DTA mechanism. Three levels of this mechanism are modelled, considering increasingly advanced behaviours ('Level 0', 'Level 1' and 'Level 2').

Let us consider the two sub-mechanisms of 4DTA. (i) DCI considers changing the cruise speed of a flight (cost index) to manage expected arrival delay, while maintaining the route. The cost index is defined as the ratio between the time and fuel costs: $CI = (\text{time costs})/(\text{fuel costs})$. Domino explicitly estimates the cost of delay considering several factors: non-passenger costs (flight crew and maintenance), passenger costs (hard costs (e.g., rebooking, Regulation 261 duty of care) and soft costs (loss of market share due to dissatisfaction)). Based on BADA4, an airline is able to estimate the cost of fuel required to recover some delay. A fixed price per kg of fuel is used in the model (0.5 EUR/kg). (ii) WFP rules consider actively delaying outbound flights to wait for delayed inbound passengers so that they do not miss their connections. This option is currently seldom used by airlines, as it impacts the on-time performance of the outbound flights and, in some cases, waiting for passengers might lead to outbound flights being regulated. However, when the optimal solution to minimise the cost of delay is sought, then this might be a relevant strategy [4]. This could be particularly important for the last flights of the day where, if passengers do not make their connection, they will need to be rebooked on next-day flights, leading to significant (hard and soft) costs for the airline. The DCI and WFP rules vary across the three levels of implementation, with the primary goal of Level 0 being to capture the most common current practices of airline operators. Levels 1 and 2 explore advanced (future) capabilities. DCI is implemented as follows across the levels:

- Level 0: The cost index (CI) is calculated before take-off and is fixed throughout the flight. The CI decision is based on the departure delay, with delays larger than 15 minutes recovered up to 5 minutes. Whether a flight intends to recover any delay is decided according to a linear probabilistic distribution, with delays larger than 60 minutes always recovered. This is a 'rule of thumb' which does not explicitly consider expected arrival delay, cost of fuel and cost of delay.
- Level 1: DCI is reassessed at the top of the climb (TOC), taking into account the estimated arrival delay. The flight performs a potential delay recovery by comparing expected fuel and time costs, and chooses the least costly option. Note that delay at departure might not represent delay at arrival due to the existence of buffers.
- Level 2: The DCI assessment and WFP strategies are coupled via a unified cost function, and the optimal decision is made before take-off (e.g. waiting for (some) passengers and then speeding up). DCI is reassessed at TOC, with the additional possibility of slowing down in cases where the expected arrival time is more than 15 minutes ahead of the scheduled arrival time, in order to consider potential fuel savings.

The assessment of the passengers' status is always performed 5 minutes before the pushback-ready event. The fol-

lowing WFP strategies have been implemented in the model:

- Level 0: Based on the latest estimate of the at-gate time for each passenger, a flight decides to wait for any passenger with a premium ticket, who is expected to be at-gate no later than 15 minutes after the flight's expected pushback time;
- Level 1: A flight decides to wait for passengers weighing two types of estimated costs: waiting v. not-waiting cost. The waiting cost includes the cost of delaying a flight for an additional n minutes due to waiting for passengers, whereas the not-waiting cost includes all the costs of having to take care of passengers missing their connection. A flight chooses the wait time that minimises the total additional cost;
- Level 2: WFP is performed in conjunction with DCI using a unified cost function, as explained above.

III. Indicators

A. Standard indicators

We first consider a set of largely classical metrics, variously used in ATM, intended to capture fundamental statistics regarding delays and costs (both flights' and passengers'). Such metrics are a useful starting point in the model evaluation. Examples include:

- average departure and arrival delay of all/delayed flights (with delay $\geq X$, where $X \in \{0, 15, 60, 180\}$);
- number of cancelled and delayed flights (on departure or arrival);
- number of flights delayed due to reactionary delays; average reactionary delay;
- average passenger delay (with delay $\geq X$, where $X \in \{0, 15, 60, 180\}$;
- average cost of compensation, passenger rebooking, duty of care, excess fuel usage (w.r.t. planned fuel usage).

Statistics have been computed over 100 iterations of the runs of the model, per scenario. For each indicator, the average, the first and the third quantile of its distribution over the iterations are considered. All the cost indicators are expressed in euros.

B. Centrality

In a networked system, such as ATM, centrality is a measure of the 'importance' of a node in terms of its role in connecting the network. Since the definition of importance depends on the context, many different metrics have been proposed in the literature. In Domino, we consider the network of airports and flights, where flights represent links and are therefore present only in some time intervals. Studying the centrality of the nodes of such a network contributes to the understanding of the impact of the new mechanisms on the whole network. Two air traffic network types can be considered: the network formed by the scheduled flights and the network of actual flights (upon their execution). Centrality metrics can be used to consider questions such as whether the introduction of a new mechanism makes the system (or parts thereof) more robust, in the sense that the actual and scheduled centralities remain more similar. This would, in fact, indicate that the mechanism mitigates disruptions in network connectivity. Existing centrality metrics, however, are not suitable for such comparisons [5]. 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 network is temporal as links (flights) appear and disappear, and connections between them are possible only if the corresponding links are in the right order. The network is multilayer because flights belong to different alliances (or single airlines), each constituting a different layer [9].

We therefore proposed two new tailored centrality metrics. Inspired by the Katz centrality [10], we consider that the centrality of a node depends on the number of itineraries in the network having that node as origin (outgoing centrality) or as destination (incoming centrality). Different from the static version of the metric, these itineraries (or 'walks') must be time-respecting and also account for the fact that travelling through a link takes non-zero time, i.e., only itineraries that can really be travelled should be counted (including airport transfers, for multi-leg trips). In Katz centrality, itineraries are weighted according to their length, with longer itineraries contributing less to the centrality. Two different ways of weighting itineraries are proposed.

First, 'Trip centrality' [9]: an itinerary of n legs outgoing from (incoming to) an airport contributes α^n to the outgoing (incoming) centrality of that airport, where $\alpha < 1$ so that longer itineraries contribute less. Additionally, if the itinerary comprises flights of different airlines, it contributes less by a factor ε^m, where m is the number of changes of layer through the itineraries, and $\varepsilon < 1$. This accounts for the fact that multi-airline itineraries are less used by passengers. If $\varepsilon = 0$, these itineraries are not counted at all. The itineraries considered by trip centrality are therefore all the possible itineraries that could be used by passengers (because they are temporally feasible), weighted more when they have fewer legs and fewer changes of airline.

Second, 'Passenger centrality': each itinerary contributes to the outgoing or incoming centrality of an airport an amount which corresponds to the number of passengers on that itinerary. Therefore, the outgoing passenger centrality of an airport corresponds to the number of passengers that depart from that airport (either as their first departure or taking a flight connection there) and are directed to another destination, either with a direct flight or with connections. The incoming centrality of an airport corresponds to the number of passengers that land in that airport, either as their final destination or with a further connection.

The damage to the network connectivity due to delays and cancellations can be estimated as the loss of centrality between the scheduled and the actual, executed network. For trip centrality, the centrality in the actual network is computed by using the actual network structure, which accounts for the delays and cancellations (see [9] for details). An airport's centrality in the actual network is therefore always smaller than its centrality in the scheduled network. The loss of outgoing trip centrality of an airport measures the loss of

potential outgoing itineraries that are not feasible anymore, therefore quantifying the decrease in the potential to access the rest of the network from that airport. For passenger centrality, in the actual network only passengers that reach their destination using their scheduled itinerary are counted. The actual outgoing passenger centrality of an airport corresponds to the number of passengers counted in the scheduled outgoing passenger centrality, who manage to follow their scheduled itinerary.

If, for example, N incoming passengers miss their connection in airport i, and are rebooked to another outgoing flight, airport i will have a loss of outgoing centrality amounting to N. The same loss would apply if N passengers depart late from i and miss their next connection at another airport. Therefore, the loss of outgoing passenger centrality of an airport accounts both for the passengers that experience a disruption in that airport and for those that experience problems downstream. This is different from the loss of outgoing trip centrality, which does not account for missed potential connections in the airport itself. The actual, incoming passenger centrality of an airport, corresponds to the number of passengers that were counted in the scheduled incoming passenger centrality and that manage to follow their scheduled itinerary up to that airport. Therefore, the loss of incoming passenger centrality can be interpreted as the damage to airport i caused by issues upstream. For both types of centrality, incoming and outgoing, we look at the average centrality loss to the entire network. Note that for trip centrality, when the centrality loss is averaged over the entire network, the loss of incoming centrality equals exactly the loss of outgoing centrality. In fact, each loss of outgoing centrality corresponds to an equal loss of incoming centrality at another airport. Therefore, in this case, we will refer to it as 'average trip centrality loss'.

C. Causality

In the ATM system, delays and congestion states propagate through the system due to the interactions between flights and the environment. Causality metrics aim to identify the channels of delay propagation, thus revealing which nodes in the network are facilitating the spreading process, in particular by forming subsystems working as amplifying feedback for delay propagation.

In statistics, detecting a (directional) causal relationship between two random variables is equivalent to assessing whether the information on the past states of one variable helps in forecasting the future state of the other. We here consider two causality metrics that have been proposed in the literature, namely Granger causality in mean [11] and Granger causality in tail [12]. The first causality metric evaluates the forecasting performance on average, thus weighting equally both 'small' and 'large' values in assessing the statistical significance of the past information of one state in forecasting the other. Sometimes, however, we are interested in restricting the causality analysis only on the dependence between 'large' states. The proposition is that departure delays that are small with respect to flight time are probably not highly relevant, as they are typically easily absorbed by buffers. Hence, the second causality metric focuses only on the prediction of more 'extreme' events, where 'extreme' refers to events which are less likely to be observed (in probabilistic terms), e.g., high congestion. Specifically, given two random variables X and Y, whose states at different times are captured by two time series, we say that:

- Y 'Granger-causes in mean' X if we reject at some confidence level the null hypothesis that the past values of Y do not provide statistically significant information about future values of X by assuming the linear VAR(p) process as the predictive model (see [11] for further details on the implementation of the method);
- Y 'Granger-causes in tail' X if we reject at some confidence level the null hypothesis that the past *extreme* values of Y, defined as states falling in the (right) tail of the distribution[1], do not provide statistically significant information about future *extreme* states of X (for further information see [12]).

In the context of the Domino project, we consider the network of airports and flights, where airports represent the nodes of the network and a (directional) link is present if there exists a causal relationship between two nodes: each node described by the state of delay of the airport, i.e., the average delay[2] of flights taking off from that airport within one-hour time windows[3]. Given N airports, the network of causal relationships is built by applying the Granger causality test ('in mean' or 'in tail') to all the possible $N(N-1)$ pairs of airports. Hence, a correction to compensate for the number of tests needs to be considered [13]: we use the Bonferroni correction, thus setting the significance level of each test as equal to $\frac{5\%}{N(N-1)}$ with $N = 255$. A similar approach has been recently considered by [5], [14].

Studying the topology of the network of causal relationships in the ATM system is crucial to understanding the dynamics of delay propagation and to investigate whether the introduction of the mechanisms represents an improvement, e.g. by disrupting some propagation channels. Hence, standard topological network metrics can then be extracted from the network of causal relationships, ranging from link density, reciprocity, clustering, and so on. Here, we consider: (i) *link density*, capturing the average level of causality in the process of delay propagation; and, (ii) the number of *feedback triplets*[4], representing subsystems of three airports where delay propagates in a circle, thus resulting in the feedback dynamics of delay amplification. A (new) ATM mechanism that tends either to decrease the level of causality or to disrupt such feedback effects, would represent an improvement for the ATM system,

[1] Thus, we can consider the time series of binary variables, which are 1 if the state is extreme, 0 otherwise.

[2] Here, flight delay is defined as the difference between actual and scheduled departure times.

[3] When no departing flights are present in the one-hour interval, we define the state of delay as equal to zero.

[4] A feedback triplet is a subgraph of three nodes, A, B, and C, where A causes B, B causes C, and C causes A.

thus we can quantify the impact of the 4DTA mechanism at the network level also by measuring the percentage changes of these network metrics from the baseline to the 4DTA scenario. However, the value of a network metric such as the number of feedback triplets depends on the link density of the network, which may change from one scenario to another. For a fair comparison, we consider the 'over-expression' of such a metric, where over-expression is defined as the ratio of the observed value of a network metric to its expected value in the random case of the Erdös-Renyi model with the same link density. Not all flight delays have the same impact on the ATM system in terms of costs, e.g. delays for connecting flights might generate higher costs. Within our simulation framework, we can explicitly assess the dynamics of cost propagation by studying the causality network built with the state of the 'cost of delay' of an airport, which is defined as the average cost of delay (at any phase, from departure to arrival) of all flights departing from that airport, within a one-hour time window. Hence, a similar causality analysis can be performed, but with a focus on the propagation of *costs* within the ATM system. Since the 4DTA mechanism aims to reduce the costs each airline incurs, by means of a dynamic cost computation during the tactical phase, the study of cost propagation may reveal patterns not observed for delay propagation, because of the supra-linear dependence of the cost of delay on the simple delay magnitude [6].

IV. Results

In this section, we first illustrate how the model behaves in different situations, before exploring the metrics associated with various stakeholders and concepts.

A. Decision process of agents

When flights are deciding whether to modify their cruise speed, they estimate the amount of expected delay, and the fuel and delay costs. This information is recorded in the model allowing us to assess the decision process of the agents.

Figure 1 presents the amount of delay (in minutes) that flights decide to recover in the three levels, both in the baseline and stressed scenarios. In Level 0 the focus is solely on the amount of delay. This leads to a 'greedy' behaviour with flights trying to recover a high percentage of the maximum possible recoverable delay. Once the cost of delay and fuel is explicitly considered, a more conservative approach is observed, in Level 1. When the possibility to reduce speed is introduced in Level 2, a more interesting behaviour emerges. In the baseline, as the system is not under high stress, many flights consider the possibility to 'generate' delay in order to save fuel at the TOC assessment[5]. In the stressed case, as more delay is present in the system, its recovery becomes a higher priority as the associated costs increase. Note that for Level 2, the information presented includes the decisions considered when assessing the WFP and is updated at TOC. If the information

[5]Note that only flights that have an expected arrival time earlier than 15 minutes with respect to their schedule have the possibility to slow down. They are therefore transferring buffer to fuel savings by reducing their cruise speed.

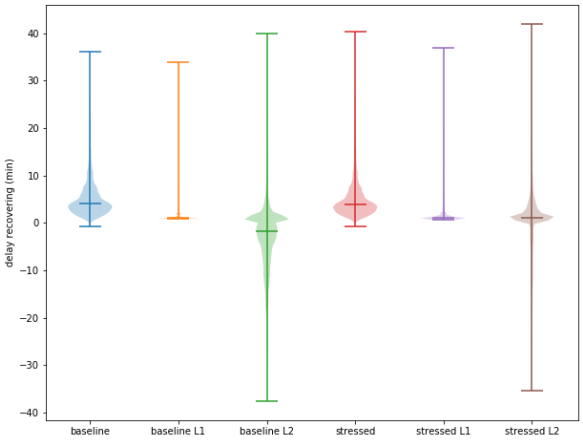

Figure 1. Expected amount of delay to be recovered by the flights under the different scenarios.

is disaggregated, it is possible to observe how some flights decide to wait for passengers and then recover all the delay, but then readjust the amount of delay to be recovered at TOC once more up-to-date information on expected arrival delay and costs are available. This might drive a lower WFP as the flight is able to estimate that this waiting might represent a high fuel cost to recover the introduced extra delay.

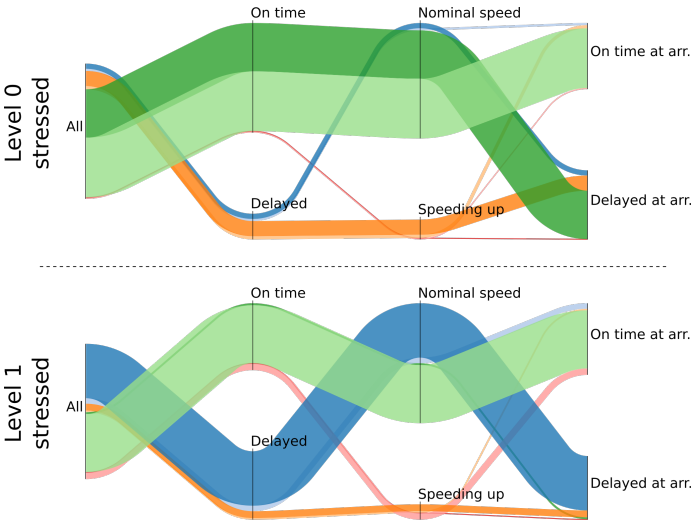

Figure 2. Flights' decisions and delay experienced for Level 0 and 1, in stressed scenarios.

Figure 2 presents the percentage of flights for the stressed cases of Level 0 and Level 1 in different categories: expected or actually 'on time' (with delay $\leq$ 20 min), expected or actually 'delayed' (with delay $>$ 20 min), maintaining their 'nominal speed', or 'speeding up'. The percentage of flights deciding to speed up is higher in Level 0 than in Level 1: the latter is considering the expected cost, not only the amount of delay. Similarly, there is no economic benefit in speeding up many of the delayed flights (in some cases it will generate

higher costs with limited savings). Therefore, in Level 1, they are maintained at their nominal speed. This leads to overall higher magnitudes of delay and reactionary delay in Level 1, but with lower total costs. Moreover, the number of flights delayed at arrival are similar in both cases. There are more flights expected to arrive with delay also due to WFP, which might increase the overall delay but reduces the total cost.
Around half of the flights that speed up in Level 1 are expected to have less than 20 minutes of delay. This speed variation ensures that they are maintained in that category. In Level 0, on the contrary, flights with small delays are maintained at their nominal speed, but then increase their delay during the flight, arriving with higher delays and costs at their destination. Overall, in Level 1, most of flights that expect to arrive with delay at TOC actually arrive delayed, while the mechanism helps to maintain on-time flights that expect to arrive with small delays (less than 20 minutes). Level 0 targets high delays: therefore, flights with small delays might end up increasing their delay and costs, while a large number of delayed flights speed up, but remain with a high delay. Level 1 considers more carefully which flights to speed up, as this comes at a (very) high fuel cost and uses WFP, increasing the departure delay, but reducing the total cost (as expensive missed connections are reduced). This is consistent with previous results, e.g. in [4]: WFP is important when the cost of delay is properly considered, as recovering delay by speeding up is usually (very) expensive.

B. Impact on flight-centric metrics

Figure 3 shows the effect of the implementation of the 4DTA mechanism on various airline metrics. It represents the relative difference, in percentage terms, for Level 1 and Level 2, for their baseline and stressed scenarios. Particularly for Level 1, flights that are already delayed are (somewhat) further delayed when stressed, as shown by the orange bars. Overall, in Level 1, airlines do not increase speed very much (as the cost of fuel is considered). For Level 2, in the baseline situation, flight arrival 'delay'[6] is as high as 12.5%. This takes place mainly for the *negative* delays, i.e. for flights that arrive before schedule. Also in Level 2, the airline slows down early flights, since they do not benefit from an earlier arrival, and saves fuel. Level 2 thus saves fuel *and*, complementarily, reduces reactionary delay (relative to Level 1).

The stressed situation is different. Indeed, Level 0 is using high amounts of fuel, since its focus is solely on reducing delay. In Level 1, the delays are still significantly larger, and the positive delays are higher than in the non-stressed situation, on average. In this case, the airline is more aggressive and tries to protect connections (as also explained in the next section, see Figure 4) at the expense of high average delays: in particular, reactionary delays. Level 1 also takes the cost of fuel into consideration, which leads to fuel savings and contributes to the high delays. At Level 2, higher savings with respect to Level 0 are observed, than for the Level 1

[6]We use the term 'delay' in the broadest sense, to include negative delay - i.e. early flights.

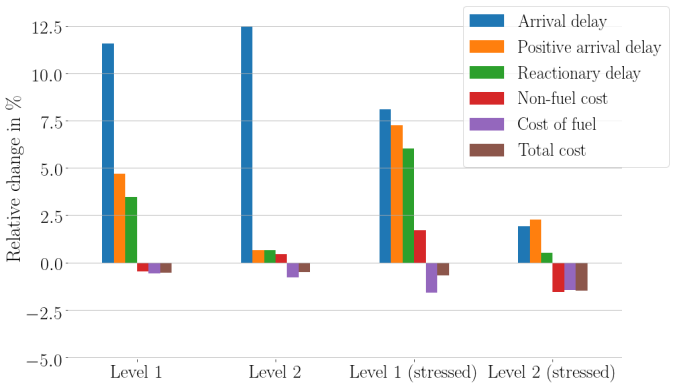

Figure 3. Changes in different metrics when various levels of 4DTA are implemented: average flight arrival delay, average arrival delay for positive delays only, average reactionary delay, average non-fuel cost, average cost of fuel, and average total cost. On the left, the results show the case where 4DTA is implemented in a normal, non-stressed situation (low delays). On the right, the stressed cases are displayed. All changes are computed with respect to their respective baseline, i.e. non-stressed on the left and stressed on the right.

implementation. This highlights the benefit of the conjoint implementation of WFP and DCI. Moreover, as the possibility to slow down is available for early arrivals, extra fuel can be saved for specific flights. WFP reduces the non-fuel associated costs, as connections are better protected. Not only are the costs decreased in this case, but also the average delay. In addition to the general trade-off between costs and delays, there can also be a trade-off between different *types* of costs, as highlighted by Level 1 in the stressed scenario.

C. Impact on passenger-centric metrics

Passenger delays tend to increase with the 4DTA mechanism: see Figure 4, which shows the percentage change in passenger-centric metrics in the scenarios, with respect to the baseline. When the mechanism is implemented at Level 1, the metrics evaluating the preservation of passenger itineraries (i.e., the number of passengers with a modified itinerary and the passenger centrality) show that passengers arrive more often at their destination according to their scheduled itinerary, than in the baseline. The average passenger centrality loss (incoming and outgoing) slightly decreases, both in the default and stressed scenarios. The percentage decrease with respect to the corresponding baseline is larger for the outgoing centrality and in the stressed case. The decrease is statistically significant (two-sample T-test, $p<0.05$) for the outgoing centrality loss (default and stressed) and for the incoming centrality loss in the stressed scenario. Given the general increase in delays and the fact that the increased loss of trip centrality tells us that potential connections (i.e., all connections that are possible, not only the ones actually used by passengers) are increasingly disrupted with respect to the baseline, the improvement from the passengers point of view must be due to the increased use of WFP and to the better evaluation of the possibility to speed up to preserve passenger connections. This is particularly visible in the average delays for connecting and non-connecting passengers, as shown in the figure. Arrival delays increase more for the latter than the former, indicating

that connecting passengers are protected to the detriment of others.

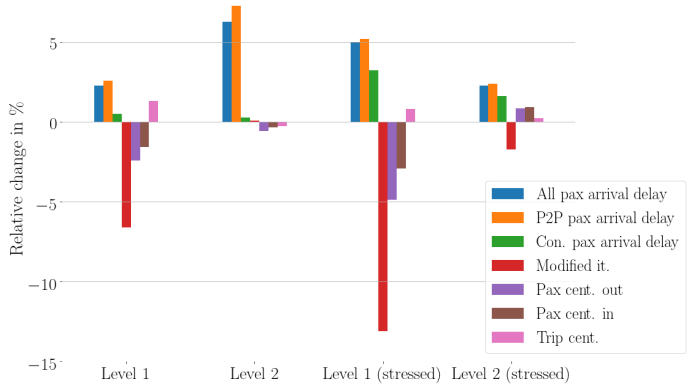

Figure 4. Changes in passenger-centric metrics in various scenarios: passenger arrival delay, arrival delay only for point-to-point (non-connecting) passengers, arrival delay for connecting passengers, number of cancelled legs for passengers (due to lost connections at the end of the day), number of modified passenger itineraries (at least one leg cancelled or using a different flight), average centrality loss. Three types of centralities are considered: passenger centrality (outgoing and incoming), and trip centrality (in this case, the average incoming and outgoing centralities coincide.

Note also the difference between Level 1 and Level 2 implementations in the stressed case. At Level 1, the airline uses WFP a lot, since it is making this decision before deciding to speed up, and thus does not balance the value of WFP against DCI. At Level 2, both decisions are made conjointly, which means that WFP is used less, as it is expensive in fuel to recover the wait by speeding up. Fewer itineraries are thus maintained, but the overall delay levels are lower.

D. Impact on network

TABLE I
LINK DENSITY OF THE CAUSALITY NETWORKS FOR THE BASELINE SCENARIO, FOR BOTH DEFAULT AND STRESSED CASES.

	Default	Stressed
GC in mean: delay	0.004	0.003
GC in mean: cost of delay	0.006	0.005
GC in tail: delay	0.16	0.34
GC in tail: cost of delay	0.23	0.34

Delays tend to increase when the 4DTA mechanism is implemented due to the more conservative strategy of delay recovery during cruise (see Section IV A). Nevertheless, 4DTA reduces the correlation between delays at any level of implementation, for both default and stressed cases, thus disrupting some channels of delay propagation. The number of these channels is captured by the link density of the causality network built for the baseline, shown in Table IV-D. The level of causality 'in tail', for both delay and the cost of delay, is much larger than the corresponding 'in mean' case, suggesting that restricting the causality analysis to the propagation of 'extreme' events is more informative. The disruption of the propagation channels can be measured by the variation of link density with respect to the baseline. The average level of Granger causality in mean (measured by link density) decreases from the baseline scenario to any other scenario with 4DTA implemented: see the top panel of Figure 5. However, this does not apply to extreme delays, which, on the contrary, become more correlated, as illustrated by the increasing link density of the Granger causality in tail network (with the exception of Level 2 for implementation in the default case). This negative impact of the 4DTA mechanism has an explanation in terms of the DCI computation: in the case of large delays, the optimal decision by airlines is not to use too much fuel, instead of speeding up, thus resulting in propagating more reactionary delay.

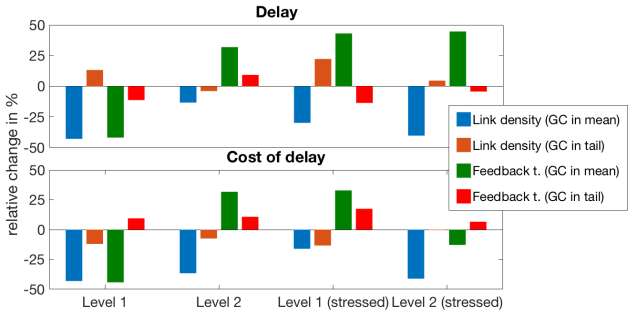

Figure 5. Percentage variations of the link density and the number of feedback triplets of the causality networks from the baseline scenario to the scenario with 4DTA implemented, at both Level 1 and Level 2, and for both the default and stressed cases. The causality analysis is shown for both the delay (top) and the cost of delay (bottom).

The opposite is observed when the causality analysis is applied to the propagation of the cost of delay: the level of causality in tail then decreases in any 4DTA scenario with respect to the baseline (see the bottom panel of Figure 5). This suggests that (at least some) propagated large delays have less impact in terms of costs, whereas some smaller (but more costly) delays, are reduced by the 4DTA mechanism, e.g. by preserving connections or avoiding the cost of compensation. Furthermore, the propagation channels of both delays and costs are quite different. To show this, we compare the two causality networks for the baseline scenario, built respectively for the delay and the cost of delay, by using the Jaccard index, which is a measure of similarity between two networks[7]. Considering Granger causality in mean, it is 0.50 (default; all values cited to 2 d.p.) and 0.40 (stressed), while it is 0.16 (default) and 0.30 (stressed) for the Granger causality in tail, thus revealing partial, or low, superposition of the propagation channels.

In Figure 5, we show the percentage variations of the over-expression of feedback triplets measured in the causality networks, from the baseline to the scenario where 4DTA is implemented at both Level 1 and Level 2. When considering Granger causality in mean for the cost of delay at Level 1

[7]The Jaccard index is defined as the ratio of the size of the intersection and the size of the union for two sets of links defining two networks to be compared.

(default), we note that the decreasing level of causality is coupled with an overall decrease of the network metric, thus suggesting the disruption of a number of propagation channels, which are involved in the feedback subsystems amplifying the propagation of costs. On the contrary, an increase of the over-expression is observed at Level 2. In this case, the feedback subsystems are less affected by the mechanism. Similar behaviour is observed for the dynamics of the propagation of 'extreme' costs, captured by Granger causality in tail. In the stressed scenarios, this is also observed 'in tail' for both Level 1 and Level 2. However, quite the opposite pattern (c.f. default case) emerges when considering Granger causality in mean[8]. In conclusion, the 4DTA mechanism is successful in reducing the propagation of the airline cost of delay. Nevertheless, the feedback processes at the network level, which have a negative impact on cost efficiency, are (largely) unaffected.

V. Conclusions and Further Work

Adjusting aircraft departure times and cruise speeds is a powerful tool in airline operations. We have explored various associated mechanisms, with a special emphasis on new rules based on an airline's network-wide cost minimisation. The combination of WFP and DCI increases the possibilities for delay mitigation, while taking into account passenger, fuel, and non-passenger costs. The Mercury agent-based model allows us to run a simulation of a single day of operations, tracking passengers and aircraft, allowing airlines to make complex decisions. The microscopic nature of the model supports the quantification of various metrics related to stakeholders, and assessing overall network performance under various scenarios.

Standard metrics show that relaxing arrival delay constraints can lead to a reduction of costs, due to savings in fuel and passenger-related costs. The best solution for airlines is achieved when optimising according to the joint objectives of WFP and DCI, and is particularly interesting in the stressed case. However, from the passenger point of view, applying these mechanisms does not necessarily have a positive impact. Most passengers experience a delay increase with respect to the baseline. Non-connecting passengers are more impacted, as airlines try to protect connections in exchange for average higher delay levels. Fewer passenger itineraries are then disrupted, however. Demonstrated by centrality metrics, the number of overall possible trips decreases, but the number of actual trips is improved, by the mechanisms. There is a clear trade-off between passenger-centric and flight-centric indicators (at least for some passengers). The existence of such trade-offs is important to highlight before deployment, and should nurture the debate on implementation priorities. Centrality metrics also show a clear decrease of the coupling of the network, in the sense that local disruptions affect fewer other parts of the network, when advanced mechanisms are implemented. This is an important systemic effect, and could lead to higher resilience and better capabilities for mitigating the propagation of local disruption.

From a project workshop with a cross-section of stakeholders, an important issue raised was the interpretation of some of the advanced metrics developed. Whilst centrality and causality are clearly helping to assess the network state, their meaning may not be immediately transparent. Even though the direction in which they should tend is clear (less causality, more centrality), the magnitude of such changes does not have an immediate interpretation. The consortium is currently working on illustrating these complex metrics and relating them to other, more standard ones. The team is also currently building an improved model, and the definition of more restricted scenarios, i.e., test cases, which will be used to assess the effectiveness of the mechanisms in specific cases, and the monetary value of schedule buffer. Attention will be paid to rendering the model user-friendly for operational stakeholders and system designers. With rewards for early movers now being progressed directly at the European Commission services level [2], assessments such as those presented in this paper, critically able to quantify the impacts of, and dependencies between, mechanisms introduced at different levels, should contribute to the implementation of the Airspace Architecture Study [1].

[8]We suspect this is because of larger random delays, thus resulting in larger statistical fluctuations affecting the network metrics, especially in the case of the Granger causality in mean network, which is characterised by a very low link density.

References

[1] SESAR Joint Undertaking, "A proposal for the future architecture of the European airspace," 2019.

[2] ——, "Future architecture of the European airspace, Transition Plan," 2019.

[3] A. Cook, H. Blom, F. Lillo, R. Mantegna, S. Micciché, D. Rivas, R. Vázquez, and M. Zanin, "Applying complexity science to air traffic management," *Journal of Air Transport Management*, vol. 42, pp. 149–158, 2015.

[4] L. Delgado, J. Martin, A. Blanch, and S. Cristóbal, "Hub operations delay recovery based on cost optimisation," in *Proceedings of the Sixth SESAR Innovation Days*, 2016.

[5] P. Mazzarisi, S. Zaoli, F. Lillo, L. Delgado, and G. Gurtner, "Towards new metrics assessing air traffic network interactions," *Proceedings of the Eighth SESAR Innovation Days*, 2018.

[6] A. Cook and G. Tanner, "European airline delay cost reference values, updated and extended values," (Version 4.1) https://www.eurocontrol.int/publications/european-airline-delay-cost-reference-values, 2015.

[7] Domino consortium, "Deliverable D5.2: Investigative case studies results," (pending formal approval), 2019.

[8] ——, "Deliverable D3.2: Investigative case description," https://s3.eu-central-1.amazonaws.com/innaxis-comm/DOMINO/Domino-D3.2-Investigative-case-studies-description.pdf, 2019.

[9] S. Zaoli, P. Mazzarisi, and F. Lillo, "Trip centrality: walking on a temporal multiplex with non-instantaneous link travel time," *Sci. Rep.*, vol. 9, no. 1, p. 10570, 2019.

[10] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

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

[12] 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.

[13] 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.

[14] 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.