Modeling Crew Itineraries and Delays in the National Air Transportation System

Keji Wei, Vikrant Vaze

Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755

{keji.wei.th@dartmouth.edu, vikrant.s.vaze@dartmouth.edu}

Abstract

We propose, optimize and validate a methodological framework for estimating the extent of the crew-propagated delays and disruptions (CPDD). We identify the factors that influence the extent of the CPDD, and incorporate them into a robust crew scheduling model. We develop a fast heuristic approach for solving the inverse of this robust crew scheduling problem to generate crew schedules that are similar to real-world crew scheduling samples. We develop a sequence of exact and heuristic techniques to quickly solve the forward problem within a small optimality gap for network sizes that are among the largest in robust crew scheduling literature. Computational results using four large real-world airline networks demonstrate that the crew schedules produced by our approach generate propagation patterns similar to those observed in the real world. Extensive out-of-sample validation tests indicate that the parameters calibrated for one network perform reasonably well for other networks. We provide new insights into the perceived tradeoff between planned costs and delays costs as reflected by actual airline crew schedules. Finally, we present a general approach to estimate the CPDD for any given network using our methodological framework under a variety of data availability scenarios.

Keywords: robust crew pairing; delay and disruption propagation; parameter estimation and validation.

1 Motivation

Flight delays and disruptions cost tens of billions of dollars annually to the world economy. The total cost of flight delays in the U.S. in 2007 was especially large, estimated to be approximately $31.2 billion (Ball et al., 2010). Over the first half of this decade, that is, from January 1st 2011 to December 31st 2015, only 79.21% of the domestic flights in the U.S. had a delay of 15 minutes or less (BTS, 2016). During the same period, 1.68% of the U.S. domestic flights were cancelled.

Determining the causes of flight delays and disruptions has been a topic of considerable interest among researchers and policymakers: The U.S. Department of Transportation (DOT) classifies flight delays into five main categories. They are Air Carrier Delays, Late Arriving Aircraft Delays, National Aviation System Delays, Extreme Weather Delays, and Security Delays. Between Jan 1st 2011 and Dec 31st 2015, Air Carrier Delays (defined as those within the control of the airline, such as those due to maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.) accounted for nearly 32% of all flight delays, while another 34% were attributed to Late Arriving Aircraft Delays (those due to late arrival of the previous flights using the same aircraft) and about 31% were classified as the National Aviation System Delays (defined as those due to non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.) (BTS, 2016). However, this public data lacks a dedicated category for flight delays due to the propagation of upstream crew delays and disruptions. Delayed or disrupted flights may generate delays and disruptions to subsequent flights because the crew for those flights is delayed, out of position, or unable to operate the scheduled flights without violating government regulations or collective bargaining agreements (CBAs). Presently, these delays (henceforth called as the *Crew-Propagated Delays and Disruptions* in this paper or CPDD for short) are considered to be a subset of the rather broad category of Air Carrier Delays. Accurate estimation of the CPDD is critical not only as a step toward fully

understanding the aviation system performance, but also for informing government policy and air carrier decisions related to airline crew scheduling.

There is yet another motivation for conducting this research. Public data sources lack information about not only the CPDD, but also the crew itineraries themselves. A prior study by Barnhart, Fearing and Vaze (2014) used a statistical approach to estimate passenger itineraries, passenger delays and disruptions. But similar estimates of crew itineraries or even a validated methodology to come up with such estimates is not available. Apart from aiding in future research studies such as those mentioned above, such estimates of crew itineraries would also be beneficial for assessing the full impact of any delay mitigation strategy being considered by airlines and/or government. In this paper, we develop a crew itinerary estimation methodology to generate a database of estimated crew itineraries that will enable accurate estimation of the CPDD consistent with their real-world values.

Note that we *do not* attempt to develop an approach to generate crew itineraries that are identical to the real-world airline crew itineraries in *every* possible aspect. Such objective would not only be extremely difficult to attain, but also likely cause overfitting issues based on the limited confidential data samples which we use for the estimation purposes. Instead, we develop a robust process to generate crew itineraries that are similar to the real-world airline crew itineraries in their potential for causing the crew-propagated delays and disruptions. There are possibly other, non-delay related, aspects of crew itineraries that could be of relevance for other purposes. However, the focus of this paper is to ensure that our process is accurate and stable in terms of the CPDD estimation.

Finally, we note that the present research project was originally motivated and funded by an aircraft delay modeling limitation faced by the U.S. Federal Aviation Administration's (FAA) Office of Performance Analysis. The FAA analyzes and forecasts, on a monthly basis, aircraft delays at the nation's major airports, with the objective of identifying airports with significant potential for delays months in advance, so that

appropriate actions may be taken to prevent or mitigate such delays. A discrete events simulation platform developed by the FAA for this purpose models aircraft-based delay propagation using public data on aircraft itineraries (BTS, 2016), but does not account for crew-based propagation effects due to the lack of crew itinerary data. This is believed to contribute to an underestimation of the propagated delays, and served as the motivation for conducting the research work presented in this paper.

1.1 Crew Pairing Optimization Problem

A crew pairing consists of a sequence of duties, where a duty is defined as the set of tasks to be performed by a crew member during a given day. Duties are connected by rest periods. Each duty is made up of a set of consecutive flights with some gaps between them. These gaps are called sit times. A pairing should begin and end at a *crew base* which is usually the domicile of a crew member. Both pairings and duties are subject to various regulations and contractual restrictions. Typically, these include the following.

- The total flying time within a duty cannot exceed an upper bound. There is also an upper bound on the total elapsed time within a duty.
- There is a lower bound on the sit time which guarantees that the crew has enough time to connect between two consecutive flights within a duty.
- The rest time between duties should be greater than or equal to a minimum rest time which ensures that the crew is sufficiently rested between duties.
- There is typically an upper limit on the number of duties within a pairing.

In addition to these rules, even when ignoring the operational cost considerations, crew pairings also have a highly non-linear pay structure. Note that the crew pay is commonly expressed in the units of hours of crew flying, which we will use throughout this paper. For a typical North American airline, the planned cost of a pairing p is the maximum of two terms: sum of the costs c_d of all its duties and a fixed fraction

(ζ) of the total time away from base ($TAFB_p$). Thus the planned cost of a pairing p (measured in the units of hours of crew flying) is given by:

$$c_p = \max\left\{\left(\sum_d c_d\right), \zeta * TAFB_p\right\} \qquad (1)$$

For each duty d, the planned cost (c_d) is the maximum of three terms, a minimum guaranteed pay (δ) per duty, flying time (fly_d) of the duty, and a fixed fraction (ε) of the duty elapsed time ($elapsed_d$). Parameters $\delta, \varepsilon, \zeta$ may vary across different carriers. Thus the planned cost of duty d (measured in the units of hours of crew flying) can be written as

$$c_d = \max\{\delta, fly_d, \varepsilon * elapsed_d\} \qquad (2)$$

The objective of the deterministic crew pairing problem is to minimize the planned crew cost and is usually modeled as a set partitioning problem (Barnhart and Vaze, 2015b). We denote the set of flights by F and the set of pairings by P. a_{ip} is 1 if pairing p contains flight i, and 0 otherwise. x_p is a binary decision variable which equals 1 if pairing p is chosen in the crew pairing solution, and 0 otherwise. Then, the crew pairing problem can be formulated (ignoring crew deadheads) as

$$Min \sum_{p \in P} c_p \, x_p$$

Subject to

$$\sum_{p \in P} a_{ip} x_p = 1, \qquad \forall i \in F, \qquad (3)$$

$$x_p \in \{0,1\}, \qquad \forall p \in P. \qquad (4)$$

1.2 **Literature Review**

As mentioned at the beginning of this section, from an application perspective, our work is motivated by the work of Barnhart, Fearing and Vaze (2014). Using one year of flight delays data and a one-quarter sample of confidential passenger booking data from an airline, they estimated passenger itinerary flows and developed insights into the factors that affect the performance of the U.S. National Air Transportation System from a passenger perspective. They developed a methodology to model historical travel and delays for passengers. From a methodological perspective however, our work is fundamentally different from theirs. While they used a statistical approach to estimate passenger itineraries, estimation of crew itineraries is considerably more complicated because of the complex rules governing what constitutes a legal crew itinerary. Thus a statistical estimation approach is unsuitable for our task. Also, while the number of possible passenger itineraries per day can be in thousands for a large airline, the number of legal crew itineraries is usually larger by several orders of magnitude, often making it very difficult or impossible to even enumerate all of them exhaustively.

Estimating delay propagation through crew connections is also much more complex than estimating the same through aircraft connections due to the more complex nature of crew work regulations than aircraft maintenance regulations (Barnhart and Vaze, 2015b). Several past studies on crew pairing optimization have tried to identify and capture one or more dimensions of a crew schedule that affect the extent of propagation. Broadly, these past studies can be divided into three main categories. First category of studies aims to incorporate one or more features that affect the ease of recovering the crew schedules after a disruption. Here, crew schedule recovery refers to the set of reactive measures available to an airline to bring its crew schedule back on track after a disruption and it typically includes alternatives such as delayed flight departures, crew swaps, reserve crews, flight cancellations, etc. Second category of studies aims to generate crew pairings that are difficult to get disrupted and/or have a low disruption

cost. Studies in the first and second categories deal exclusively with crew schedules without capturing the relationship between crew-based and aircraft-based propagation. The third category attempts to capture this interdependence.

Studies in the first category usually focus only on one or two specific factors that improve recovery potential. Shebalov and Klabjan (2006) maximize the number of move-up crews, wherein a move-up crew for a flight is a crew that is not actually assigned to that flight but can be feasibly and legally assigned to it. On the other hand, Gao, Johnson and Smith (2009) extend the fleet purity idea proposed by Smith and Johnson (2006) to crew base purity. The crew base purity idea restricts the number of crew bases allowed to serve each airport in order to increase the opportunities to find a move-up crew in crew recovery. Shebalov and Klabjan (2006) as well as Gao, Johnson and Smith (2009) capture the potential for crew swaps, which is an important dimension of crew recovery process, but do not explicitly capture the extent of delay propagation through crew connections.

Studies in the second category apply a variety of robust planning approaches to airline crew scheduling. Yen and Birge (2006) develop a two-stage stochastic programming model that implements a simplified recovery model for the second stage. Schaefer et al. (2005) adjust the cost of each crew pairing to include a combination of planned costs and a linear approximation of delay costs. The delay cost is assumed to be a function of four attributes, namely, 1) the sit time between consecutive flights within a duty, 2) the rest time between consecutive duties, 3) the total flying time in a duty, and 4) the total elapsed time in a duty. The rationale behind these choices is that the potential for the propagation of delays and disruptions is greater when the crew's sit and rest times are too short and when the per-duty flying and elapsed times are too long. Yen and Birge (2006) as well as Schaefer et al. (2005) account for the differences in the delay propagation potential of different crew pairings, but neither captures recovery actions such as crew swaps.

Because delays can propagate due to both late arriving/unavailable aircraft and late arriving/unavailable crew, there are many interdependencies between the effects of aircraft schedules and crew schedules on the propagation of delays and disruptions through the overall flight network. The aircraft scheduling and crew scheduling stages of the airline schedule planning process are conventionally solved in a sequential manner. However, recognizing the interdependencies between the two stages, in terms of the planned costs and delay propagation potential, some recent studies (such as Dunbar, Froyland and Wu (2012), and Weide, Ryan and Ehrgott (2009)) have developed integrated robust optimization models. Weide, Ryan and Ehrgott (2009) attempt to increase the buffer in crew connection times when the crew changes aircraft. Dunbar, Froyland and Wu (2012) focus explicitly on minimizing the delay costs while ignoring the planned costs. Cacchiani and Salazar-González (2016) solve an integrated fleet assignment, aircraft routing and crew pairing problem with a weighted average objective function that incorporates robustness solely as measured by the number of aircraft changes between successive flights in a crew itinerary. Mercier, Cordeau and Soumis (2005) also incorporate aircraft changes by the crew as a measure of robustness in their integrated aircraft routing and crew scheduling model. Ehrgott and Ryan (2002) and Tam et al. (2011) describe and evaluate a bi-criteria optimization approach to balance the planned crew costs and a single robustness measure which penalizes crew connections with aircraft changes and small crew sit times. Studies in this category usually emphasize aircraft changes and crew sit times but do not focus much on crew recovery potential, crew rest times, duty flying times, or duty elapsed times.

In summary, past research studies in airline crew scheduling have identified the various features of airline crew schedules that affect the crew-propagated delays and disruptions (CPDD). However, no prior study has combined these different features into a single optimization model. Additionally, while some past studies, such as Yen and Birge (2006), have attempted to incorporate the actual delay costs into the crew pairing optimization models, these models have been highly simplified due to computational tractability issues. Finally, and most importantly, all aforementioned studies have focused, implicitly or explicitly, on

finding an “optimal” crew schedule with respect to a known optimization formulation. The problem that we solve in this paper can be thought of as the inverse of this problem. Given an actual crew pairing sample, our goal is to reverse engineer the process used, and the problem solved, by the airlines to generate the crew pairings that the airlines actually used. This will enable us to generate similar crew pairings for other airlines, and/or other aircraft families, and/or other time periods than the ones for which the crew pairing data sample is available. It is common knowledge that the major airlines typically use advanced optimization solvers to generate crew pairings. Furthermore, in addition to minimizing planned costs, most airlines are known to directly or indirectly attempt to reduce delay and disruption costs as well. However, the exact models and algorithms used by a particular airline for crew pairing optimization are proprietary. Therefore, in this paper we reverse engineer airlines’ crew pairing generation process with the objective of generating pairings that are similar to the actual airline-generated crew pairings in terms of their CPDD potential.

1.3 Contributions and Outline

This paper makes four main contributions. First, we propose a comprehensive crew pairing optimization formulation that minimizes the combination of planned costs and the various features that make the crew schedules vulnerable to propagation of delays and disruptions. Second, we solve this model to near-optimality by combining known ideas such as branch-and-bound and delayed column generation as well as a sequence of new heuristic ideas developed by us. The sizes of the networks in the problem instances solved by us far exceed those solved in past studies on robust or recoverable crew pairing optimization. Third, we embed this crew pairing generation problem in an upper-level calibration framework wherein a parameterized crew pairing optimization problem is solved repeatedly by varying the parameters until the resulting crew pairings are *similar* to those used by the airlines. This upper-level calibration problem represents the inverse crew pairing generation problem mentioned in Section 1.2. We employ a local-

search heuristic for solving the upper-level calibration problem. Our algorithm is motivated by that of Schaefer et al. (2005) and borrows some features from theirs. Ours, however, is the first study to formulate and solve this inverse crew pairing generation problem. Finally, we generate and validate crew pairing solutions that are similar to those used by the airlines in the real world in terms of their potential for the crew-propagated delays and disruptions (CPDD). The out-of-sample testing results demonstrate the accuracy and stability of our modeling framework and algorithms. An important conclusion is that the ratio between the planned crew cost and approximate delay costs is found to be stable across airlines and aircraft types.

The rest of the paper is structured as follows. Section 2 describes our overall modeling approach and problem formulation. Section 3 describes the solution approach including the exact algorithms and heuristic ideas developed by us to solve this challenging problem. Section 4 describes our computational case studies in terms of data and pre-processing, and presents evidence of the computational tractability of our approach. Section 5 describes the calibration and validation results obtained from our series of computational experiments. Finally, Section 6 describes how to use our results for estimating the CPDD for any given network and discusses the main conclusions and the directions for future research.

2 Modeling Framework

Our objective is to generate crew itineraries that are similar to the real-world crew itineraries as measured by the extent of the crew-propagated delays and disruptions (CPDD). Therefore, we first need to develop an appropriate similarity metric for comparing two crew pairing solutions with each other, for any given flight network. Defining similarity directly based on the actual costs of propagated delays and disruptions is problematic for multiple reasons. Propagated delays and disruptions depend on not only the crew schedules but also the underlying root (i.e., non-propagated) delays and disruptions, as well as the operational recovery actions used by the airline. Exact recovery actions used by the airlines are typically

not public knowledge. Hence these are difficult to model accurately. Moreover, while planning the crew schedule, the airline itself is unaware of the exact set of root delays and disruptions that it will face on a given day of operations. For these reasons, accurate calculation of the CPDD costs is impossible. Instead, we measure the similarity between crew pairing solutions in terms of their *CPDD potential*. As explained in Section 1.2, the CPDD potential is a function of various features of a crew schedule. In Section 2.1, we classify these features into four categories and select six representative features for inclusion in our model. Then, in Section 2.2, we provide the mathematical formulation for the crew pairing optimization problem used as the basis of our calibration framework. Finally, in Section 2.3, we give a mathematical formulation for our calibration problem of minimizing the distance (i.e., maximizing the similarity) between the estimated and actual crew pairing solutions in terms of their CPDD potential as quantified by these six features.

2.1 Representative Features

In the absence of sufficient schedule buffers and recovery opportunities, delays and disruptions propagate to downstream flights leading to additional operating costs. Therefore, besides the planned crew costs, airlines often consider some of these buffers and/or recovery opportunities during crew scheduling to reduce these extra operational costs. There are a variety of mechanisms through which delays and disruptions affect downstream flights. When the sit time buffer (defined as the scheduled sit time minus the minimum required sit time) or the rest time buffer (defined as the scheduled rest time minus the minimum required rest time) between two consecutive flights is less than the arrival delay of the first flight, delay propagates to the second flight unless some recovery action, such as a crew swap, is able to prevent it. Thus the sit time buffers, the rest time buffers, and the crew recovery potential affect the crew-propagated delays and disruptions (CPDD). However, if these two flights are scheduled to be operated by the same aircraft, then this delay to the second flight would be unavoidable due to aircraft-based

propagation, irrespective of whether the crew is on-time. Note that, as per the DOT classification, delay propagation in such situations is classified as aircraft-based propagation causing the CPDD to be counted as zero to avoid double-counting. Thus, whether or not the crew travels with the same aircraft affects the CPDD. Finally, if flight delays result in violation of any of the crew duty regulations and/or CBA rules, such as the total flying time in a duty or the total elapsed time in a duty, then the later flight becomes inoperable by its scheduled crew, resulting in either a flight cancelation or some crew recovery action. Thus, the available buffers (defined as the maximum allowable value minus the scheduled value) in the total flying time or the total elapsed time in a duty also affect the CPDD. This discussion motivates our classification of features affecting the CPDD as well as our choice of the representative features.

We divide the features affecting the CPDD into four categories: Aircraft Change, Push-Back, Crew Legality, and Crew Swaps. This categorization highlights the variety of ways in which delays and disruptions can propagate through crew connections, and it facilitates any future revisions or extensions of the feature-set based on the methodologies we have developed.

2.1.1 Aircraft Change

We first motivate this category with an example. Consider two flights, Flight 1 and Flight 2, scheduled to be operated consecutively by the same crew within the same duty. If Flight 1 and Flight 2 are scheduled to be operated by the same aircraft, then irrespective of whether Flight 1's is delayed or not, by the time the aircraft is ready to operate Flight 2, the crew will typically be ready as well. Thus, there will be either no delay propagation or there will be some delay propagation attributed to the late arriving aircraft. However, no *crew-propagated* delay or disruption will occur. On the other hand, if Flight 1 and Flight 2 are scheduled to be operated by different aircraft, then to avoid delay propagation from Flight 1 to Flight 2, the crew on Flight 1 will need to exit that aircraft, reach the aircraft scheduled to operate Flight 2 and get ready to start operating it before the scheduled departure time of Flight 2. In this scenario, delay might

propagate through the crew connection. Therefore, if the crew needs to change aircraft between consecutive flights within a duty, there is a potential for the CPDD to occur. Hence whether or not the crew stays with the aircraft is an important factor affecting the CPDD (U.S. G.A.O, 2008). Therefore the number of times a crew switches aircraft within a duty is included as one of the representative features in our model.

2.1.2 Push-Back

When a flight's arrival is delayed, and the same crew within the same duty is scheduled to operate a subsequent flight, which is not scheduled to be operated by the same aircraft, a simple policy is to delay the subsequent flight until its scheduled crew is ready to operate it, regardless of how severe the delay is. We call this as the push-back strategy (Rosenberger et al., 2002). Similarly, when the arrival of the last flight in a crew duty (which is not the last duty in the crew pairing) is delayed, push-back strategy may be used to delay the departure of the first flight in the crew's next duty regardless of how severe the delay is and irrespective of whether or not the same aircraft is scheduled to operate the two flights. Note that, under the push-back strategy, delay propagates through the crew connection when the buffer in the crew sit time or the crew rest time exceeds the arrival delay of the first flight. Thus, the crew sit time buffer (between flights not scheduled to be operated by the same aircraft) and the crew rest time buffer are important factors affecting the CPDD potential, and hence both are used as representative features in our model.

2.1.3 Crew Legality

When developing crew schedules, airlines must adhere to FAA crew safety regulations and CBAs regarding the maximum flying time in a duty and the maximum elapsed time in a duty. For example, if FAA regulations limit a pilot to a maximum of 8 hours of flying time during a duty, and if the scheduled flying time is exactly 8 hours or just under 8 hours, then even a small delay to one of the earlier flights could

cause the actual flying time to exceed 8 hours. This would disallow the pilot to operate the last flight in the duty until the completion of a rest period, either leading to a flight schedule disruption such as cancellation or large delay, or triggering a crew recovery action such as a crew swap or the use of reserve crews. Note that, under this scenario, the CPDD occur even when the crew connection time buffer is large and/or the crew is not scheduled to change aircraft between flights. A similar argument holds when the scheduled elapsed time in a duty is equal to or just under the maximum allowable duty elapsed time. Thus, the buffer in the flying time and the elapsed time in a crew duty are important factors affecting the CPDD potential, and hence both of these are used in our model as representative features.

2.1.4 Crew Swaps

As mentioned in Section 1.2, crew schedule recovery actions include alternatives such as delayed flight departures, crew swaps, reserve crews, flight cancellations, etc. While delaying flight departures is the default alternative, under significant disruption events, it can result in very large and expensive delays. A flight cancellation cannot be done in isolation; typically it leads to cancellation of one or more other flights scheduled to be operated by the same aircraft and requires extensive amounts of passenger rebooking. Use of reserve crews is constrained by their availability and is typically an expensive strategy as well. A crew swap involves assigning a late arriving crew to operate a flight with a later departure time than its originally scheduled flight and instead using a different crew to operate the earlier flight. To allow a crew swap, the two swapped pairings must be from the same crew base, must end on the same day, and either crew must be qualified (in terms of equipment, route and airport certifications) to operate the subsequent flights in both pairings (Shebalov and Klabjan, 2006).

Compared to other crew recovery actions such as cancellations or reserve crews, swaps are typically less expensive, and therefore airlines find it beneficial to increase the crew-swapping opportunities. Gao, Johnson and Smith (2009) introduced the concept of crew base purity to restrict the number of crew bases

serving each airport. They found that improving the crew base purity can significantly increase crew-swapping opportunities and thus reduce the cost of crew recovery. They describe the idea of using an adjacency graph to quantify the extent of crew swapping potential. In an adjacency graph, airports are represented by nodes and the existence of an arc implies that there is at least one flight connecting the two nodes. For a specific airline's network, distance between two airports in an adjacency graph is defined as the minimum number of arcs that need to be traversed to go from one airport to the other. Crews serving airports that are more distant from the crew base, lead to fewer crew swapping opportunities and thus lower recovery potential. In our model, the number of times a crew visits an airport which is at a distance of 2 or more from its base is used as a feature representative of the crew swapping potential and hence representative of the CPDD potential. Other features indicating the crew recovery potential, such as, the number of reserve crew members available at various airports, could also be potentially included as representative features. However, we did not include them because of the lack of data on the availability of reserve crews in our dataset.

2.2 **Robust Crew Pairing Formulation**

The six representative features identified in Section 2.1 were integrated into a mathematical model that generates crew pairings that are similar to the real-world airline crew pairings. Our model formulation is motivated by the work of Schaefer et al. (2005), which used a penalty method for quantifying and maximizing the robustness of a crew schedule. They optimize the total expected operational cost of a crew pairing solution, which is defined as the sum of the planned cost and a linear function of four attributes of each crew pairing serving as proxy measures of its robustness. They assume that the aircraft are always available and hence no delay propagates through the aircraft connections. Also, the recovery method is assumed to be push-back only. Finally, they assume that the operational cost of a crew pairing solution is the sum of the operational costs of the individual chosen pairings, and that interaction between

crew pairings does not have an effect on the operational costs. We retain this last assumption, but partially relax the first and second assumption as follows. Similar to the four attributes chosen by Schaefer et al. (2005), we also include, in our representative features set, the scheduled sit time when crew changes planes, the scheduled rest time between duties, flying time in a duty, and elapsed time in a duty. Additionally, we also include as one of our representative features, the number of times a crew changes aircraft between successive flights within a duty. This provides a partial proxy for the additional delay that may result from the late arriving aircraft. Similarly, we also include the crew base purity, as measured by the number of times the crew arrives at an airport whose distance from the base is 2 or greater in the adjacency graph. We define these as the instances of violation of the crew base purity. Crew base purity provides a proxy for the crew recovery potential through crew swaps, as described in Section 2.1. Thus, we used the following six features.

Feature 1: Scheduled sit time when the crew changes aircraft.

Feature 2: Scheduled rest time between duties.

Feature 3: Flying time in a duty.

Feature 4: Elapsed time in a duty.

Feature 5: Number of crew base purity violations.

Feature 6: Number of aircraft changes by the crew within a duty.

Our method of incorporating these features into the crew pairing optimization model is an extension of the penalty method developed by Schaefer et al. (2005). For any pairing p, let c_p be its planned cost, and f_p be the penalty cost as a function of feature i. Then the total cost ($\bar{c}_p$) of pairing p (measured in the units of hours of crew flying) is defined as

$$\bar{c}_p = c_p + \sum_{i=1}^{6} f_p(i) \qquad (5)$$

For Features 3 and 4, as their scheduled value approaches the largest acceptable value, the potential for the crew-propagated delays and disruptions (CPDD) increases. For instance, the FAA requires that a crew must rest if it has already flown for 8 hours in a duty. As the scheduled flying time in a duty increases, the chances of this pairing becoming illegal during operation increase because of increased likelihood of violation of this rule. Similarly, for Features 1 and 2, as their scheduled value approaches the smallest acceptable value, the CPDD potential increases. For Feature i, let δ_i denote the relevant bound, that is, lower bound for Features 1 and 2 and upper bound for Features 3 and 4. For example, for Feature 2, δ_2 is the minimum rest time as allowed by the FAA regulations and the CBAs. For δ_2 of 10 hours, rest periods shorter than 10 hours in length are not permitted. Let $Count(i,p)$ be the number of times that Feature i occurs in pairing p, and let $V_{i,p}^{j}$ be the value of the j^{th} occurrence of Feature i in pairing p. For instance, if pairing p has three duties with elapsed times of lengths 10, 12, 5 hours respectively, then $Count(4,p) = 3$, $V_{4,p}^{1} = 10, V_{4,p}^{2} = 12$, and $V_{4,p}^{3} = 5$. We use parameters α_i to represent the maximum penalty, and β_i to represent the slope in Feature i 's penalty function. So, for the first four features, the function $f_p(i)$ is defined as:

$$f_p(i) = \sum_{j=1}^{Count(i,p)} \max\left(\alpha_i - \beta_i\left|V_{i,p}^{j} - \delta_i\right|, 0\right), \forall i \in \{1,2,3,4\} \qquad (6)$$

The form of function $f_p(i)$ described by Equation (6) is similar to that used by Schaefer et al. (2005). It assumes that $f_p(i)$ is additive across the effects of all occurrences of feature i in pairing p. Also, it assumes that, within a range, the effect of the value of the feature in each occurrence is linear and increases as the value of the feature gets increasingly closer to the relevant bound δ_i. At the bound, the effect has the maximum value α_i, because this leaves zero buffer in case of any prior delays or disruptions, and hence creates the maximum CPDD potential. Farthest away from the bound (i.e., at a distance of $\frac{\alpha_i}{\beta_i}$), the effect is zero. This is because large enough buffers almost fully eliminate any CPDD potential.

For defining the penalty function for Feature 5, we observe that if most airports are directly connected to the crew base, the airline has a greater potential for recovery by finding a move-up crew. Also, as for Feature 6, we observe that if most crews stay with the aircraft, then most of the delay propagation would be attributed to late arriving aircraft, rather than being counted as part of the CPDD. So we penalize the number of occurrences of crew changing aircraft and the number of occurrences of crew base purity violations. Let parameters γ_i, $i \in \{5,6\}$ denote the penalty weights for Features 5 and 6. With $Count(i,p)$ defined the same way as that for Features 1 through 4, the function $f_p(i)$ for Features 5 and 6 is defined as:

$$f_p(i) = \gamma_i * Count(i,p), \forall i \in \{5,6\} \quad (7)$$

Note that this expression is simpler than that for Features 1 through 4 because the number of aircraft changes and the number of crew base purity violations directly have an effect on the CPDD potential, as against the effects of Features 1 through 4 which depend on the difference between the feature value and the relevant bound. This results in a crew pairing optimization model given by

$$Min \sum_{\mathrm{p}\in\mathrm{P}} \left(c_p + \sum_{i=1}^{6} f_p(i) \right) x_p \quad (8)$$

Subject to

$$\sum_{p\in P} a_{ip} x_p = 1, \qquad \forall i \in F \quad (9)$$

$$x_p \in \{0,1\}, \qquad \forall p \in P. \quad (10)$$

2.3 Calibration Framework

There are several ways of conceptualizing our calibration problem. Given the optimization model (8-10), we could consider the calibration problem as one of estimating the parameters $\alpha_i, i \in \{1,2,3,4\}, \beta_i, i \in$

$\{1,2,3,4\}$ and $\gamma_i, i \in \{5,6\}$. Thus it is an inverse optimization problem. While an inverse linear optimization problem has been shown to be another linear optimization problem (Ahuja and Orlin, 2001), and hence is easy to solve, similar results do not exist for an inverse integer optimization problem (IIOP). Recently, Lamperski and Schaefer (2015) developed an approach to formulate the IIOP as an integer optimization problem with exponentially larger size. Others have proposed heuristic approaches for solving variants of the IIOP (see Duan and Wang, 2011; and Wang, 2013; for recent examples). However, these are computationally intensive and deal with only small-sized problems. A crew pairing optimization problem, on the other hand, typically consists of millions of (or more) variables, and is typically solved using complex, resource-intensive algorithms such as branch-and-price (Barnhart et al., 1998). Therefore, solving an inverse version of such a problem is extremely challenging for realistic problem sizes and no existing study has addressed this challenge successfully.

Alternatively, the calibration problem could also be considered a type of supervised machine learning problem where the goal is to generate crew pairing solutions similar to those in the labeled training data by learning the parameters $\alpha_i, i \in \{1,2,3,4\}, \beta_i, i \in \{1,2,3,4\}$ and $\gamma_i, i \in \{5,6\}$. This labeled training data, represented by a set of crew pairings, is in the form of a set of sequences of flights. This is in a non-standard structure for supervised machine learning, and the mechanism through which the parameters affect the labels is also very complicated. Thus, none of the typical supervised learning approaches, such as support vector machines or neural networks, to name a few, are directly applicable.

This discussion suggests that our calibration problem has several unique attributes, and is computationally much more expensive compared with what existing methods have been shown to solve. Therefore, we propose a new mathematical framework and a solution heuristic for solving this calibration problem. First, in this section we describe the framework and the relevant mathematical notation. Then, in Section 3, we describe the solution heuristic.

Let us denote the set of parameters by $PARAMS$. Thus, $PARAMS = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_5, \gamma_6\}$. Let $\hat{x}(PARAMS)$ be the crew pairing solution generated by solving the optimization model (8-10) for a given set of $PARAMS$ values, and let x be the real-world airline's scheduled crew pairing solution in our sample data. So, for each set of parameters, we have

$$\hat{x}(PARAMS) \in \underset{x_p \in \{0,1\}\, \forall p \in P}{\operatorname{argmin}} \left\{ \sum_{\mathrm{p}\in\mathrm{P}} \left(c_p + \sum_{i=1}^{6} f_p(i) \right) . x_p : \sum_{p \in P} a_{ip} x_p = 1, i \in F \right\} \qquad (11)$$

Also, let $F^{\hat{x}}(i) = \sum_{\mathrm{p}\in\mathrm{P}} f_p(i)\, \hat{x}_p$ and $F^{x}(i) = \sum_{\mathrm{p}\in\mathrm{P}} f_p(i)\, x_p$ be the values of the i^{th} components of the penalty functions corresponding to the crew pairing solutions $\hat{x}$ and x respectively. Then the calibration problem is formulated as follows:

$$min \sum_{i=1}^{6} \left| F^{\hat{x}}(i) - F^{x}(i) \right| \qquad (12)$$

Subject to

$$\hat{x} \in \underset{x_p \in \{0,1\}\, \forall p \in P}{\operatorname{argmin}} \left\{ \sum_{\mathrm{p}\in\mathrm{P}} \left(c_p + \sum_{i=1}^{6} f_p(i) \right) . x_p : \sum_{p \in P} a_{ip} x_p = 1, i \in F \right\} \qquad (13)$$

$$f_p(i) = \sum_{j=1}^{Count(i,p)} \max\left(\alpha_i - \beta_i \left| V_{i,p}^{j} - \delta_i \right|, 0 \right), i \in \{1,2,3,4\} \qquad (14)$$

$$f_p(i) = \gamma_i * Count(i,p), i \in \{5,6\} \qquad (15)$$

Note that this formulation minimizes the L1 norm of the difference between $F^{\hat{x}}(i)$ and $F^{x}(i)$. Alternatively, we could consider minimizing other norms (such as L2 norm) as well. Our computational experiments with L1 and L2 norms showed that these two alternative formulations did not lead to any significant changes in our results.

3 Solution Approach

In order to generate crew pairings that are similar to those scheduled by the airline, we need to solve the calibration optimization problem represented by (12-15). The similarity or closeness between the two crew pairing solutions provides a measure of the success of the calibration process. However, in order to truly assess the stability of this approach, we need to perform out-of-sample testing. As described in Section 4, we use one set of sample data to calibrate the parameters and then use the same calibrated values with another set of sample data (from a different airline, and/or different aircraft family, and/or different time period) to assess the stability of our approach. But before that, we need to develop a heuristic to solve this very difficult problem represented by (12-15). Note that the right hand side of constraint (13), in itself, is a very challenging problem for large network sizes. It is a type of robust crew pairing optimization problem, and no prior study in the literature has solved such problems of size as large as those of the networks used in this paper. Therefore, we develop and implement new heuristic approaches to solve the inverse of this already very difficult problem. In this section, we describe the solution approach. Then, in Section 4, we present our computational results.

We begin this section by describing, in Section 3.1, our overall heuristic for solving the calibration problem. This involves repeatedly solving instances of the model (8-10). Section 3.2 summarizes the solution approach for model (8-10), which itself includes repeatedly solving instances of the LP (linear programming) relaxation of this integer optimization problem. The solution to the LP relaxation of model (8-10) involves repeatedly solving instances of a sub-problem called the pricing problem. The process for solving this pricing problem is described in Section 3.3.

3.1 Local Search Heuristic for the Calibration Problem

We use a local search method for solving the optimization problem given by (12-15). It starts with all parameters in the set $PARAMS$ initialized to 0. It then varies the parameters corresponding to each

feature, one feature at a time, using a simple grid search to identify any opportunities for improving the objective function (12). The algorithm terminates when no better solution can be found in an iteration and returns the current best solution. Note that these grid-searches require us to examine various combinations of $PARAMS$ values to calculate the $\sum_{i=1}^{6}\left|F^{\hat{x}}(i) - F^{x}(i)\right|$ value. Examining each combination of $PARAMS$ values requires solving the robust crew pairing optimization problem given by (8-10). Next, we discuss the process for solving this problem.

3.2 Crew Pairing Solution Approach

Typically, the deterministic crew pairing optimization problem is solved by techniques such as branch-and-price (Barnhart et al., 1998), which combine ideas from the branch-and-bound algorithm for solving integer optimization problems with the delayed column generation ideas for solving large-scale linear optimization problems. The reader is referred to Kasirzadeh et al. (2015) for a detailed review of the state-of-the-art techniques in this area. Unlike previous studies, our goal is not just to solve the robust crew pairing problem once. Instead, its solution constitutes a sub-problem within our overall calibration optimization process described in Section 3.1. The overall calibration algorithm requires solving hundreds of these individual crew pairing optimization problems. Therefore, our computational performance requirements are far more stringent than those of most prior studies in the literature. We cannot afford to wait for several hours to solve the crew pairing optimization problem. Instead of using column generation at each node of the branch-and-bound tree, which is very time consuming, we use a heuristic strategy to solve this problem. As explained in Section 4, this strategy helps us in obtaining solutions that are within a small optimality gap. This strategy can be summarized as follows. It refers to two other algorithms, *Algorithm A* and *Algorithm B*, which are described in Section 3.3.

Heuristic Solution Strategy for the Robust Crew Pairing Optimization Problem

Step 1: Form the Restricted Master Problem (RMP) by including only a small subset of columns and relaxing the integrality constraints.

Step 2: Solve the RMP to find a set of dual variable values.

Step 3: Using the dual variables from Step 2, solve the pricing problem with Algorithm B to identify if one or more variables have negative reduced costs. If so, add *all* variables with negative reduced costs to RMP's column pool and go back to Step 2; else go to Step 4.

Step 4: Using the dual variables from Step 2, solve the pricing problem with Algorithm A to identify if one or more variables have negative reduced costs. If so, add all variables with negative reduced costs to RMP's column pool and go back to Step 2; else go to Step 5.

Step 5: Fix the largest fractional variable to 1 and check if an integer solution is obtained. If not, go back to Step 2; else stop.

This algorithm was developed after experimenting with various alternative heuristic ideas, and each step was chosen carefully based on the computational performance with and without it. Our computational experiments revealed that Step 5 helps by improving the computational performance substantially while increasing the optimality gap by very little or nothing. Also, we found that decomposing the pricing problem's solution process into two steps, i.e., using Algorithm B in Step 3 and Algorithm A in Step 4, was a vital part of the computational speedup that we achieved. Without this, we would not have been able to finish all our experiments in reasonable amounts of time to accomplish this research project. More details about this two-step approach are provided in Section 3.3.

3.3 Solution to the Pricing Problem

Researchers have proposed and implemented a variety of methods for solving the pricing problem. *Multi-Label Shortest Path* (MLSP) is a commonly used pricing algorithm in the crew pairing context (Desaulniers et al. 2005; Vance et al. 1997). Unlike the deterministic version, our robust crew pairing problem involves a more complicated objective function including the planned costs and six types of penalty costs. Furthermore, as explained in Section 4, our network size is the largest among all existing research studies addressing any variety of the robust crew pairing problem. Therefore, we cannot directly use an existing method to solve the problem to near optimality in a limited time. Therefore, we develop a new two-step approach to solve the pricing problem to optimality. This approach is presented in Appendix A.

4 Case Study

In this section, we apply the models presented in Section 2 and the solution methods presented in Section 3 (and the Appendices) to four networks from two airlines across multiple time periods. We use confidential airline data containing crew scheduling samples acquired from these two airlines to calibrate and validate our parameterized crew pairing models. The data sources and data preprocessing steps are described in Section 4.1, while the computational performance of our models is highlighted in Section 4.2.

4.1 Data Source and Data Preprocessing

We acquired crew schedule samples from one major regional carrier (RC) and one major network legacy carrier (NLC) in the U.S. The RC has a homogenous fleet consisting of only one fleet family and the data available to us spanned two full months, namely, March and April 2014. The NLC's operations consisted of several different fleet families. However, for our computational experiments we chose only the three largest networks, namely, those operated by A320, B737 and B757 aircraft types, because the others were much smaller in size. The NLC's crew scheduling data sample spanned one full year, from August 2013 to July 2014.

Aside from this confidential data, we also used the Airline On-Time Performance (AOTP) database from the BTS website (BTS, 2016) which contains on-time arrival data for domestic flights by all major U.S. carriers. Most importantly for our purposes, AOTP provides tail number for each flight, which is a key piece of information useful to track aircraft rotations in real-world airline schedules. Since our data is obtained from two separate sources, some data preprocessing steps, including data cleaning and merging, needed to be performed before using the data for model calibration and validation.

We obtained the true values of the planned crew cost parameters, such as δ, ε, and ζ, as well as the values of the lower limits on the crew sit times and the crew rest times, and the upper limits on the maximum duty flying time and the maximum duty elapsed time from both the airlines represented in our data samples. Finally, note that all our data is related to the cockpit (and not the cabin) crew schedules, which are more stringent in their regulations and hence are expected to be responsible for a majority of the crew-propagated delays and disruptions. So our analysis is restricted to the cockpit crew schedules only, and hence deals with a large part, but not all, of the crew-propagated delays and disruptions. Note that this is a limitation of the available data and not of our methodology which would be valid if we were to perform a similar analysis with the cabin crew scheduling data.

4.2 Computational Experiments

CPLEX 12.5 solver with its default settings is used to solve all the linear and integer optimization problems. An 8-thread / 4-core Intel® i7-X5600 CPU with 8GB RAM and Windows 7 Professional as the operating system was used for all computational experiments.

Table 1. Computational Performance of Our Heuristic

Network Size (Flights)	Pricing Approach	Root LP Lower Bound	Our Integer Solution	Gap	Solution Time (hours)
102	Algorithm A Only	398.36	398.37	0.025%	<0.1
	Algorithm B + Algorithm A	398.36	398.37	0.025%	<0.1

3300	Algorithm A Only	12760.43	12832.12	0.56%	10
	Algorithm B + Algorithm A	12760.43	12832.12	0.56%	2

For demonstrating the computational performance of our crew pairing optimization approach, we consider two networks: one with 102 flights and the other with 3300 flights. We first solve the root node LP relaxation to optimality to get a lower bound on the optimal objective function value. This is listed in the third column of Table 1. Then using the method described in Section 3.2, we obtain a feasible, but not necessarily optimal, solution of the integer optimization problem. Its objective function value is listed in the fourth column. Fifth column gives the gap between the values in the third and fourth columns by dividing the difference between the two by the value in the third column. Note that this gap gives an upper bound on the true optimality gap of our heuristically obtained solution. The last column gives the total runtime for obtaining the solution in the fourth column.

Second column lists the solution approach. We first list the performance of our overall heuristic using the exact SPPRC method (as described in Appendix A), i.e., without using Algorithm B. We also list the performance of our modified method, i.e., when using both Algorithm A and Algorithm B. Across all cases in Table 1, the gap was at most 0.56%. For the small network, Algorithm B does not help in speeding up because Algorithm A alone is sufficient to solve it within a few minutes. However, in case of the large network with 3300 flights, the combined use of Algorithm A and Algorithm B, as described in Section 3.2, significantly reduces the overall computational time from 10 hours to 2 hours. Similar improvements were observed in all our large network instances. This demonstrates the value of using our modified two-step pricing approach.

5 Calibration and Validation Results

5.1 Calibration Results

Table 2 lists the estimated parameters resulting from our calibration process as described in Section 4.1, while Table 3 lists the penalty function values corresponding to each feature.

Table 2. Parameter Results

Feature Type	Parameter	RC	NLC-A320	NLC-B737	NLC-B757
Type 1	α_1	1	0.3	0.3	0.8
	β_1	1.5	0.65	1	3
Type 2	α_2	0.5	0	0	1.5
	β_2	0.15	0	0	1.1
Type 3	α_3	0.4	1.1	3	1
	β_3	1.4	0.4	1.25	1
Type 4	α_4	2	1.65	2	3.8
	β_4	1.5	0.5	0.7	1.3
Type 5	γ_5	0	0.025	0.08	0
Type 6	γ_6	0.05	0.07	0.4	0

Tables 2 and 3 present results using four distinct networks, namely, regional carrier's complete network (*RC*) excluding the flights that were filtered out in pre-processing, and the network legacy carrier's networks using the A320 fleet family (*NLC-A320*), the B737 fleet family (*NLC*-B737), and the B757 fleet family (*NLC-B757*). The numbers of flights in the RC, NLC-A320, NLC-B737 and NLC-B757 networks were 2432, 1200, 1840, and 147, respectively. All experiments were conducted over a seven day time horizon and the maximum number of duties allowed in a single crew pairing was 4 in all cases.

Table 3. Penalty Function Values

Cost Type		Planned	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6	%age
RC	Airline Sample	4590.40	105.87	42.05	0	3.15	0	33.90	3.87%
	Without Calibration	4143.38	293.72	46.94	4.08	93.60	0	43.34	10.41%
	With Calibration	4244.99	91.23	40.66	0.41	8.44	0	34.80	3.97%
NLC-A320	Airline Sample	4800.45	5.26	0	15.91	19.39	3.42	28.07	1.48%
	Without Calibration	4515.51	13.05	0	62.24	148.29	5.30	37.45	5.57%
	With Calibration	4539.24	5.27	0	9.48	12.53	3.38	28.00	1.28%
NLC-B737	Airline Sample	7448.00	0.77	0	53.74	32.75	9.44	182.4	3.61%
	Without Calibration	6696.80	9.35	0	300.36	280.52	12.56	280.00	11.65%
	With Calibration	6773.94	2.05	0	15.77	12.21	10.88	175.20	3.09%
NLC-B757	Airline Sample	976.18	0.44	1.08	0.47	3.52	0	0	0.56%

	Without Calibration	925.08	0.41	2.17	2.28	22.25	0	0	2.85%
	With Calibration	927.09	0	0	0.77	6.26	0	0	0.75%

In addition to the penalty function values corresponding to each feature, Table 3 lists the planned cost values for comparison purposes. In the last column, the total penalty cost as a percentage of the planned cost is listed. For each network, there are three rows. All three rows provide the components of the objective function evaluated using the calibrated parameter values listed in Table 2. The first row uses the actual airline-provided crew pairings. The second row uses the crew pairings obtained by solving the crew pairing optimization problem by setting all parameters to 0. Finally, the third row uses the crew pairings obtained by solving the crew pairing optimization problem by setting all parameters to their calibrated values (listed in Table 2). Note that the calibration algorithm does not explicitly attempt to match the planned cost values, because our aim is to match the CPDD potential alone. Yet, for all four networks, the planned costs of the crew pairing solution generated by our approach are found to be closer to the actual airline-provided crew schedules with calibration than without calibration. Across networks and cost types, the cost values with calibration were found to be closer (in most cases significantly closer), than the cost values without calibration, to the airline-provided crew pairing solutions in 22 out of the 23 network-cost type combinations in Table 3. Note that we have excluded the network-cost type combinations where all three values are zeros, which happens in 5 instances. In Section 5.2, we provide a metric for an easy comparison of this degree of closeness in the form of a percentage error measure.

Tables 2 and 3 exhibit several differences between the four networks. Some of these differences reflect the differences in the crew pay and crew legality rules. For the NLC-B737 and NLC-A320 networks, crew pay does not depend on the time away from base (i.e., parameter ζ in (1) equals 0). As a result, there is no tradeoff associated with the length of the rest period. For networks with nonzero ζ values, having shorter rest periods can cause additional delay propagation while having longer rest periods can add to the planned crew costs. Absent this tradeoff, for the NLC-A320 and NLC-B737 networks, the optimization

simply sets the rest period lengths such that the Type 2 penalty function value is zero irrespective of the values of α_2 and β_2. For the RC network, we find that irrespective of the values of the Type 5 parameters α_5 and β_5, Type 5 penalty function value equals zero. Recall that Type 5 penalty function penalizes the number of crew-base purity violations. Because of the simple hub-and-spoke structure of the RC network, most crew travel from a hub to a spoke and back, and there isn't much opportunity for changing the Type 5 penalty cost by varying γ_5. Therefore, for the RC network, γ_5 value remains 0 even after calibration. Finally the NLC-B757 network is the smallest among the four, due to which crews usually don't have too many alternatives other than staying with the aircraft and the crews do not end up going more than a distance of 1 unit away from the crew base in the adjacency graph. This simplified structure of the network explains why both Type 5 and Type 6 parameters and the corresponding penalty function values are set to 0 for the NLC-B757 network.

Although these four flight networks vary in size, and although the absolute CPDD level cannot be directly compared across the four networks, the values in the last column of Table 3 range between 0.75% and 4% across all four networks, for the airline-provided crew schedules and also for the solution generated by our calibrated model. These numbers are much higher for the crew-schedules generated using the uncalibrated model. These results demonstrate that our approach generates crew schedules whose balance between planned and operational costs is similar to that of the actual crew-scheduling solution used by the airlines. Previous studies involving robust crew pairing optimization, such as Yen and Birge (2006), have emphasized the importance of finding the right tradeoff between the planned and operational costs. They test effects of different penalty parameters to control this tradeoff, but do not provide explicit insights into the right tradeoff values. Our results, for the first time, allow us to get a measure of the perceived balance between the planned costs and the penalty costs as reflected by the airlines' actual crew scheduling practices. Table 3 suggests that the right balance between the planned

and penalty costs across the four networks is in a relatively narrow range of 0.75% to 4% and is thus quite stable across airlines and aircraft families.

Unlike previous studies, such as Schaefer et al. (2005), Shebalov and Klabjan (2006), Gao, Johnson and Smith (2009) which focus on minimizing a subset of the factors affecting the CPDD, we use a more comprehensive approach by including a wider variety of factors. Unlike Yen and Birge (2006) who consider the total expected cost of future disruptions, our approach can give a separate ratio between the penalty cost corresponding to each feature of the operational cost and the planned cost. By allowing penalty costs of each component to be assessed separately, we get a clearer understanding of the relative importance of each component as perceived by the airlines.

Table 3 provides some preliminary evidence of the effectiveness and accuracy of our calibration framework. However, there are several shortcomings of using the in-sample penalty costs to assess the similarity of our solutions to the airline-provided crew schedules. First, this in-sample comparison has an inherent bias because we are using the same data samples to calibrate and to test the accuracy. We address this concern in Section 5.2 by presenting results of computational experiments where the parameter calibration is performed using one dataset and then other datasets are used to assess the out-of-sample accuracy of our approach. Second, we are measuring the closeness of the two solutions using penalty functions, which themselves depend on the calibrated parameter values. To make our comparisons more meaningful, we compare the distributions of the actual feature values in Section 5.3. Finally, all methods used by us for evaluating the accuracy of our approach depend on the features that we deem to be good proxies for the CPDD. While many of these were chosen and are well-supported by previous research studies, they are not likely to be precise measures of the CPDD. Therefore, a true test of the performance of our approach can only be conducted by comparing the actual CPDD. This concern is addressed in Section 6.1.

5.2 Out-of-Sample Validation Results

This section demonstrates the accuracy and stability of our results through out-of-sample validation. First, in Tables 4 through 7, we present the results where the calibration and validation datasets belong to two different time periods for the same airline and for the same fleet family. For the RC network, we choose March 2014 as the calibration set and April 2014 as the validation set. For the three NLC networks, we select the first week of one month from each quarter to represent flight schedules through a full year. Specifically, we use January 2014 data for calibration and perform validation using datasets from April 2014, July 2014 and October 2013. Additionally, February 2014 dataset is also used to perform validation for a scenario where the calibration and validation datasets are not too far apart in time from each other. The intent of this validation is to test the validity of using parameters calibrated using one time period to predict crew schedules for another time period for the same airline and the same aircraft family. If the results are found to be stable across time periods, then this allows us to use crew scheduling data samples from one period to estimate crew schedules for other periods and thus reduces our data requirements if we were to estimate the CPDD across long periods of time.

Let C_i, $i \in \{1, \dots, 6\}$, be the Type i penalty cost associated with the crew schedule generated by our approach, and $C_i^{Airline}$ be the Type i penalty cost associated with the corresponding airline-provided crew schedule. Then we define the Absolute Percentage Error (APE) as $\frac{|C_i - C_i^{Airline}|}{\sum_i^N C_i^{Airline}}$, where N is the total number of components of the penalty cost function, i.e., the total number of robustness features. Note that this is not a commonly used method of error representation, but is chosen because it offers certain advantages for our problem setting. First, the choice of denominator ($\sum_i^N C_i^{Airline}$) in the APE expression guarantees that we do not have issues related to division by zero. Contrast this choice of denominator with a more standard $C_i^{Airline}$ term as the denominator, which would have led to division-by-zero issues in many cases, such as the Type 3 error for RC network, as presented in Table 3. Second, since we use the

same denominator for all features, it is easy and meaningful to compare the percentage errors across the different features.

Tables 4 through 7 list the APEs for each feature and also the average and maximum values across features. The columns titled “Before” and “After” list the errors for crew pairing solutions generated using uncalibrated and calibrated parameters respectively. Note that as described in Section 5.1, in all cases the penalty function evaluation is performed using the calibrated parameters. Looking at the results presented in Tables 4 through 7, several observations can be made. Errors are substantially lower in most cases after calibration than before. The improvement is especially clear when looking at the average or maximum values of the APEs across types. Average and maximum APEs are reduced substantially by the calibration process and a reduction is observed across all calibration and validation datasets. In many cases the reduction is by one or more orders of magnitudes. The APEs are slightly lower in the out-of-sample validation datasets compared with the in-sample calibration datasets, especially for the RC network. However, the out-of-sample APEs are consistently reduced by the calibration process demonstrating the stability and effectiveness of our approach. Moreover, seasonality is not found to play a significant role in terms of the errors. The out-of-sample validation errors did not worsen and stayed stable as the time between the calibration and validation datasets increased from 1 month (for February) to 6 months (for July). The consistently lower error values with the calibrated parameters, as measured individually, using averages, or using maximum values, indicate that our approach produces crew pairings that are stable across time periods of up to several months.

Table 4. Cross-Time Period Validation Results: RC

Data	Calibration (Mar)		Validation (Apr)	
	Before	After	Before	After
Type 1	101.6%	7.9%	58.9%	28.8%
Type 2	2.6%	0.8%	1.9%	0.3%
Type 3	2.2%	0.2%	1.6%	0.1%
Type 4	48.9%	2.9%	26.2%	0.2%
Type 5	0	0	0	0

Type 6	5.1%	0.5%	2.9%	0.4%
Average	26.7%	2.0%	15.3%	5.0%
Maximum	101.6%	7.9%	58.9%	28.8%

Table 5. Cross-Time Period Validation Results: NLC-A320

Data	Calibration (Jan)		Validation (Feb)		Validation (Apr)		Validation (Jul)		Validation (Oct)	
	Before	After	Before	After	Before	After	Before	After	Before	After
Type 1	10.8%	0.0%	12.0%	3.1%	9.3%	2.5%	11.3%	2.6%	21.1%	5.5%
Type 2	0	0	0	0	0	0	0	0	0	0
Type 3	64.3%	8.9%	59.8%	4.0%	39.2 %	12.9%	40.8%	8.6%	70.9%	11.5%
Type 4	178.9%	9.5%	228.9%	8.2%	140.3%	31.0%	134.6%	23.1%	299.2%	0.3%
Type 5	2.6%	0.1%	1.9%	0.7%	3.0%	1.6%	3.2%	2.6%	4.4%	3.3%
Type 6	13.0%	0.1%	12.0%	1.5%	9.4%	1.9%	8.2%	3.7%	18.9%	5.5%
Average	44.9%	3.1%	52.4%	2.9%	33.5%	8.3%	33.0%	6.8%	69.1%	4.4%
Maximum	178.9%	9.5%	228.9%	8.18%	140.3%	31.0%	134.6%	23.1%	299.2%	11.5%

Table 6. Cross-Time Period Validation Results: NLC-B737

Data	Calibration (Jan)		Validation (Feb)		Validation (Apr)		Validation (Jul)		Validation (Oct)	
	Before	After	Before	After	Before	After	Before	After	Before	After
Type 1	3.1%	0.5%	3.9%	0.9%	2.0%	0.3%	2.6%	0.9%	3.1%	0.7%
Type 2	0	0	0	0	0	0	0	0	0	0
Type 3	88.4%	13.6%	76.3%	7.8%	48.9%	13.4%	68.1%	12.8%	93.8%	6.9%
Type 4	88.8%	7.4%	86.9%	5.2%	60.7%	10.3%	70.9%	10.5%	100.2%	8.2%
Type 5	1.1%	0.5%	1.1%	1.0%	2.0%	0.6%	2.3%	0.7%	0.4%	1.1%
Type 6	35.0%	2.6%	30.2%	14.1%	26.3%	7.4%	28.0%	6.8%	35.3%	18.0%
Average	36.1%	4.1%	33.1%	4.8%	23.3%	5.3%	28.7%	5.3%	38.8%	5.8%
Maximum	88.8%	13.6%	86.9%	14.1%	60.7%	13.4%	70.9%	12.8%	100.2%	18.0%

Table 7. Cross-Time Period Validation Results: NLC-B757

Data	Calibration (Jan)		Validation (Feb)		Validation (Apr)		Validation (Jul)		Validation (Oct)	
	Before	After	Before	After	Before	After	Before	After	Before	After
Type 1	0.5%	8.0%	0.0%	2.9%	0.0%	0.0%	3.7%	0.8%	3.8%	3.8%
Type 2	19.8%	19.6%	0.0%	0.0%	161.6%	0.0%	16.0%	0.8%	105.5%	10.2%
Type 3	32.8%	5.4%	14.1%	3.9%	54.7%	0.0%	17.4%	19.8%	20.3%	9.6%
Type 4	339.9%	49.7%	185.1%	92.0%	1080.8%	25.1%	309.9%	79.1%	321.7%	39.4%
Type 5	0	0	0	0	0	0	0	0	0	0
Type 6	0	0	0	0	0	0	0	0	0	0
Average	65.5%	13.8%	33.2%	16.5%	216.2%	4.2%	57.8%	16.8%	75.2%	10.5%
Maximum	339.9%	49.7%	185.1%	92.0%	1080.8%	25.1%	209.9%	79.1%	321.7%	39.4%

Tables 8 through 11 present the cross-validation results where the validation is performed on a dataset which belongs to a different airline, a different fleet family, and in some cases, a different time period compared with the calibration dataset. This constitutes an important test of our calibration approach because realistically we cannot expect crew scheduling samples to be available for all combinations of

airlines, fleet types and time periods. Instead, if we are able to access a small crew scheduling sample from one airline, for one fleet family, and for one time period, it is desirable to use that sample to calibrate parameters of a model that can then be used to generate crew schedules for other airline types, other fleet types and/or other time periods. Tables 8 through 11 present these cross-validation results where the four chosen combinations of networks and time periods are RC March 2014, NLC-B737 January 2014, NLC-B757 January 2014, and NLC-A320 January 2014. In each table, we use the parameter sets obtained by calibration over the networks listed in the top row.

Each table represents results of validation using a single network and time period combination specified in the table caption. The intent of this validation is to test whether our parameters calibrated for one combination of airline, fleet family and time period still perform well for other combinations of airlines, fleet families and time periods.

Table 8. Validation across Airline, Fleet Family and Time Period for the RC Network for Mar 2014

Data	RC (Calibration)		NLC-A320		NLC-B737		NLC-B757	
	Before	After	Before	After	Before	After	Before	After
Type 1	101.6%	7.9%	21.4%	6.7%	6.1%	0.6%	43.2%	1.1%
Type 2	2.6%	0.8%	22.0%	16.5%	0	0	0.8%	7.5%
Type 3	2.2%	0.2%	28.8%	11.6%	31.1%	16.8%	18.5%	1.8%
Type 4	48.9%	2.9%	50.8%	21.1%	26.4%	7.4%	149.5%	51.2%
Type 5	0	0	0.3%	1.1%	0.4%	2.3%	0	0
Type 6	5.1%	0.5%	6.5%	1.4%	15.7%	7.7%	0	0
Average	26.7%	2.1%	21.6%	9.7%	13.3%	5.8%	35.3%	10.3%
Maximum	101.6%	7.9%	50.8%	21.1%	31.1%	16.8%	149.5%	51.2%

Table 9. Validation across Airline, Fleet Family and Time Period for the NLC-A320 Network for Jan 2014

Data	NLC-A320 (Calibration)		RC		NLC-B737		NLC-B757	
	Before	After	Before	After	Before	After	Before	After
Type 1	10.8%	0.0%	36.9%	38.2%	2.4%	0.2%	34.1%	4.8%
Type 2	0	0	22.0%	16.5%	0	0	6.6%	6.3%
Type 3	64.3%	8.9%	1.1%	0.0%	52.6%	5.3%	29.7%	0.5%
Type 4	178.9%	9.5%	85.9%	4.0%	73.6%	3.0%	895.2%	52.4%
Type 5	2.6%	0.1%	0	0	2.9%	1.7%	0	0
Type 6	13.0%	0.1%	8.2%	0.6%	25.8%	16.6%	0	0
Average	44.9%	3.1%	25.7%	9.9%	26.2%	4.5%	160.9%	10.7%
Maximum	178.9%	9.5%	85.9%	38.2%	73.6%	16.6%	895.2%	52.4%

Table 10. Validation across Airline, Fleet Family and Time Period for the NLC-B737 Network for Jan 2014

Data	NLC-B737 (Calibration)		RC		NLC-A320		NLC-B757	
	Before	After	Before	After	Before	After	Before	After
Type 1	3.1%	0.5%	60.2%	28.9%	17.1%	7.4%	27.6%	2.9%
Type 2	0	0	7.2%	22.6%	0	0	38.6%	0.3%
Type 3	88.4%	13.6%	2.6%	0.3%	120.1%	13.4%	37.0%	3.3%
Type 4	88.8%	7.4%	118.7%	0.2%	242.9%	20.6%	685.7%	80.6%
Type 5	1.1%	0.5%	0	0	3.5%	3.6%	0	0
Type 6	35.0%	2.6%	12.1%	5.8%	44.0%	33.8%	0	0
Average	36.1%	4.1%	33.5%	9.6%	71.3%	13.1%	131.5%	14.5%
Maximum	88.8%	13.6%	118.7%	28.9%	242.9%	33.8%	685.7%	80.6%

Table 11. Validation across Airline, Fleet Family and Time Period for the NLC-B757 Network for Jan 2014

Data	NLC-B757 (Calibration)		RC		NLC-A320		NLC-B737	
	Before	After	Before	After	Before	After	Before	After
Type 1	0.5%	8.0%	21.4%	25.0%	2.5%	2.6%	0.0%	0.0%
Type 2	19.8%	19.6%	50.4%	19.6%	0	0	0	0
Type 3	32.8%	5.4%	3.1%	1.4%	69.4%	6.1%	70.4%	6.4%
Type 4	339.9%	49.7%	87.0%	33.5%	101.5%	16.5%	48.3%	0.1%
Type 5	0	0	0	0	2.3%	1.8%	3.2%	5.2%
Type 6	0	0	6.3%	0.9%	5.7%	2.5%	13.9%	4.0%
Average	65.5%	13.8%	28.0%	13.4%	30.2%	4.9%	22.6%	2.6%
Maximum	339.9%	49.7%	87.0%	33.5%	101.5%	16.5%	70.4%	6.4%

Tables 8 through 11 show that the average and maximum errors (APEs) after calibration are much smaller when compared to those before calibration for all combinations of the calibration and validation datasets. However, when compared with the calibration errors, the validation errors are typically larger. This is especially obvious in Table 8 where the calibration is performed using the RC network for March 2014 and the validation is performed using the three NLC networks for January 2014. This seems to suggest that the three NLC networks are more “similar” to each other in terms of their calibrated parameters than the similarity between NLC and RC networks. This is not surprising given that the RC network exhibits many differences in the network structure, schedules and flight durations when compared with the three NLC networks. Moreover, Tables 9 and 10 together suggest that the parameters for the NLC-A320 and NLC-B737 networks are especially similar to each other as reflected by their low cross-validation errors. This phenomenon can also be explained by the fact that A320 and B737 aircraft families are similar to each

other. They are both single aisle, twin-engine aircraft with similar seating capacity and range capabilities causing their flight networks to also look similar to each other.

Thus Tables 8 through 11 provide several interesting insights. First, they demonstrate that the out-of-sample validation errors are considerably lower using the calibrated than the uncalibrated parameters even when the calibration was performed using a crew scheduling sample from a different airline type and/or fleet family. However, we also note that the error reduction by using the calibrated rather than the uncalibrated parameters is greater when the calibration and validation datasets are more similar, in terms of airline type and fleet family. This suggests that, on the one hand, when estimating crew schedules for a given flight network, it is advisable to use a parameter set that has been calibrated using a flight network that shares as many of its attributes as possible. On the other hand though, using *any* set of calibrated parameters is still likely to be considerably better than using uncalibrated parameters. Even if the calibrated parameters are from a different time period, different airline and/or different fleet family, they improve the accuracy considerably compared with the uncalibrated parameters, i.e., compared with solving the deterministic crew scheduling problem. Thus, while it is advisable and beneficial to have a wide variety of airline crew schedule samples, our calibration approach enhances the degree of similarity of the generated crew pairing solution with the actual pairing solution used by the airline even when crew sampling data is relatively scarce.

5.3 Validating Crew Pairing Distributions

In this section, we perform additional validation of our results by directly comparing the distributions of the features that affect the crew-propagated delays and disruptions (CPDD) for our results against the distributions of those features for the airline-provided crew pairing solutions. Our goal is to ensure that the distributions of the features affecting the CPDD are similar between our solution and the airline-provided solution so that the two crew pairing solutions possess similar CPDD potential. This validation

approach is similar in spirit to that used by Barnhart, Fearing and Vaze (2014) to compare the distributions of features of passenger itinerary flows.

We consider the following distributions for validation purposes.

1. Distribution of the flying time in a duty.
2. Distribution of the elapsed time in a duty.
3. Distribution of the scheduled sit times.
4. Distribution of the scheduled rest times.

These correspond to Features 1 through 4 described in Section 2.2. Chi-squire statistic and the Kolmogorov-Smirnov statistic are two commonly used metrics for comparing two distributions to each other. The lower the values of these statistics, the more similar are the two distributions. Table 12 compares the distributions of these four features. For the RC network, the calibration is performed using the March 2014 dataset and the validation is performed using the April 2014 dataset while for the three NLC networks, the calibration is performed using the January 2014 dataset and the validation is performed using the February 2014 dataset. Note that we do not present the rest time distributions for the NLC-A320 and NLC-B737 networks for the reasons mentioned in Section 5.1. These results in Table 12 further reinforce our conclusion that the calibrated models generate crew-pairing solutions that are very similar to those provided by the airline in terms of the distributions of the CPDD potential, when tested on both in-sample and out-of-sample datasets. In almost all cases, the calibrated parameters yield a better fit to real-world distributions compared with the uncalibrated ones and in many cases the improvement is large.

Table 12. Validating Distributions of Crew Pairing Solution Features

Dataset	Feature	Chi-Square		Kolmogorov-Smirnov	
		Before	After	Before	After

		Statistic	p-value	Statistic	p-value	Statistic	p-value	Statistic	p-value
RC, Calibration (March 2014)	Flying Time	48.86	0	0.15	0.9852	0.67	0.03	0.33	>0.2
	Elapsed Time	94.19	0	36.22	0	0.40	>0.2	0.20	>0.2
	Sit Time	89.67	0	106.84	0	0.50	0.18	0.50	0.18
	Rest Time	6.95	0.0735	1.27	0.7363	0.33	>0.2	0.33	>0.2
RC, Validation (April 2014)	Flying Time	40.55	0	7.26	0.0641	0.67	0.03	0.33	>0.2
	Elapsed Time	83.15	0	3.58	0.3105	0.40	>0.2	0.20	>0.2
	Sit Time	157.16	0	7.64	0.89	0.75	<0.01	0.50	0.12
	Rest Time	6.63	0.0847	17.74	0.0005	0.33	>0.2	0.33	>0.2
NLC-A320, Calibration (January 2014)	Flying Time	73.47	0	1.25	0.741	0.75	<0.01	0.25	>0.2
	Elapsed Time	171.22	0	1.32	0.7244	0.75	<0.01	0.25	>0.2
	Sit Time	42.13	0	13.3	0.0099	0.40	>0.2	0.20	>0.2
NLC-A320, Validation (February 2014)	Flying Time	60.36	0	8.95	0.03	0.75	<0.01	0.25	>0.2
	Elapsed Time	160.34	0	10.75	0.0132	0.50	0.17	0.25	>0.2
	Sit Time	55.16	0	31.90	0	0.40	>0.2	0.20	>0.2
NLC-B737, Calibration (January 2014)	Flying Time	281.96	0	49.77	0	0.67	0.03	0.33	>0.2
	Elapsed Time	284.30	0	21.63	0.0001	0.60	0.07	0.40	>0.2
	Sit Time	81.07	0	27.98	0	0.75	<0.01	0.50	0.12
NLC-B737, Validation (February 2014)	Flying Time	113.51	0	20.09	0.0002	0.50	0.17	0.17	>0.2
	Elapsed Time	262.42	0	11.15	0.0109	0.60	0.07	0.20	>0.2
	Sit Time	54.66	0	33.15	0	0.25	>0.2	0.25	>0.2
NLC-B757, Calibration (January 2014)	Flying Time	3.45	0.3273	0	1	0.25	>0.2	0	>0.2
	Elapsed Time	4.85	0.1831	3.40	0.334	0.40	>0.2	0.20	>0.2
	Sit Time	3.67	0.4525	0.37	0.9849	0.50	0.12	0.25	>0.2
	Rest Time	3.92	0.2702	9.23	0.0264	0.33	>0.2	0.33	>0.2
NLC-B757, Validation (February 2014)	Flying Time	2.38	0.4974	2.31	0.5106	0.25	>0.2	0.25	>0.2
	Elapsed Time	3.32	0.3449	2.23	0.5261	0.20	>0.2	0.20	>0.2
	Sit Time	0.22	0.9944	2.25	0.6899	0.25	>0.2	0.25	>0.2
	Rest Time	7.10	0.0688	2.98	0.3947	0.33	>0.2	0.33	>0.2

6 Conclusion and Future Research

In this paper, we developed an approach to generate crew pairing solutions that are similar to the actual crew pairing solutions used by the airlines in the real world, in terms of their potential for the crew-propagated delays and disruptions (CPDD). As mentioned in Section 1, this work has at least three main types of applications. First, it is the first step toward estimating the extent to which delays and disruptions propagate through crew connections. Second, it allows us to assess and compare the effectiveness of various operational recovery strategies used by the airlines. Finally, it allows us to evaluate and compare the full impact of various candidate strategies for congestion and delay mitigation that are being

considered by the airlines, airports, air traffic control system, and the government. In following two paragraphs, we briefly describe how each of these objectives can be achieved using our results.

Table 2 provides four different sets of parameters representing four different airline networks. The robust crew pairing optimization model (8-10) can be solved for each of these four sets of parameters to come up with an estimated crew schedule for any given airline network of interest. As our results in Section 5 indicate, when picking the right set of parameters, it is advisable to choose a set that corresponds to a network which is the most similar (in terms of the airline type, fleet type and time period) to the network of interest. However, no matter which parameter set is picked, using calibrated parameters gives a far better fit than solving the deterministic crew pairing model in all cases. Alternatively, the calibration approach described in Sections 3 and 4 could be used to generate more suitable parameters in case a better-matching airline crew schedule sample is available.

Once the crew schedules are estimated, they can be used to estimate the CPDD. Note that, for accurately estimating delays in a historical dataset, some knowledge or assumption regarding the recovery strategies used by the airlines is necessary. For a given set of root delays and for a given operational recovery strategy, our crew schedules can be used to estimate the historical CPDD values in a relatively straightforward manner. Note that, in all cases, total propagated delays and disruptions should be measured by accounting for the propagation through aircraft connections as well as crew connections. However, the aircraft connections are publicly available and hence are not a bottleneck in this overall process.

It would be the first paper to use the inverse of the robust crew pairing generation problem in order to gain insights into the extent of the robustness of the real-world airline crew scheduling practices. The problem was formulated as one of learning the parameters of the robust optimization objective function using real-world airline crew scheduling samples. A heuristic solution approach was developed and

implemented. It involved solving the forward problem (the robust crew pairing problem) repeatedly to minimize a similarity measure between the solution of the robust crew pairing problem and the actual airline crew schedule samples by identifying the optimal set of objective function parameters. The forward problem minimizes the sum of the planned cost and the penalty costs which penalize the crew pairings for six different features that make them vulnerable to the propagation of delays and disruptions. A sequence of exact methods and heuristic ideas was used to solve this robust crew pairing problem to near-optimality. This allowed the overall parameter calibration problem to be solved in a reasonable amount of time.

Several new insights were obtained into the airline crew pairing generation process. First, compared with the crew pairings obtained by solving the deterministic crew pairing problem the calibrated parameters led to crew pairings that are considerably closer to the actual airline crew schedules in all our experiments. In most cases, the accuracy improvement was substantial. This suggests that airlines do take into account robustness or the potential for propagation of delays and disruptions when creating their crew schedules. Furthermore, we found that the crew pairings calibrated using four different airline networks performed similar to each other, and much better than the deterministic crew pairing solutions, in terms of their closeness to the actual crew schedules, even when the calibration and evaluation is not conducted on the same network. This suggests that the calibrated parameters are relatively stable. Thus, even in cases where the data available for model training is for a network somewhat dissimilar to the one of interest, it is better to use the calibrated parameters than the uncalibrated ones. However, for maximizing estimation accuracy, whenever possible, it advisable to use parameters calibrated with a network that is as similar to the one of interest as possible, in terms of airline type, fleet type and time period. Finally, this paper presented, for the first time in the literature, a measure of the tradeoff as perceived by the airlines between the crew salary costs and the costs of the crew-propagated delays and disruptions (CPDD) as reflected by the calibrated robust crew pairing objective functions. Across the four networks, the ratio of

the penalty costs (representing the costs of the CPDD) and the crew salary costs was found to lie between 0.5% and 4%. Note that this is inferred based on the crew pairings used by the airlines and not based on the actual costs of these delays and disruptions.

Moreover, these crew pairing estimates do give a starting point to estimate the CPDD, the important next step toward accurately estimating historical delay propagation is to develop an understanding of the crew recovery strategies used by the airlines in the real world. Once we have access to a historical sample of actual crew recovery actions, a framework similar to the one developed in this paper could be used to learn the airline crew recovery optimization process as well. This will be the next step in our research project.

Appendix A

This appendix details the two-step approach used to solve the pricing problem to optimality.

Algorithm A: This is a dominance algorithm with an exact implementation, similar to that described by Irnich and Desaulniers (2005), wherein only a path starting with the same crew base can dominate another path.

Algorithm B: This is a dominance algorithm with an implementation similar to that described by Irnich and Desaulniers (2005) except that a path starting with either the same or a different crew base can dominate another path.

The exact set of labels used by Algorithm A in our robust crew pairing implementation is as follows. Note that Algorithm B uses all but the last label listed below.

1. The number of duties covered so far by the path.

2. The total flying time so far in the current duty of the path.

3. The total elapsed time so far in the current duty of the path.

4. A constant multiple (ζ) of the total elapsed time so far in the path minus the sum of the dual contributions of all flights included so far in the path.

5. The total flying time so far in the current duty plus the sum of the costs of the previous duties in the path minus the sum of the dual contributions of all flights included so far in the path.

6. The minimum guaranteed pay of the current duty plus the sum of the costs of the previous duties in the path minus the sum of the dual contributions of all flights included so far in the path.

7. A constant multiple of (ε) of the total elapsed time so far in the current duty plus the sum of the costs of the previous duties in the path minus the sum of the dual contributions of all flights included so far in the path.

8. Crew base (the starting point) of the path.

Acknowledgments

This research is sponsored by the Federal Aviation Administration's National Center of Excellence for Aviation Operations Research (NEXTOR). We are also thankful to ILOG for providing CPLEX licenses to conduct our computational experiments.

Reference

AhmadBeygi B, Cohn A, Guan Y, Belobaba P (2008) Analysis of the potential for delay propagation in passenger airline networks. *J. Air Transport Management* **14**(5):221–236.

Ahuja RK, Orlin JB (2001) Inverse optimization. *Oper. Res.* **49**(5):771–783.

Barnhart C, Fearing D, Vaze V (2014) Modeling passenger travel and delays in the national air transportation system. *Oper. Res.* **62**(3): 580–601.

Barnhart C, Johnson E, Nemhauser G, Savelsbergh M, Vance P (1998) Branch-and-price: Column generation for solving huge integer programs. *Oper. Res.* **46**(3): 316–329.

Barnhart C, Vaze V. (2015a) Ch.10. Irregular Operations: Schedule Recovery and Robustness. Belobaba P, Odoni A, Barnhart C, eds. *The Global Airline Industry,* 2nd ed (John Wiley & Sons, West Sussex), 263-287.

Barnhart C, Vaze V. (2015b) Ch.8. Airline Schedule Optimization. Belobaba P, Odoni A, Barnhart C, eds. *The Global Airline Industry*, 2nd ed (John Wiley & Sons, West Sussex), 189-222.

Beatty R, Hsu R, Berry L, Rome J, (1998) Preliminary Evaluation of Flight Delay Propagation through an Airline Schedule. In: Proceedings of the 2nd USA/ Europe Air Traffic Management R &D Seminar, Orlando, Fl.

Bureau of Transportation Statistics (BTS) (2016) TranStats. On-Time Performance: http://www.transtats.bts.gov/Fields.asp?Table_ID=236

Cacchiani V, Salazar-González J-J (2016) Optimal Solutions to a Real-World Integrated Airline Scheduling Problem. *Transportation Sci*. Articles in Advance.

Desaulniers G, Desrosiers J, Solomon M (2005) *Column Generation* (Springer, New York).

Duan Z, Wang L (2011) Heuristic algorithms for the inverse mixed integer linear programming problem. *Journal of Global Optimization.* **51**(3):463-471.

Dunbar M, Froyland G, Wu C (2012) Robust airline schedule planning: Minimizing propagated delay in an integrated routing and crewing framework. *Transportation Sci*. **46**(2):204–216.

Engel F (1995) Summary over test runs, Internal report, Carmen Systems AB, Gothenburg, Sweden.

Ehrgott M, Ryan DM (2002) Constructing robust crew schedules with bicriteria optimization. *J. Multi-Criteria Decision Anal*. **11**(3):139–150.

Gao C, Johnson E, Smith B (2009) Integrated airline fleet and crew robust planning. *Transportation Sci.* **43**(1): 2–16.

IATA (International Air Transport Association) Annual Review 2015. Available at https://www.iata.org/about/Documents/iata-annual-review-2015.pdf.

Irnich S, Desaulniers G (2005) Shortest path problems with resource constraints. Desaulniers G, Desrosiers J, Solomon MM, eds. *Column Generation* (Springer, New York), 33–65.

Jacquillat A, Odoni A (2015) An Integrated Scheduling and Operations Approach to Airport Congestion Mitigation. Oper. Res. 63(6):1390-1410.

Klabjan D, Johnson E, Nemhauser G (2001) Solving large airline crew scheduling problems: Random pairing generation and strong branching. *Computational Optim Appl.* **20**(1):73–91.

Kasirzadeh, A, Saddoune, M, Soumis, F (2015) Airline crew scheduling: Models, algorithms, and data sets. *EURO Journal on Transportation and Logistics*. doi:10.1007/s13676-015-0080-x.

Lamperski J, Schaefer A (2015) A polyhedral characterization of the inverse-feasible region of a mixed-integer program. *Oper. Res. Lett.* **43**(6): 575-578.

Mercier A, Cordeau JF, Soumis F (2005) A computational study of Benders decomposition for the integrated aircraft routing and crew scheduling problem. *Comput. Oper. Res*. **32**:1451–1476

Petersen JD, Sölveling G, Clarke J-P, Johnson EL, Shebalov S (2012) An optimization approach to airline integrated recovery. *Transportation Sci*. **46**(4):482–500.

Pyrgiotis N, Malone KM, Odoni A (2013) Modeling delay propagation within an airport network. *Transportation Res. Part C* **27**:60–75.

Rosenberger J, Schaefer A, Goldsman D, Johnson E, Kleywegt A, Nemhauser G (2002) A stochastic model of airline operations. *Transportation Sci.* **36**:357–377.

Shebalov S, Klabjan D (2006) Robust airline crew pairing: Move-up crews. *Transportation Sci.* **40**(3): 300–312.

Schaefer A, Johnson E, Kleywegt A, Nemhauser G (2005) Airline crew scheduling under uncertainty. *Transportation Sci*. **39**(3):340–348.

Smith B, Johnson E (2006) Robust airline fleet assignment: Imposing station purity using station decomposition. *Transportation Sci*. **40**(4):497–516.

Swaroop P, Zou B, Ball MO, Hansen M (2012) Do more U.S. airports need slot controls? A welfare based approach to determine slot levels. *Transportation Res. Part B* **46**(9):1239–1259.

Tam B, Ehrgott M, Ryan D, Zakeri G (2011) A comparison of stochastic programming and bi-objective optimization approaches to robust airline crew scheduling. *OR Spectrum*. **33**(1):49-75.

U.S. G.A.O (2008) Commercial Aviation: Impact of Airline Crew Scheduling on Delays and Cancellations of Commercial Flights. GAO-08-1041R. United States.

Vance P, Atamturk A, Barnhart C, Gelman E, Johnson E, Krishna A, Mahidhara D, Nemhauser G (1997) A heuristic branch-and-price approach for the airline crew pairing problem. Technical Report LEC-97-06, Georgia Institute of Technology.

Vaze V, Barnhart C (2012) Modeling airline frequency competition for airport congestion mitigation. *Transportation Sci.* **46**(4): 512–535.

Wang L (2013) Branch-and-bound algorithms for the partial inverse mixed integer linear programming Problem. *Journal of Global Optimization.* **55**(3):491-506.

Wong J, Tsai S (2012) A survival model for flight delay propagation. *J. Air Transp. Manage*. **23**: 5–11.

Weide O, Ryan D, Ehrgott M (2010) An iterative approach to robust and integrated aircraft routing and crew scheduling. *Comput. Oper. Res.* **37**(5):833–844.

Xu N, Donohue G, Laskey K, Chen C-H (2005) Estimation of delay propagation in the national aviation system using Bayesian networks. *Air Traffic Management Res. Development Seminar*, Baltimore, 1–11.

Yen JW, Birge JR (2006) A stochastic programming approach to the airline crew scheduling problem. *Transportation Sci.* **40**(1):3–14.