Quantifying Delay Propagation in Airline Networks *

Liyu Dou[†] Jakub Kastl[‡] John Lazarev[§]

First version: September 2016
This version: January 2020

We develop a framework for quantifying delay propagation in airline networks. Using a large comprehensive data set on actual delays and a model-selection algorithm (elastic net) we estimate a weighted directed graph of delay propagation for each major airline in the US. We use these estimates to decompose the airline performance into "luck" and "ability." We find that luck may explain about 38% of the performance difference between Delta and American in our data. We further use these estimates to describe how network topology and other airline network characteristics (such as aircraft fleet heterogeneity) affect the expected delays. Finally, we propose a model of aircraft scheduler who decides which flights to delay and by how much. We then use the estimated model to evaluate counterfactual scenarios of investments in airport infrastructure in terms of their impact on delays.

Keywords: Airline Networks, Shock Propagation, Elastic Net **JEL Classification**: C5, L14, L93

*We thank Jan de Loecker, Aureo de Paula, Jeremy Fox, Bo Honoré, Eduardo Morales, Jim Powell, and seminar participants at 2018 CEPR IO meeting, Boston College, Harvard, Northwestern, Penn State, Princeton, Purdue, Rice, Rochester, Toronto, Toulouse, UC Berkeley, UT Austin and Yale for useful feedback. Kastl is grateful for the financial support of the NSF (SES-1352305) and the Sloan Foundation. All remaining errors are ours.

[†]School of Management and Economics, The Chinese University of Hong Kong, Shenzhen
[‡]Department of Economics, Princeton University, NBER and CEPR
[§]Department of Economics, University of Pennsylvania

1 Introduction

In this paper, we study how delays propagate in airline networks. Our first goal is to understand how exogenous shocks experienced by distinct parts of the airline network (e.g. morning snow in New York) affect the performance of the entire airline's network. The main challenge to our analysis is the fact that airlines *choose* which flights to delay. Our solution is to treat the observed day-to-day realizations of flight delays as an outcome of a (perhaps, very complicated) single-agent optimization problem of deciding which flights to delay and by how much. To implement this revealed preference approach, we develop a simple model that formally defines the data generating process. We do that to achieve several goals. First, there could be multiple reasons why delays of different flights throughout the network may be correlated. We show what sources of variation in the data identify the causal impact of an individual flight's delay on the performance of the entire network. Second, using the revealed preference approach, we recover the airline's perceived costs of an individual flight's delay from the observed joint distribution of delays. Flights that are relatively less expensive to delay get delayed longer and more often. We separate these delay costs into direct and indirect parts. Directly, delays inconvenience passengers on board of the current flight. We refer to these costs as "direct costs of delay." Delays also make maintaining downline on-time performance harder, as destination airports will have fewer planes, crews, or other resources than originally scheduled. We call these costs "indirect costs of delay." We show that the network structure of the problem allows us to identify these two types of costs separately. Finally, we use tools developed for our network analysis to answer three distinct economic questions.

First, we explore the reasons why some airlines systematically perform better overall than others. In our setting, the on-time performance of an airline is driven by two contributing factors: the distribution of exogenous shocks ("luck") and the properties of the airline's network that determine the shock propagation coefficients ("hard work"). We quantify the relative importance of each factor using an example of Delta and American and find that luck may explain about 38% of Delta's performance advantage. Second, we estimate the global network effect of local improvements. We show that the overall network effect of a delay-reducing investment may qualitatively differ from its local effect. Finally, we quantify pro-competitive benefits from a merger between two airline

networks. These benefits (or "efficiencies") represent an important part of a prospective merger analysis. However, to be considered by an antitrust enforcer, these benefits must be quantifiable (US Department of Justice (2010)). We show how to quantify them in our setting. The unifying conclusion of our three counterfactuals is simple: network effects matter. An analysis without network effects leads to qualitatively different answers.

The primary object of our analysis is a conditional distribution $D|X$, where D is a vector of realized delays of all flights that an airline is operating during a day, and X is the airline's network characteristics. Both objects are high dimensional. Major U.S. airlines operate thousands domestic flights a day. So, the dimensionality of D is several thousands. The airline's network characteristics are an object of potentially higher complexity. First, it encodes all flight specific demand and cost factors that the airline may take into account when it decides whether to delay a flight and by how much. Second, it includes information on the airline's entire domestic schedule: each flight's scheduled departure and arrival time, origin and destination airport, availability of spare planes in case of mechanical delays, the distribution of mechanical and weather-related shocks and so on. We have little theoretical guidance on which part of this information ends up being crucial. Our goal is to propose a set of tools that can figure that out. The scope and quality of available data determine which economic questions we can address with the tools we propose. For example, since delays are directly observed in the data, our tools can be used to determine what would happen if a shock is exogenously introduced to one part of the network without the airline being able to re-optimize the network ("one-time shock"). At the same time, we will not be able to say how much an airline would benefit if it adds a spare plane in one of the hubs because we don't see a credible shifter that would exogenously affect the number of available planes so that we could use any such observable variation to identify this effect causally.

In our application we start with a network in which a flight is a node and a (directional) link between two flights exists whenever a delay is systematically transmitted in that particular direction. The strength of this link is determined by the strength of the delay transmission. An appealing feature of this network of delays is that delays arise both for endogenous reasons (airlines slowing flights down to wait for incoming aircraft, connecting passengers, or incoming crews) and

for exogenous reasons (such as inclement weather or air traffic control). More importantly, it is reasonable to assume that the shocks on the various links are correlated within the day, but are independent across days (to the first order approximation), and the airline schedule is fixed over longer horizon (e.g., a quarter). This allows us to follow a natural asymptotic argument in our estimation step. Utilizing variation in network geography across aircraft types, airlines, and over time, we are also able to speak to how different airline network designs may alleviate or exacerbate the shock propagation. In the case of airline networks, there are other network characteristics such as the heterogeneity of the aircraft fleet that play an important role and we quantify this role as well. Of course, we need to exercise care when interpreting such results due to the lack of random variation in network characteristics.

Our analysis proceeds in the following steps. Relying on the industry specific details, we first present a very simple structural model of the data-generating process. We do this with three goals in mind. First, we get a tractable mechanism of shock propagation in networks. Second, the model allows us to explicitly determine under what conditions the estimates of descriptive regressions can be given a causal interpretation. Finally, we show how to identify the fundamental parameters of the model off the observed data.

We then derive the reduced form of our structural model that defines the observed delay. We proceed with the descriptive analysis of the joint distribution of delays. We regress the delays of each flight on the realized delays of incoming flights, the realized delays of the incoming flights for the incoming flights, and so on, up to four lags. These regressions resemble the textbook vector-autoregression (VAR) analysis with one important distinction. The asymptotic assumption of the textbook VAR analysis implies that the number of lags grows with the sample size. In our setting, each new observation reveals the entire distribution of delays for all flights, which allows us to keep the number of lags fixed as the sample size grows. Using the structural model of the data generating process, we identify a possible source of reverse causality that could potentially bias the estimates of the VAR regressions. We propose a formal statistical test to determine whether this reverse causality effect is present. We find evidence of this spurious correlation in the data. The model, however, suggests a natural identification strategy that relies on instrumental variables, whose

validity and relevance are derived from the assumptions placed on the structural representation of the data generating process. We reestimate the VAR regressions using these instruments and use these reduced form coefficients to perform a counterfactual analysis to address the first of our economic questions. We compare the overall performance of Delta Air Lines to that of American Airlines and conclude that Delta's advantage can be attributed both to its superior network *and* to a more favorable distribution of shocks in its major hubs. In other words, both "luck" and "hard work" are important to Delta's "on-time machine" brand. This result contributes to the growing literature that studies causes for widespread differences in management practices and productivity among firms and countries (see Syverson (2011) and Bloom and Van Reenen (2010)).

At the final step of our analysis, we estimate the fundamentals of the structural model using a method of moments. Taking into account the suggestive evidence of potential endogeneity in the data, we impose the same orthogonality restrictions as we did for our IV-VAR results.

We use the estimated coefficients to simulate two counterfactual scenarios to answer the other two economic questions. Importantly, these questions cannot be answered based on the IV-VAR coefficients because, as we show, the reduced form derived in the paper will change in the corresponding counterfactuals. We find that the global effect of a local infrastructure improvement can qualitatively differ from its local effect. First, although somewhat counterintuitive, it is not necessarily true that airlines experiencing fewer delays benefit less from a delay-reducing improvement. On the contrary, shorter and less frequent delays indicate the importance of the flights to the airline's network and associated higher costs of delaying these flights. These airlines will benefit from a delay reducing improvement because that improvement will result in cost savings. Similarly, an airline with the largest presence in an airport may not be the one that benefits from such an improvement the most. For example, our calculations show that even though JetBlue is currently the largest airline in Boston Logan Airport, the airline that would benefit the most from a delay-reducing improvement there is in fact American Airlines. This finding is particularly reassuring since it turns out that American Airlines operates the same aircraft type between Boston and JFK as it does between JFK and LAX and between JFK and SFO. This scheduling decision has exposed the stability of American's premium transcontinental New York service to shocks in Boston,

even though Boston is neither the origin nor the destination for these premium routes. Finally, we quantified the network benefits from American–US Airways (2015) and Alaska–Virgin (2016) mergers. We found that the relative benefit from network integration is quite small (decrease in overall delay-related costs of less than 0.3%).

There is a rich recent literature on shock propagation in networks arguing that network topology is one of the crucial determinants of the strength of spillovers of shocks between nodes (see e.g., Acemoğlu, Carvalho, Ozdağlar and Tahbaz-Salehi (2012), Acemoğlu, Ozdaglar and Tahbaz-Salehi (2015), Elliot, Golub and Jackson (2014), Carvalho, Nirei, Saito and Tahbaz-Salehi (2016)). In this paper we propose a framework for thinking about aircraft scheduling problem, which takes into account heterogeneous cost of effort necessary to avoid delays and the impact of the airline route network topology on shock propagation and thus on (expected) implied costs of delays. Using this framework we build an empirical model, in which we utilize a model-selection algorithm to reduce the dimensionality of the problem and evaluate the impact of a delay of an individual flight on the rest of that airline's network. There is a burgeoning literature on econometrics of networks (see de Paula (2017) for a survey, and de Paula, Richards-Shubik and Tamer (2018b), de Paula, Rasul and Souza (2018a), Menzel (2015), Manresa (2016), or Graham (2017) for further examples).

In a paper studying financial networks, Bonaldi, Hortaçsu and Kastl (2013) propose to use elastic net to estimate the network of spillovers of funding costs and use it to define systemic risk. One important additional contribution of our paper relative to Bonaldi et al. (2013) is that using our application on airline networks we can better gauge whether the estimation method based on the elastic net algorithm works reasonably. As we will argue below, unlike in the case of financial network where the links between individual institutions are largely unobserved, in the case of airline network, we do observe a very important piece. In particular, the entire sequence of flights that each physical aircraft performs on given day is known. Each aircraft has a unique identifier, called tail number, and both the scheduled and the realized path of each tail number during the course of a day is known. We can thus evaluate how much of the observed delays can be attributed to the purely "mechanical" delay transmission due to the flights serviced by the same tail number being scheduled too close to one another and how much is due to unobserved factors: either due to crew

scheduling or to real-time airline optimization where airlines try to minimize delay cost by taking into account connecting passenger itineraries etc.

The remainder of this paper proceeds as follows. We give a brief overview of the institutional details as well as data sources in Section 2. We use these details to develop a structural model of the data generating process that we outline in Section 3. From this model, we derive a reduced form that defines the joint distribution of the observed delays. We describe our data in Section 4. Section 5 presents a reduced form analysis of the data and its results. Section 6 presents the results of our structural analysis. We conclude in Section 7.

2 Industry Background and Data Sources

2.1 Why Airlines?

Understanding how shocks propagate in networks is crucial in many economic settings. Acemoğlu et al. (2012) illustrate how input-output linkages can transmit shocks through the whole economy and thus have real macroeconomic implications for the business cycle. Burzstyn, Ederer, Ferman and Yuchtman (2014) show that social learning among friends and peers can causally impact financial asset purchases. Hence, positive shocks to some central nodes can trigger a cascade of purchases. Conley and Udry (2010) show that farmers in Ghana adjust input usage based on what their successful neighbors do, which is again direct evidence that shocks elsewhere on the (social) network matter for allocations. Chaney (2014) shows that exports of French firms also respond to shock realizations of connected firms and that new trading partners are often found using existing contacts. These and many other applications show that the network structure is of utmost importance in many settings of interest. We focus on the airline industry for two reasons.

First, the airline industry is an important part of the U.S. economy. For every dollar of U.S. gross domestic product, the industry contributes 5 cent. Driving more than 10 million American jobs, the industry remains in the focus of government attention. Before 1978, almost all the industry was regulated. A federal government agency, the Civil Aeronautics Board (CAB), used to decide where airlines can fly, how many flights they could offer, and how much they could charge. Deregulation

decentralized these decisions. Even though it is indisputable that the prices went down following deregulation, the effect on non-price characteristics of air travel is often disputed. Time and time again consumers and policy makers raise concerns about systematic delays and cancellations and quality in general. There is a general consensus that some of these problems can be alleviated with additional investments in travel infrastructure. However, in order to spend these resources efficiently, it is crucial to understand how delay shocks propagate through airline networks.

Second, the comparative advantage of the airline industry is the availability of great data. For example, due to its commercial sensitivity, there is little public information on many financial transactions. Historical on-time performance data for all major airlines are publicly available, generally accurate, disaggregated, and very detailed. Forbes and Lederman (2009) used these data to study patterns of vertical integration in the US airline industry.

Airline networks remain stable over longer periods of time (generally, several months). At the same time, the realizations of delays are observed daily. To first approximation, each day can be treated separately. Shocks that last multiple days (e.g. winter storms) are rare. The industry itself makes a distinction between flights that end the day ("remain-overnight," or RON fights) and flights that arrive early enough to serve as inbound flights to some other flights later that day. Thus, we have multiple, oftentimes many, realizations of shocks for the same network structure.

Importantly, there is a natural source of exogenous variation in shocks that causes the day-to-day variation in observed delays: mechanical problems and weather. The data on network characteristics have rich variation as well: there is plenty of cross-sectional variation in network topology (e.g, contrast Southwest and United) as well time-series variation in network topology within airlines due to new entry/exit or mergers.

2.2 Industry Background

Scheduling in commercial aviation is possibly one of the most complex problems that companies need to solve. Aircraft are expensive assets that are extremely costly to leave idle. This fact forces the airlines to invest in making scheduling as efficient as possible by minimizing times when aircraft are not transporting passengers. This then makes it difficult to absorb any kind of unforeseen

shocks, such as delays due to air-traffic control or due to weather, as the typical schedules leave relatively little time for on-the-fly adjustments.

The schedule itself is an outcome of a much bigger problem. First, an airline chooses which routes to serve. Then it assigns to each route capacity (the total number of available seats) and aircraft type, which determines frequency. Given the schedule, the airline scheduler solves the fleet assignment problem, which determines a sequence of flights to be performed by each aircraft. The airline then develops the schedules of crews. While it is clear that an optimum should involve solving these problems simultaneously, the problem is too complicated for the industry to solve it that way.

Therefore, the literature traditionally assumes that the problem can be separated and solved sequentially. In other words, when the airline develops its schedule, it does not take fully into account how it would affect the flight assignment problem. In this paper we will point to some results from the operations research (OR) literature, but our main objective is to build a tractable empirical model for a subproblem: the real-time scheduling of airplanes and crews that will allow us to quantify the extent of delay externalities and attribute them to the various network characteristics. The data we use allow us to study both cross-sectional differences between various airlines and time-series differences. We will then try to relate these differences to differences in route network characteristics.

2.3 Data Sources

The main data set for our study comes from Airline On-Time Performance Database collected by the Bureau of Transportation Statistics.[1] This database collects flight-level data reported by U.S. certified air carriers that account for at least one percent of domestic scheduled passenger revenues. It includes scheduled and actual arrival and departure times for most of the commercial flights in the U.S. airspace. In particular, it contains on-time departure and arrival data for non-stop scheduled domestic flights by major U.S. air carriers.[2] The Office of Airline Information in DOT defines a major carrier as a U.S.-based airline that posts more than $1 billion in revenue during

[1] Available here: `http://www.transtats.bts.gov/Tables.asp?DB_ID=120`.

[2] The criteria for classifying a U.S. air carrier as major are unfortunately not consistent between DOT's own grouping and the one used in the on-time performance database description.

a fiscal year. They regularly publish accounting and reporting directives that explicitly state the following calendar year's air carrier groupings, according to which each airline files so-called Form 41 reports.

To keep the size of the data set manageable, we focus on Jan-Jun 2010-2015 and on eight major airlines: United (merged with Continental in April 2010), American (merged with US Airways in October 2015), Delta, Alaska Air (merged with Virgin in December 2016), US Airways, Virgin America, Jetblue and Southwest. These airlines account for the overwhelming majority of daily scheduled domestic flights and of daily transported passengers. As we will argue below, this set of airlines provides us with rich variation in the network characteristics: while most airlines operate on a hub-and-spoke network (UA, DL, AA etc.), few airlines operate a spoke-to-spoke (Southwest, Jetblue). Airlines also differ in the number of hubs they employ, their location, density of their routes and in the heterogeneity of employed aircraft. One of our goals is to relate these characteristics of the network to how delay shocks propagate through the flight network on a given day.

2.4 The OR approach to the Problem

There is an extensive literature on aircraft scheduling in operations research (OR). Mathematically, it is a many-to-one assignment problem that can be informally defined as follows. A discrete set of planes has to be assigned to a (larger) set of scheduled flights. The objective is to minimize the total costs of delay. A feasible assignment has to satisfy a number of natural constraints. First, whenever a plane is assigned to two consecutive flights, the destination airport of the first flight must be the origin airport of the second flight. Second, the departure time of the second flight cannot be earlier than the arrival time of the first flight plus some minimum turnaround time. Third, there are constraints on how long a plane has to stay on the ground for routine maintenance after certain number of flights. After all these constraints are specified, the solution to the assignment problem can be found numerically within reasonable time. We, however, will not be using an OR type of model in our analysis. We chose to do so for a number of reasons. First, the solution to the problem is likely not unique. Apart from trivial relabeling, delaying a

given aircraft by a minute is likely not going to change the optimal value of the objective function. Second, to obtain a non-generate distribution of realized delayed, we need to introduce stochastic shocks to the model. Adding them to the minimum turnaround time would be a natural way to augment the model. The problem of this approach, however, is the fact that the observed delays will likely be a discontinuous and hard-to-integrate function of these shocks, which can limit the extent to which the model generated distribution of delays can approximate the one observed in the data. Third, many important variables that are crucial to the decisions of the airline scheduler (e.g. the number and readiness of substitute planes) are not recorded in the public data.

Although the OR literature typically treats this assignment problem as static, the actual problem is inherently dynamic. As new information on mechanical and weather related shocks continuously arrives, the airline's irregular operations team adjusts the assignment trying to minimize the overall impact of these shocks on the airline's system. The assignment that looked optimal in the morning may be revised several times during the day as delay shocks and cancellations propagate though the system. We do not study this aspect of scheduling primarily due to data limitations. We have very little information on how the decisions of the scheduling team changed throughout the day. The data only record the realized assignment. Additionally, the scheduling team has far superior real-time information on mechanical and weather related shocks that gets revealed over time. A mechanical problem that looked minor at the beginning may end up being more serious than expected. Of course, one could model the continuous process of shock realization and then match the solution to this (very complicated) problem to the observed data. It is unlikely, however, that modeling this process is a first order issue for understanding the performance of an entire airline network. That is why we proceed with a simpler setting in which the scheduler's problem is static and all shocks are known at the beginning of the day. It is unlikely that the fundamentals of this simpler problem are going to change in the counterfactual we consider. This assumption will be less palatable, however, if the dynamic aspect of the process were the core of the counterfactual of interest (e.g. the overall network effect of a more accurate weather forecast).

3 Model of Flight Delays

We develop a model of delay propagation in airline networks with two main goals in mind. First, we will use this model to interpret the coefficients of our main descriptive regressions defined in the subsequent section. In particular, we will be able to state explicitly what assumptions we need to place on the sources of variation in the data so that the estimated coefficients of delay propagation have causal interpretation. Second, once we estimate the primitives of the model, we will be able to perform a set of counterfactual simulations for which the impact of the network externalities is first order and needs to be taken into account. The leading example is investment in airport infrastructure, which allows for easier delay avoidance.

Our focus here is on the day-to-day adjustment in aircraft scheduling (routing). Hence, we view both the competitive environment and the planned schedule (which included all scheduled flights and the assigned physical airplanes and crews) as fixed and we are interested in analyzing how the daily assignments of planes to routes scheduling proceeds as various random shocks get realized.

3.1 Fundamentals

Airline and Flight Schedule. Let n be the number of flights that are scheduled to be performed during a day t ($t = 1, \ldots, T$).[3] Assume that the day is divided into S non-overlapping discrete intervals, "time slots" (e.g. 30-minute intervals). The set of scheduled flights is denoted by $\mathcal{I} = \{1, \ldots, n\}$ and indexed by $i = 1, \ldots, n$. The airline serves A airports from set $\mathcal{A} = \{1, \ldots, A\}$, whose elements are indexed by $a = 1, \ldots, A$. Each flight i has origin airport $\underline{a}_i \in \mathcal{A}$, destination airport $\overline{a}_i \in \mathcal{A}$, scheduled departure time $\underline{s}_i$, and scheduled arrival time $\overline{s}_i$.

Effort, Delays, and Cancellations. Delays (and, in their extreme form, cancellations) are endogenous. In our model, they are determined by the amount of effort exerted by the airline. We assume that the realized delay of flight i, d_i is a (strictly) decreasing deterministic function of airline's effort e_i. We denote this function by $\phi(\cdot)$, i.e. $d_i = \phi(e_i)$.

Effort is costly. The costs of effort may depend on the particular airport and the time slot. Let

[3] Unless it may create confusion, we will suppress index t. However, this is the index that denotes a single observation. Our asymptotics relies on $T \to +\infty$.

e_{as} denote the total effort that airline exerts on all flights departing from airport a in period s:

$$e_{as} = \sum_{i \in \{i: \underline{a}_i = a,\ \underline{s}_i = s\}} e_i$$

We assume that the costs of effort have constant returns to scale. The marginal cost function is therefore a constant that we denote as c_{as}. These costs will depend on the aggregate delay of incoming aircraft.

Delays are costly too. We distinguish between direct and indirect costs of delay. Direct costs are costs that an airline has to incur because this flight is delayed. We denote them by $c_i(d_i)$. A delayed inbound flight also means that fewer aircraft will be available at the destination airport. This shortage makes the problem of the aircraft scheduling team harder. In our model, that means that the costs of effort at the destination airport go up as an indirect result of incoming delay. If the destination airport relies on this aircraft to operate subsequent flights, then a delay in the origin airport leads to higher costs of effort at the destination. We refer to these costs as the indirect costs of delays and cancellations.

3.2 Objective Function and Optimization Problem

Airline's goal is to minimize the total costs, which is a sum of the costs of effort and the costs of delays. Formally, airline solves the following unconstrained problem:

$$\min_{e_i,\ i=1,\ldots,n} C = \sum_{i \in \mathcal{I}} c_i(d_i(e_i)) + \sum_{s=1,\ldots,S} \sum_{a \in \mathcal{A}} c_{as} e_{as}$$

Optimality Conditions. Differentiating the objective function with respect to all e_i gives us n first order conditions:

$$\underbrace{c_i'(d_i) \times \phi'(e_i)}_{\text{direct costs of delay}} + \underbrace{\frac{\partial c_{\overline{a}_i \overline{s}_i}}{\partial d_i} \times e_{\overline{a}_i \overline{s}_i} \times \phi'(e_i)}_{\text{indirect costs of delay}} + \underbrace{c_{\underline{a}_i \underline{s}_i}}_{\text{costs of effort}} = 0.$$

These first order conditions state that airline should exert effort as long as the marginal benefit of effort exceeds its marginal costs. The marginal benefit of effort is a reduction in costs caused

by delays. Fewer minutes of delay means less costs—both direct and indirect—that airline has to incur. The multiplier $\phi'(e_i)$ there is simply an "exchange rate" that converts units of delay into units of effort. The marginal costs of effort is simply $c_{\underline{a}_i \underline{s}_i}$.

Since the marginal costs of effort are the same for all flights departing from the same airport in the same time slot, these first order conditions lead to an important restriction. Two flights scheduled to depart in the same time period should have the same marginal costs of delay. Formally, for $i \in \mathcal{I}$ and $j \in \mathcal{I}$ such that $\underline{a}_i = \underline{a}_j$ and $\underline{s}_i = \underline{s}_j$, in equilibrium:

$$\left[c_i'(d_i) + \frac{\partial c_{\overline{a}_i \overline{s}_i}}{\partial d_i} e_{\overline{a}_i \overline{s}_i}\right] \phi'(e_i) = \left[c_j'(d_j) + \frac{\partial c_{\overline{a}_j \overline{s}_j}}{\partial d_j} e_{\overline{a}_j \overline{s}_j}\right] \phi'(e_j).$$

Intuitively, suppose that the marginal costs of effort are different for different flights departing from the same airport in the same time slot. If that was the case, airline could be better off by increasing its effort on the flight with lower marginal costs and decreasing its effort on the flight with higher marginal costs, by the same amount. Only when such reallocation is not possible, airline will achieve the optimum assignment.

In the data, we do not observe the amount of effort exerted by airlines.[4] Nor do we have direct information on the costs of delay. Out result, however, suggests that the joint distribution of realized delays for different flights should contain information on how the costs of delay for different flights relate to each other. Intuitively, if one of two flights gets consistently delayed more often than the other, that should imply that the costs of delay for this flight are lower than for the other.

3.3 Observables and Stochastic Structure

To establish identification formally, we first must describe the data generating process. The direct costs of delay and the costs of effort are fundamentals that we seek to identify, while the indirect costs of delay arise endogenously: a flight that arrives late (or does not arrive at all) increases the costs of effort at the destination airport.

We impose the following stochastic structure.

[4] For that reason, one could propose an alternative representation of the model, in which the total amount of delays is given to the scheduler who needs to allocate them among flights. This model would be equivalent to our model after an appropriate change in notations is made. The first order conditions will have to include the corresponding shadow value of the overall amount of delays.

Direct Costs of Delay. We assume that the unobservable part of the direct costs of delay is additively separable. For each flight i, the direct costs of delay are defined as follows:

$$c_i'(d_i) \times \phi'(e_i) = g(d_i) + \epsilon_i,$$

where g is an invertible deterministic function that may depend on some observable characteristics and ϵ_i is a mean-zero idiosyncratic cost-shifter that varies from day to day independently of the observable characteristics.

Costs of Effort. For each airport a and time period s, the marginal costs of effort are defined as follows:

$$c_{as}(e_{as}) = f(z_{as}; \beta_z) + \varepsilon_{as},$$

where f is a deterministic function, and ε_{as} is a random mean-zero shock whose realization varies from day to day independently of other shocks. Cost shifters z_{as} include observable characteristics such as realized inbound delay by period s, inbound cancellations, or the number of spare airplanes on the ground. For example, if $f(z_{as}; \beta_z) = z_{as}\beta_z$, then the parameter β_z determines the marignal impact of these observable shifters on the costs of effort.

Indirect Costs of Effort. The indirect costs of delay arise endogenously in the model. They are defined as the impact of a delay on the costs of effort at the destination airport: if one or several inbound flights are delayed, it will become more difficult for the airport to ensure on-time departure of its flights. Formally, the indirect costs of delay:

$$\frac{\partial c_{\overline{a}_i\overline{s}_i}}{\partial d_i} \times e_{\overline{a}_i\overline{s}_i} \times \phi'(e_i) = \frac{\partial f(z_{\overline{a}_i\overline{s}_i})}{\partial d_i} \times e_{\overline{a}_i\overline{s}_i} \times \phi'(e_i) = h_{\overline{a}_i\overline{s}_i}(d_i),$$

where $h_{\overline{a}_i\overline{s}_i}$ is a deterministic function that depends on the delays of originating flights at the destination airport.

An Observation. We assume that each day airline faces new realizations of both costs of delay and costs of effort. This assumption implies that the scheduling problem of airlines is separable over days. Even though it is fully consistent with the airline lingo that distinguishes between RON

("remain overnight") and non-RON flights, there are notable exceptions that may violate it. Some flights are scheduled overnight ("red-eyes"). However, they are typically between hubs and have little effect on morning flights. The effect of extended (typically weather related) disruptions may last several days, which would violate the separability assumption as well but such disruptions are infrequent to have any significant impact on our results.

Discussion. Even though airline schedules do not change significantly from day to day, there is a lot of variation in the time of actual departure and arrival. In our model, this variation is caused by two sets of random variables: ϵ_i and ε_{as}. The first set of shocks, ϵ_i, affects the idiosyncratic performance of an individual flight. Negative realizations of ϵ_i imply that delaying this particular flight i is less costly compared to other flights. Therefore, flight i will more likely be delayed, which makes further delays at the airport of its destination more likely. Mechanical delays are a good example of this type of shocks. Variation in ϵ_i identifies costs of effort at the destination airports, and, therefore, the indirect costs of delay.

The second set of shocks, ε_{as}, are airport-specific shocks. Higher realizations of this type of shocks imply that all flights departing from this airport are likely to be delayed. An example of this type of shocks are weather-related factors. Exogenous variation in the costs of effort caused by these shocks identifies the direct costs of delay.

The purpose of the model is to explain how shocks propagate in networks. To illustrate the shock-propagation mechanism implied by our model, consider the impact of shock ϵ_i on the rest of the network. A negative realization of shock ϵ_i will lead to a delay of flight i and its late arrival to airport $\overline{a}_i$. This delay in turn will increase the cost of effort $c_{\overline{a}_i\overline{s}_i}$. This increased costs will affect all flights departing from $\overline{a}_i$ in slot $\overline{s}_i$ but to a different degree. Flights that are more costly to delay (based on the sum of their direct and indirect costs) will be delayed less. Similarly, flights that have lower costs of delay will be impacted more.

3.4 The Reduced Form

The optimality conditions that rationalize the observed delay of each flight i define the structural form of the model:

$$\underbrace{g(d_i) + \epsilon_i}_{\text{direct costs of delay}} + \underbrace{h_{\bar{a}_i \bar{s}_i}(d_i)}_{\text{indirect costs of delay}} + \underbrace{f(z_{as}) + \varepsilon_{as}}_{\text{costs of effort}} = 0.$$

The observed delay, d_i, is a solution to this system of equations that depends on the unobserved costs of effort, unobserved direct costs of delay, and the delays at the destination airport (through the indirect costs of delay).

Assuming that the total cost of delay is an invertible function, the optimality conditions lead to the following reduced form:

$$d_i = \mathcal{C}(f(z_{as}) + \varepsilon_{as} + \epsilon_i),$$

where $\mathcal{C}$ is an unknown transformation.

Combined together for all flights, these expressions for the optimal delay define the joint distribution of realized delays across the entire airline's network (D) conditional on various network characteristics (X). This distribution is highly multi-dimensional. To make it analytically tractable, we will take two alternative, yet complementary approaches to analyzing it.

First, we look at the joint distribution of delays through the lens of a vector autoregression model. Since not all types of aircraft can perfectly substitute each other, delay propagation will be relatively sparse. We leverage this sparsity to estimate the propagation coefficients of the VAR model. To give a causal economic interpretation of these propagation coefficients, we rely on the underlying structural model developed in this section. We then proceed with estimating the structural model directly by imposing a parametric specification.

We show that these methods complement each other. The first method is simpler and require fewer parametric assumptions. It can be used to address industry relevant economic questions for which the propagation coefficients do not change. The second method, to be implemented efficiently, may require a more restrictive parametric structure, even though, as we show next, the model is non-parametrically identified.

3.5 Non-Parametric Identification

We begin by showing that the reduced form of the model is non-parametrically identified. As previously derived, assuming invertibility, the optimality conditions lead to the following reduced form:

$$d_{it} = \mathcal{C}(f(z_{ast}) + \varepsilon_{ast} + \epsilon_{it}),$$

where $\mathcal{C}$ is an unknown transformation. Keep in mind that we observe delays of the same exact flight day-after-day ($t = 1, \ldots, T$) over a relatively long period of time (T).

This is a familiar class of econometric models known as "regression models with an unknown transformation of the dependent variable." The asymptotic argument in our application assumes that the number of delay observations for the same flight goes to infinity. A set of identification results was first derived by Horowitz (1996). Chiappori, Kristensen and Komunjer (2015) further extends the analysis of these models by providing a set of sufficient conditions that guarantee that the unknown functions $\mathcal{C}(\cdot)$ and $f(\cdot)$, together with the distributions of the unobserved shocks are non-parametrically identified. We will not restate these conditions and the associated theorem here explicitly.[5] Rather, we will discuss the intuition behind them.

Broadly speaking, the identification argument requires two familiar conditions: relevance and validity of the cost shifter z_{as}. The first condition ("relevance") ensures that the observable part of the costs of effort at the origin airport, $f(z_{as})$, varies from observation to observation. Without such variation we cannot identify the function $f(\cdot)$. The second condition ("validity") requires the unobservable part, $\varepsilon_{as} + \epsilon_i$, to be independent of the cost shifter, z_{as}. Without this condition, the observed and unobserved sources of variation in delays cannot be separately identified. As long as these two conditions are satisfied, the model can be non-parametrically identified (provided that some technical assumptions that ensure the differentiability of the unknown functions hold).

Which observables can satisfy these conditions? We need to find a shock that affects the costs of effort at the airport but is independent of the unobservable shocks that move delay. If the costs of effort were fully observed (no ε_{as}), observed delays to other flights that leave from the same

[5]Interested reader can consult Appendix A which restates these conditions within the context of our model, discusses the relevant assumptions and presents the formal Identification Theorem.

airport in the same time slot would satisfy both conditions (provided that the shocks to direct costs of delay are in fact independent). The assumption that the costs of effort are fully observed is unfortunately unlikely to be satisfied in practice.

Observed delays to inbound flights whose aircraft can be assigned to serve the flight in question naturally satisfy the relevance condition: more inbound delays imply higher costs of effort. However, the validity of this shifter may raise some concerns. By construction, the delay of an inbound flight is a function of (anticipated) aggregate delay at the destination airport. If the unobserved shocks to cost of effort and unobserved shocks to direct costs of delay are known before the decision to delay the inbound flight is made, the validity condition will fail, creating the endogeneity problem.

There is however a way to overcome this problem. Consider all inbound flights whose aircraft can be assigned to the flight in question. Even though the observed delay to these flights can be endogenous, any shifter of this delay that is independent of the unobserved shocks ε_{as} and ϵ_i will be both relevant and valid. Such shifter will in turn affect the realized delays of all other flights that depart from these airports at the same time as the inbound flights but to different destination.

To illustrate this argument, consider an example. Suppose we want to identify the costs of delay of the 2pm flight from Dallas to San Francisco. As discussed above, we cannot directly use flights that arrive to Dallas shortly before 2pm because their delays are likely correlated with the unobserved shocks to the San Francisco flight. Suppose these inbound flights are coming from Chicago, Boston, and Miami. What we can use instead are the observed delays of flights that departed from Chicago, Boston, and Miami at the same time as the flight to Dallas, but to any other destination than Dallas. These delays will provide a valid source of identification if costs of effort across airports in different time slots are not correlated.

The nonparametric identification of the reduced form does not necessarily guarantee that the structural form is identified as well. Indeed, the argument above establishes that the unknown transformation $\mathcal{C}(\cdot)$ is identified. Going back to the structural form, we have:

$$\mathcal{C}^{-1}(d_i) = -\left[g(d_i) + h_{\overline{a}_i\overline{s}_i}(d_i)\right].$$

Thus, to establish the nonparametric identification of the structural form, we need to show that

the direct, $g(d_i)$, and indirect costs of delay, $h_{\overline{a}_i\overline{s}_i}(d_i)$, are *separately* identified. To do so, we need to find an observable that moves the indirect costs of delay separately from the unobserved shocks.

If the unobserved shocks at the destination airport are unknown at the time the decision to delay is made, we could use the realized delay at the destination airport as a source of variation. However, this assumption is likely unrealistic.

If the shocks are known, any shock that increases the costs of effort at the destination airport that is independent of the endogenous delay at this airport will satisfy the two conditions. In particular, shocks to the costs of other inbound flights that are independent of the unobserved delay shocks at the destination airport will work.

To see the argument, suppose now that we want to identify the indirect costs of delay of the 2pm flight from Dallas to San Francisco. Consider the set of all flights that are scheduled to arrive to San Francisco at the same time as the flight from Dallas. Consider their origins. Suppose those are Los Angeles, Chicago, and Seattle. The delays of all flights that depart from Los Angeles, Chicago, and Seattle at the same time as the flights to San Fransisco but with a different destination are both valid and relevant and therefore move the indirect costs of delay of the Dallas - San Francisco flight.

Thus, the structural model explicitly defines the joint distribution of observed delays, the shock-propagation mechanism, and specifies what sources of variation can be used to identify the primitives of the model.

4 Data: Definitions and Stylized Facts

4.1 Measure of Delay

Table 1 reports the summary statistic of the key variable: the delays. Flight delays can be measured at departure or at arrival. While the arrival delays are perhaps more important from a passenger's perspective, the table suggests that at least at the aggregate level, it makes virtually no difference which one we use.[6] What may be important, however, is how to treat cancellations. The table summarizes delays where cancellations are top coded (as the longest observed delay conditional on

[6] In fact, most airlines themselves set internal goals that target on-time departure rather than on-time arrival.

non-cancellation) in columns (1) and (2) and conditional on non-cancellation in columns (3) and (4).

Table 1: Means of Delay (per flight) in minutes: Jan-Jun 2010-2015

	Dep Delay[a]	Arr Delay[a]	Dep Delay 2[b]	Arr Delay 2[b]	Obs.
B6	28.08	28.80	14.79	15.50	689,965
VX	17.08	17.48	12.25	12.63	111,078
AS	10.72	11.64	5.91	6.82	443,441
UA	14.07	13.36	14.01	13.30	1,314,227
AA	26.82	27.38	13.11	13.62	1,600,135
DL	18.02	18.49	10.23	10.70	2,242,915
WN	20.95	19.50	12.93	11.45	3,466,391
US	7.98	9.79	7.85	9.67	1,202,425

[a] Delays are topcoded,
[b] Delays are conditional on non-cancellation.

4.2 Sources of Variation

There are several sources of variation that we will exploit in our analysis. Day-to-day variation in observed delays comes from both exogenous factors and endogenous decisions. Flights may be delayed due to weather, air-traffic control, industrial action, mechanical problems, delayed inbound flights, airport congestion.[7]

To illustrate the variation in the recorded delays, we look at different slices of the data. Figure 1 shows a time series of delays for United Airlines, which shows quite a bit of heterogeneity at monthly level, with some evidence of seasonality. In contrast, however, the corresponding figure for American Airlines displayed in Figure 2 exhibits little seasonality. Figure 3 shows the time series of delays of Southwest which also does not exhibit much of a seasonal pattern. These graphs are useful when thinking about the appropriate definition of a period to choose for the estimation. While according to some industry sources, airlines' schedules are typically set at for a quarter, we will opt for assuming that the network is formed and stays fixed for one-month at a time.

The airline networks exhibit useful time-variation in their characteristics. For example, Figure 4

[7] Although the data do record the historical reason for delays and cancellations for each flight, few experts consider them reliable. One large airline was caught maintaining two distinct databases with reasons for delays and cancellations: one for public reporting, and the second one for internal use.

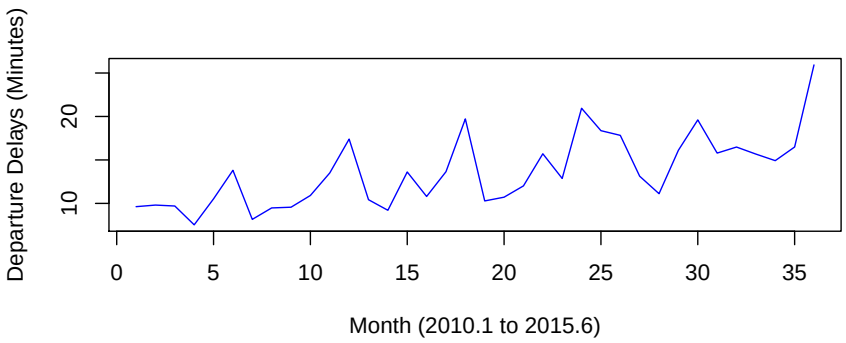


Figure 1: Monthly Average of United Airlines Departure Delays in minutes

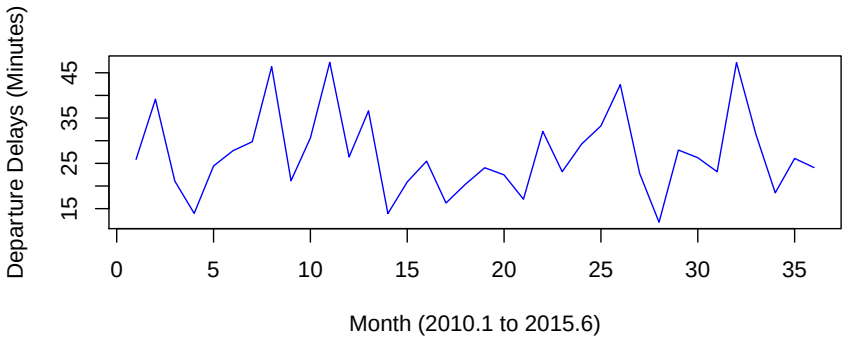


Figure 2: Monthly Average of American Airlines Departure Delays in minutes

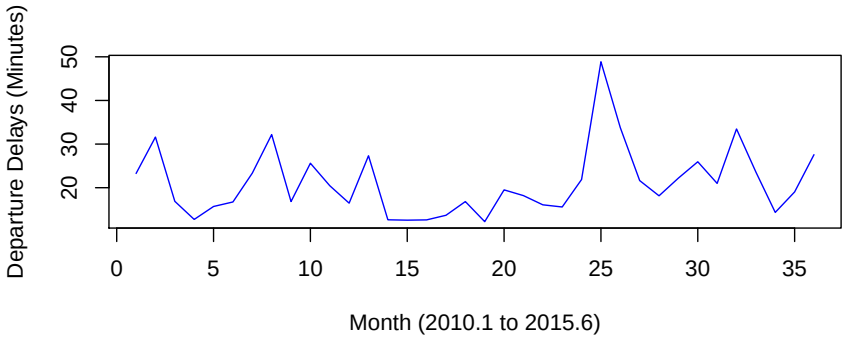
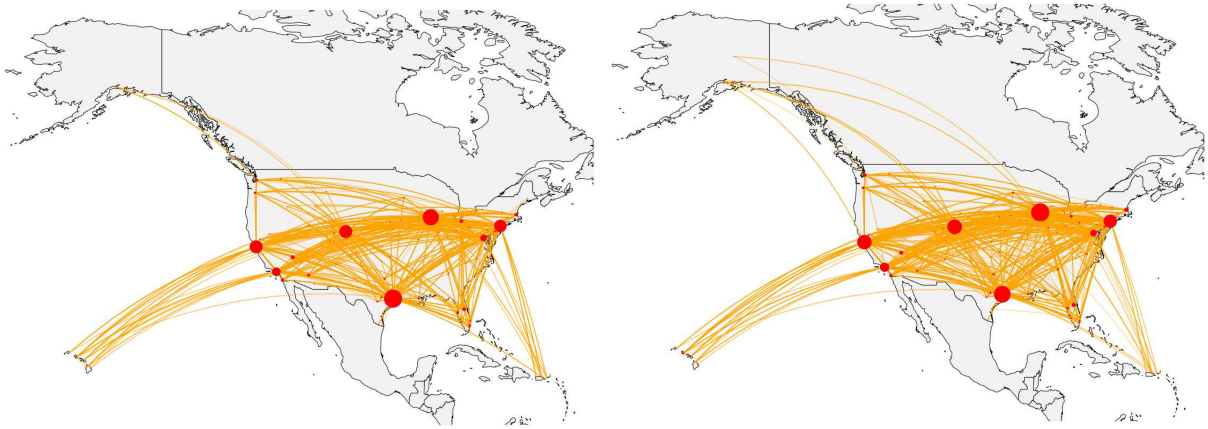


Figure 3: Monthly Average of Southwest Airlines Departure Delays in minutes



(a) January 2012

(b) June 2015

Figure 4: Network of United in January 2012 and June 2015

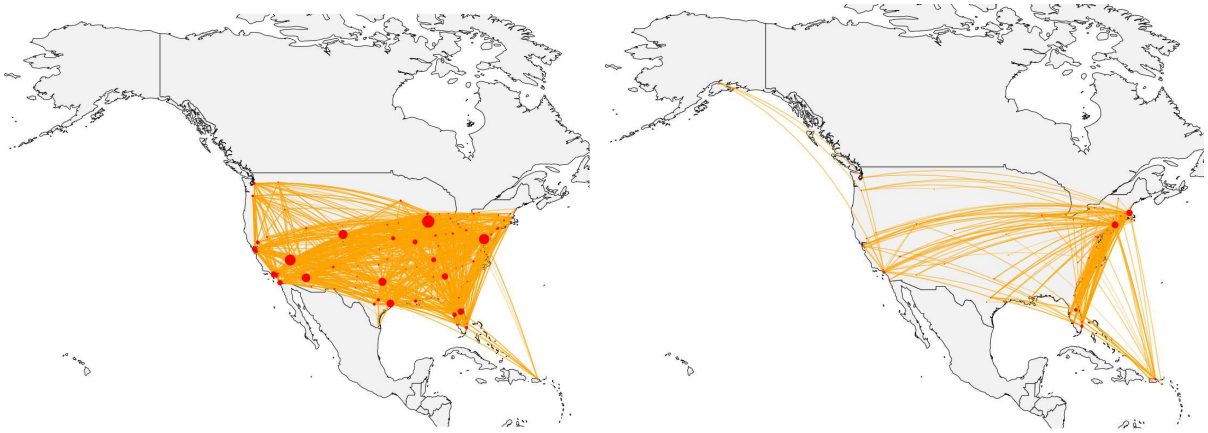
shows that over time, United's network became much denser. There are more flights in the right panel, and some new airports were added. There is also fair amount of cross-sectional variation in network characteristics. Figure 5 shows that Southwest has a very dense network with fairly short flights, whereas Jetblue specialized in serving just a few airports.

We now turn to observed variation in networks. Table 2 reports one of our key airline network characteristics: the degree distribution. These measures are defined formally in Section 5.1.4. The degree distribution can be roughly viewed as the expected number of links from a randomly chosen node.

	UA	AA	B6	AS	DL	VX	WN	US
1001	5.0	6.6	5.2	4.0	7.1	3.0	15.4	5.6
1002	5.4	6.1	5.1	3.8	6.7	3.0	16.1	5.1
1501	6.9	5.5	6.2	4.6	5.8	3.4	15.6	5.2
1502	6.9	5.9	6.6	4.3	6.2	3.5	15.1	5.6
Time Avg.	6.7	6.8	6.1	4.0	6.5	3.3	15.4	4.9

Table 2: Mean of network degree distributions by airlines, month

Then, consider figures 6 and 7. It depicts the flight network of United Airlines in June 2015 and of Southwest Airlines in June 2015. There is also some variation in the time-series: Figure 8 depicts the network of United Airlines in January 2012 (immediately after the merger with Continental).



(a) Southwest, June 2015 (b) Jetblue, June 2015

Figure 5: Networks of Southwest and Jetblue in June 2015

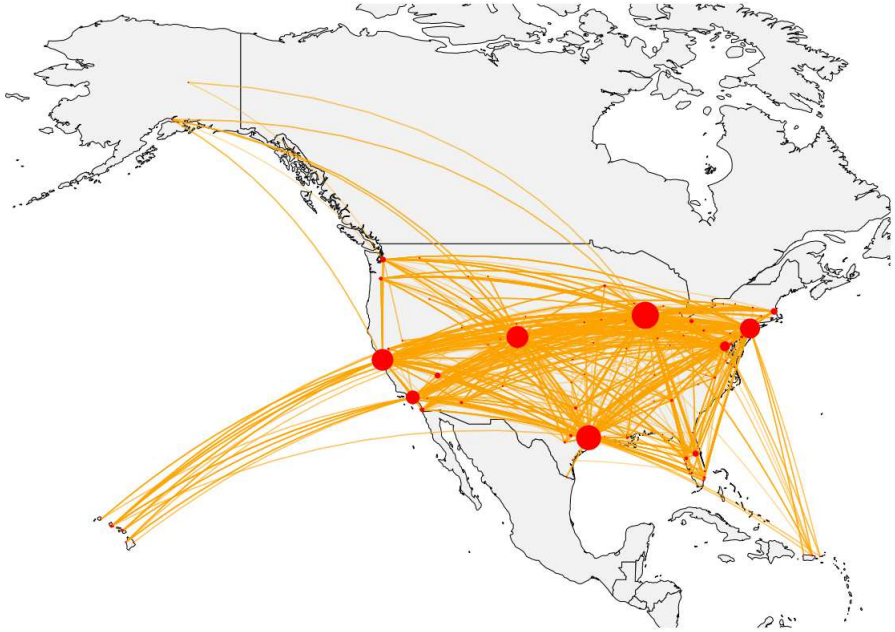


Figure 6: United Airlines, June 2015

It is reasonable to expect that delays might be affected by mechanical problems and that for a fixed number of planes, larger fleet heterogeneity might make it harder to substitute planes in real-time if such need arises as both pilots and mechanics are typically licensed only for one type of planes. Table 3 shows the top plane models used by the airlines in our study and Table 4 reports the (monthly) Hirschman-Herfindahl Index which summarizes how concentrated individual airlines'

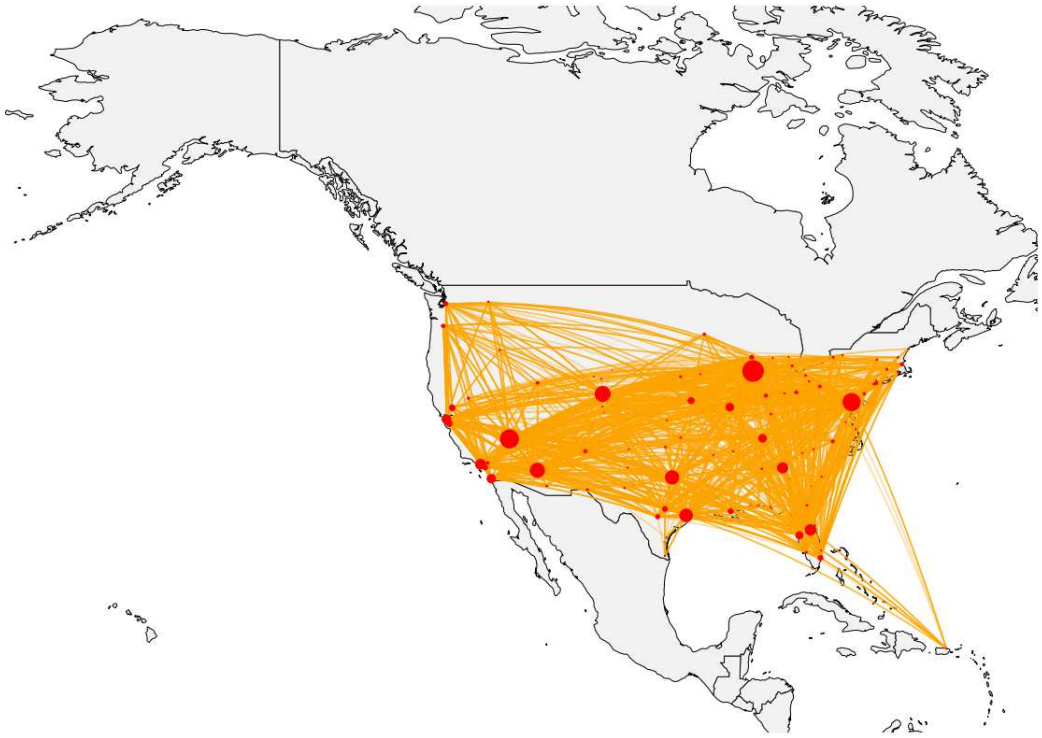


Figure 7: Southwest Airlines, June 2015

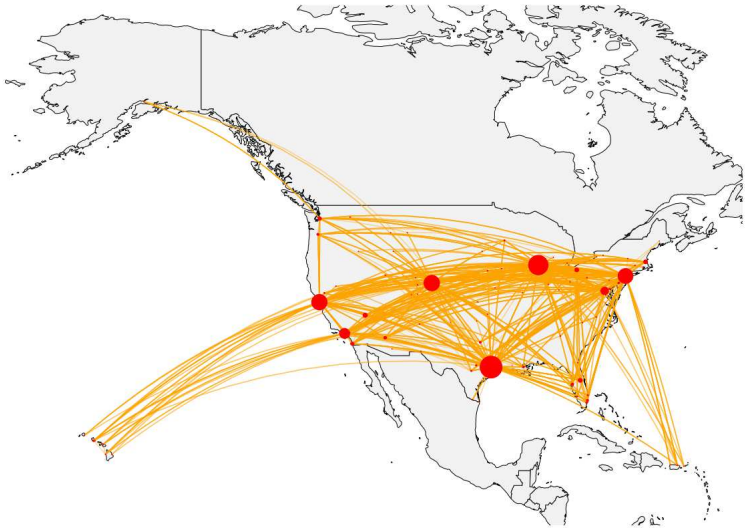


Figure 8: United Airlines, January 2012

usage of planes is. It shows that Jetblue, Virgin, and Southwest use significantly fewer models than the legacy airlines. It also shows that there is some time series variation within airlines, especially in United's case after its merger with Continental, which was ultimately implemented in January

2012.

We complement these data on delays by data on passengers from T100-Segment database, which will be useful for scaling our results appropriately, i.e., converting minutes of delay of a flight into passenger-minutes.

Aircraft Type	Performed Flights	Avg. Avail. Seats
BOEING 737-700/700LR/MAX 7	2,427,622	136.5
AIRBUS INDUSTRIE A320-100/200	1,377,364	148.0
MCDONNELL DOUGLAS DC9 SUPER 80/MD81/82/83/88	1,297,214	143.5
BOEING 737-800	1,281,841	160.6
BOEING 757-200	940,946	181.3
BOEING 737-300	915,987	138.0
AIRBUS INDUSTRIE A319	841,536	123.7
EMBRAER 190	341,723	99.7
AIRBUS INDUSTRIE A321	307,658	183.3
BOEING 737-400	248,609	130.5
BOEING 737-900	231,597	171.5
MCDONNELL DOUGLAS MD-90	185,502	158.9
BOEING 737-500	141,600	121.1
BOEING 767-300/300ER	108,309	229.6
MCDONNELL DOUGLAS DC-9-50	88,595	123.7
BOEING 757-300	85,994	221.3
BOEING 717-200	78,843	110.0
BOEING 737-900ER	33,921	180.0
BOEING 767-200/ER/EM	29,772	183.0
BOEING 777-200ER/200LR/233LR	27,034	281.1

Table 3: Number of aggregate performed departures and averaged available seats per flight by aircraft type

	UA	AA	B6	AS	DL	VX	WN	US
1001	0.292	0.437	0.572	0.308	0.153	0.553	0.509	0.182
1002	0.297	0.436	0.563	0.307	0.153	0.549	0.508	0.184
1201	0.151	0.373	0.550	0.322	0.164	0.649	0.533	0.213
1202	0.153	0.372	0.541	0.329	0.162	0.649	0.536	0.213
1501	0.176	0.347	0.476	0.293	0.137	0.649	0.506	0.272
1502	0.180	0.347	0.475	0.288	0.135	0.657	0.511	0.275
Time Avg.	0.214	0.376	0.529	0.305	0.153	0.631	0.516	0.226

Table 4: HHI index of aircraft type use competition

5 Reduced Form Analysis: the Joint Distribution of Delays

We now turn to the joint distribution of delays. We condition this distribution on network characteristics and ask the following question: how the delays of incoming flights affect the delays of outgoing flights? We first address this question descriptively by performing a VAR type of analysis

of the observed correlations among flights. We then use the model defined in Section3 to give these estimated correlations a causal interpretation.

5.1 Econometric framework

5.1.1 L-operator

We begin by defining an operator that allows us to put a useful order-like structure on all flights scheduled by an airline on a given day. Let us define an operator $L_1(\Delta)$ which for a flight determines which flights are its immediate predecessors (in the sense of arriving at the same airport within Δ minutes before the scheduled departure), and then we will define recursively the predecessor of the predecessor and so on.

Consider the collection of all flights on a given day, $\mathcal{I} \equiv \{1, \ldots, i, \ldots, n\}$ in a given order, for instance, by origin $\underline{a}_i$, destination $\overline{a}_i$, ordered by scheduled departure time. Suppose $|\mathcal{I}| = n$. We first define a binary $n \times n$ matrix $L_1(\Delta)$. A "prior flight" is defined by matching of destination-origin and the difference between corresponding scheduled arrival time and departure time of the subsequent flight being less than a lag difference, Δ. Whenever a (i,j)- element $L_{1,ij}$ is equal to one, the j-th flight in $\mathcal{I}$ is a "prior flight" of i-th flight in $\mathcal{I}$ and $L_{1,ij} = 0$ otherwise.[8] Using this notation, we can then define the "lag 2" matrix L_2, indicating flights that are "prior to the prior" flights. This can be defined as an adjusted square of L_1-matrix (i.e., essentially applying the L_1 operator twice), where all entries of L_2 are equal to $\text{sgn}\left(L_1^2\right)$, where $\text{sgn}(\cdot)$ denotes the sign function. The logic is $L_{2,ij} = \sum_{m \in \mathcal{I}} L_{1,im} L_{1,mj} = 0$ if and only if there does not exist a flight as m-th element of $\mathcal{I}$ such that $L_{1,im} = 1$ and $L_{1,mj} = 1$. Thus, as long as j is a prior flight of a prior flight of i, $L_{2,ij} \neq 0$ and the sign function makes all such non-zero elements of L_1^2 equal to 1. "Lag k" matrix L_k can be defined recursively in a similar manner.

5.1.2 System of Delay Propagation

Using the notation described in the previous section, we are now ready to specify equations that we will take to the data and the estimation approach that we employ. We will proceed by analyzing the

[8] In our empirical results, we set the lag difference to be one hour.

delay spillovers on each airline's network separately, and subsequently we will relate thus obtained results to the features of the network, its topography, and to competition the airline is facing at various nodes of its network. Denote by D the n by 1 vector of delays for all flights in $\mathcal{I}$. Let the maximum depth a shock can propagate be K lags. This may be the longest sequence of "hops" according to the above-defined order on $\mathcal{I}$.[9] We now specify the following statistical model for delays on a network by an airline:

$$D = \text{const} + \begin{pmatrix} \sum_{m\in\mathcal{I}} \beta_{1,1m} L_{1,1m} D_m \\ \vdots \\ \sum_{m\in\mathcal{I}} \beta_{1,nm} L_{1,nm} D_m \end{pmatrix} + \cdots + \begin{pmatrix} \sum_{m\in\mathcal{I}} \beta_{K,1m} L_{K,1m} D_m \\ \vdots \\ \sum_{m\in\mathcal{I}} \beta_{K,nm} L_{K,nm} D_m \end{pmatrix} + \eta \tag{1}$$

$$= \text{const} + \sum_{l=1,\ldots,K} (\beta_l \circ L_l) D + \eta, \tag{2}$$

where const is a vector of (possibly unequal) constants with length n, β_l is a $n \times n$ matrix for $l = 1, \ldots, K$. $\beta_{k,ij}$ denotes the k-lag delay effect of flight j on flight i if j is a prior flight of i and there is a delay in flight j. η denotes a vector of exogenous delay shocks to each flight in $\mathcal{I}$. Notation "$\circ$" denotes element-wise matrix product (Hadamard product).

Equation (2) can be written in a long regression form as

$$D = c + W\beta + \eta, \tag{3}$$

where

$$W = (W_1, W_2, \ldots, W_K)_{n \times Kn^2}$$

[9] In our estimation, we will impose $K = 4$ due to computational constraints for most airlines and $K = 3$ for Southwest.

and

$$W_l = \begin{pmatrix} L_{l,1.} \circ D' & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & L_{l,n.} \circ D' \end{pmatrix}_{n \times n^2}$$

$$\beta = \left(\text{vec}(\beta_1')', \text{vec}(\beta_2')', \ldots, \text{vec}(\beta_k')'\right)'_{kn^2 \times 1}.$$

$L_{l,1.}$ denotes the first row of L_l. $\text{vec}(\cdot)$ denotes vectorization operator. Note that this is a very high dimensional problem as $\dim(\beta) = Kn^2$ where n is essentially the number of flights scheduled on a given day and K the number of lags allowed. Since as we discussed above the vector of coefficients β is sparse, we will estimate the long regression given by (3) by an elastic-net regression (Zou and Hastie 2005), which is a mixture of a Ridge Regression with the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani 1996). Some sparsity is directly imposed by assuming that the flight delays are independent across days.[10] The elastic net estimator is then simply a solution to:

$$\hat{\theta}_{enet} = \left(1 + \frac{\lambda}{2}(1 - \alpha_e)\right)\left(\underset{\theta \in \Theta}{\text{argmin}} \|D - Z\theta\|_2^2 + \lambda\left(\frac{(1-\alpha_e)}{2}\|\theta\|_2^2 + \alpha_e\|\theta\|_1\right)\right) \tag{4}$$

where $Z = \begin{bmatrix} 1 & W \end{bmatrix}, \theta = \begin{bmatrix} c & \beta \end{bmatrix}$, and $\|\cdot\|_1$ and $\|\cdot\|_2$ denote the L^1 and L^2 norms, respectively. Parameters λ and α_e determine the shadow value of the constraint and the relative weight on the norms, respectively.[11] The term $\left(1 + \frac{\lambda}{2}(1 - \alpha_e)\right)$ is a bias correction factor added by Zou and Hastie (2005) to lessen the downward bias due to double penalization. Appendix B presents the consistency properties of our elastic net estimator for our context. Note that our specification does not allow for contemporaneous effect, since it would obscure the interpretation of the reduced form coefficients introduced above, but such a model would in principle be identifiable and estimable as

[10] If the researcher were really worried about such dependence, the asymptotic argument can easily be adapted, for example, assuming independence across weeks rather than days.

[11] The parameter λ is typically set by cross-validation. From our experience the particular choice of α_e has little effect on results as long as it is away from the extremes of $\alpha_e = 0$ or $\alpha_e = 1$. We impose non-negativity constraints on the parameters by setting the *lower.limits* argument in the *cv.glmnet* function of the *glmnet* package in R.

proposed by de Paula et al. (2018a).

5.1.3 Matrix of Systemicness

Notice that equation (2) can also be written as:

$$D = \left(I_n - \sum_{l=1,\ldots,k} (\beta_l \circ L_l) \right)^{-1} \text{const} + \left(I_n - \sum_{l=1,\ldots,k} (\beta_l \circ L_l) \right)^{-1} \eta. \tag{5}$$

This allows us to define a key matrix of interest:

$$K = \left(I_n - \sum_{l=1,\ldots,k} (\beta_l \circ L_l) \right)^{-1} - I_n \tag{6}$$

An element of this matrix K_{ij} can be interpreted as the long run effect of a minute delay shock to flight j on flight i. Then $k_j = \frac{1}{n} \sum_{i \in \mathcal{F}} K_{ij}$ can be used to measure average effect a minute delay in flight j on the rest of flights in $\mathcal{F}$. Note that the matrix K is a key ingredient in the calculation of systemicness and vulnerability of financial institutions in Bonaldi et al. (2013) and various centrality calculations in Diebold and Yilmaz (2014).

5.1.4 Network Characteristics

Now that we have estimated the weighted directed graphs of delays, which allows us to assign a "systemicness" score to each individual flight, we will proceed to link these scores with the properties of the airline network in the usual sense: nodes being airports and flights being links. We will mainly be interested in two different classes of characteristics: those related to the network topology and those related to homophily. We begin by defining these variables.

Given the focus of this paper is on airline networks, we will start our list of network characteristics with the natural ones: the *number of airports* served and the *number of hubs* that an airline operates. Furthermore, we borrow from network literature several standard definitions describing the topology of the network. A *degree distribution* is the frequency of number of links belonging to each node. Jackson and Rogers (2007) relate this object to spreading of infections over the

network, which is quite fitting in our application. A closely related measure is called network density, P_N. It is defined as the frequency of drawing any random pair of connected nodes (or a dyad): $\binom{N}{2}^{-1} \sum_{i=1}^{A} \sum_{j<i} B_{ij}$, where $B_{ij} = \mathbf{1}[i \text{ and } j \text{ are connected}]$. The average degree then simply equals $(N-1) P_N$. We will also use the standard deviation of the degree distribution as a measure of asymmetry of airports within the network.

A transitivity index (or clustering coefficient) is defined as the fraction of (three times the) transitive triads (i.e., transitive triplets) or the number of triads where we add those triads that are either transitive or would become transitive if a single link were added. As Graham (2015) notes, this measure should be close to the network density for random graphs, but could substantially deviate for non-random graphs.

5.1.5 Interpretation of the Coefficients

Recall that the optimality conditions imply:

$$d_i = \mathcal{C}(f(z_{as}) + \varepsilon_{as} + \epsilon_i).$$

Differentiating with respect to the delay d_j of an inbound flight j (and ignoring endogeneity) yields:

$$\frac{\partial d_i}{\partial d_j} = \mathcal{C}'(f(z_{as}) + \varepsilon_{as} + \epsilon_i) \frac{\partial f(z_{as})}{\partial d_j}.$$

The delay propagation coefficients hence approximate the local average value of the left-hand side of this equation. Other things equal, we should expect higher delay propagation coefficients when inbound flight j has higher impact on the costs of effort at the destination airport and when outbound flight i has lower total costs of delay.

These two forces can be isolated from each other if we consider the ratio of coefficients scheduled to depart from the same airport in the same time slot. Since these flights share inbound flights, the ratio of the delay propagation coefficients will be equal to the inverse of the ratio of the corresponding total costs of delay. For example, if, according to the estimated coefficients, a minute of delay of the inbound flight "causes" 10 seconds of delay of flight A and only 5 seconds of

delay of flight B, then the direct and indirect marginal costs of delay of flight B is twice as much as the costs of delay of flight A.

5.1.6 Potential Reverse Causality

The model developed in the previous section allows us to explicitly state conditions under which the coefficients of the descriptive delay propagation regressions can have causal interpretation. The delay of inbound flights causes the delay of originating flights only if the unobserved shocks to costs of effort and delay are not known to the airline at the time it chooses the delay of the inbound flights. Arguably this assumption is strong and probably unrealistic.

Without this assumption, however, we will have an endogeneity problem. To see that, suppose that flight i received a favorable realization to the direct costs of delay that makes the delay less costly and, therefore, more likely. At the same time, that shock will decrease the indirect costs of delay for the inbound flight, since the total effort at this airport will go down. This decrease in indirect costs increases the delay of the inbound flights creating a somewhat mechanical correlation. It is not the case that the inbound flight "caused" the delay of flight i. Instead, lower realization of delay costs of flight i caused both the delay of flight i and the inbound flights. To estimate the effect of inbound delays, we need a shock that affects the delay of inbound flights independently of the cost shocks of flight i.

To construct a test and a procedure for correcting for the reverse causality effect described above, consider the following setting. Our null hypothesis is the absence of this effect: the delay of incoming flights is exogenous. The alternative hypothesis is the presence of correlation between the unobserved costs of delay to outgoing flights and the delay of incoming flights.

Our test is simple. Under the null hypothesis (and all other assumptions of the delay model), the realized delay of the incoming flights is a sufficient statistic for the costs of effort. In other words, any additional information about what happened earlier in the rest of the airline network should not matter. The delay of the incoming flights to the incoming flights (lag two delay) should not affect the delay of the outgoing flight conditional on knowing the delay of incoming flights only (lag one delay). Importantly, under the alternative, correlation between the delay of outgoing flight

and the delay of lag-two incoming flight (conditional on the delay of lag one incoming flight) will show up in the data. If the airline delays the incoming flight *because* it needs to (or wants to) delay the outgoing flight, it will delay the lag-two incoming flight as well.

In our data, we see some evidence for the statistical significance of the higher order delays suggesting that the reverse causality is likely to present a challenge for a causal interpretation of our estimates and should be addressed.

To address this issue, our model of aircraft scheduler tells us where to look for suitable instruments. In particular, the delays of "adjacent" flights to an incoming flight can be used as instruments for the delay of this incoming flight. For example, consider the delay of a flight from DFW to SAN, a plane for which is arriving from LGA, i.e., we want to estimate the effect of the delay to the LGA-DFW flight on DFW-SAN flight. Then all flights that depart from LGA to other destinations at the same time slot are valid instruments for the delay of the LGA-DFW flight as they are affected by the realizations of the shock to effort at LGA, but not by the realization of the shock at DFW.

We re-estimated our delay propagation model instrumenting in this way. If the reverse causality described above were present, we should see the coefficients decline as part of the attributed effects would be due to the "reverse." We find that to be the case. Qualitatively, the effects remain intact, the magnitudes, however decline by about 40%.

5.2 Estimation Results

We implement the estimation method described above on the sample of realized delays separately for each airline/month. By doing so, we allow for networks to differ by airlines and for scheduling adjustments on a monthly level. We thus have a K_{at} matrix summarizing the effect of a minute delay to a flight on the whole system of airline a in month t. We can now aggregate the K matrix along various dimensions to present the main results. For example, one can aggregate to an airport level by averaging over all flights departing from that airport. Doing so, we obtain a three-way panel of aggregated matrices K^O_{at} (defined in (6)) indexed by airline/month/origin (airport).

As an example of our estimation results, Figure 9 depicts (a subset of) the results of the elastic

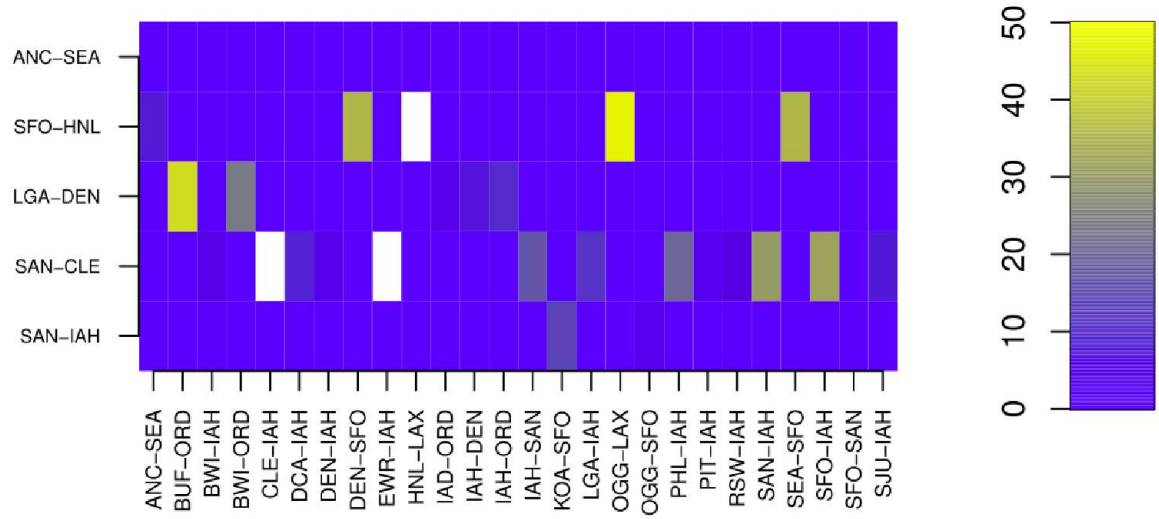


Figure 9: Overall effects: Jan 2012, United Airlines

net estimation for United in Jan 2012. It shows the effects of the 5 flights on the y-axis on the 26 flights on the x-axis.

5.2.1 Patterns of Delay Effects over Time and in the Cross-Section

Table 5 reports 10 airports with the largest effects based on our aggregated K-matrices during the early years of our data, i.e., 2010-2011. The column labeled *Total* reports the numbers of interest: for example, if we were to delay all flights at Seattle Airport on a random day by 1 minute, there would subsequently be additional 6,087 passenger minutes lost because of that.[12] Table 6 reports the same exercise for the later part of the data, i.e., 2012-2015. It is immediately visible not only that delays are becoming worse over time (average departure delays increased), but also that the indirect effects of delays (i.e., the effects of other flights down the road) became much more pronounced - with the large airports being the major sources of these effects.

Table 7 displays the passenger-weighted sum of Katz centrality measures (as defined in (6)) of individual flights aggregated over flights, months and airlines to annual level. An approximate interpretation is that a 1-minute delay to all flights by an airline from an airport translates into X

[12]Note that "own" effects are not counted. Furthermore, note that these estimates might occasionally "double-count" as some passengers may have a connected flight and delay of the first lag is not really a minute lost as long as they make their connection.

Table 5: Total delay effects: highest 10 airports from 2010 to 2011

Origin	Total[a]	Avg. Pass.[b]	Avg. Depdelay[c]	Depdelay2[d]
SEA	6086.9	37264.4	12.3	8.3
MIA	2675.2	25404.2	19.2	12.5
LAX	2087.6	59380.2	15.6	10.2
ORD	2059.1	72321.2	30.5	17.0
JFK	1960.0	31256.2	25.8	13.6
MSP	1789.3	39163.7	15.8	8.9
BOS	1763.2	31578.3	25.6	11.5
FAI	1611.4	1243.4	16.3	6.3
DEN	1595.0	67181.8	17.3	10.5
MCO	1585.4	44770.0	17.4	11.8

[a] Total avg daily passenger minutes delay effect at origin
[b] Avg. pass. is average daily passengers at origin
[c] Avg. Depdelay is the averaged topcoded departure delay per flight
[d] Last column is the departure delay conditional on non-cancellation.

Table 6: Total delay effects: highest 10 airports from 2012 to 2015

Origin	Total	Avg. Pass.	Avg. Depdelay	Depdelay2
BOS	8180.4	34489.0	22.1	11.7
SLC	7212.0	27056.2	14.6	10.7
SEA	7146.1	41235.6	16.2	10.5
PDX	6799.2	19320.8	11.7	8.8
SMF	6369.3	12057.7	14.5	9.6
LAX	5887.4	67220.0	14.7	10.6
JFK	5437.5	33332.4	21.9	13.4
ORD	4538.0	75261.6	24.7	16.8
SFO	4267.5	46942.8	18.9	12.4
ANC	3479.8	6006.9	14.0	11.1

minutes (reported in the table) of total passenger delay minutes down the road - not including this immediate delay. While these numbers seem small, one has to recognize that such averages involve a lot of very small numbers which may of course mask substantial heterogeneity.

Table 7: Daily total delay effects in passenger minutes: time average

Month	2010	2011	2012	2013	2014	2015
Jan	250.75	90.27	2437.79	161.13	201.20	224.71
Feb	76.05	154.93	165.07	163.14	149.68	369.69
Mar	201.99	180.39	188.21	285.10	253.64	143.88
Apr	209.71	193.73	220.68	147.35	259.00	298.44
May	134.19	169.91	184.36	171.73	199.71	179.23
Jun	147.92	187.51	248.89	174.21	192.30	229.60

5.2.2 Effects of Network Characteristics

Equipped with the estimates of matrix K (defined in (6)), we can now project the estimates on various characteristics of the network defined in Section 5.1.4. In Table 8 we present a projection of centrality measures on these various network characteristics. Most of the coefficients are qualitatively similar to what one might expect: hubs are more important and delays in hubs and more connected airports tend to spread more, delays at larger airports (in terms of passengers) are more important, networks in which nodes are more alike (those that have low standard deviation of the degree distribution) tend to have smaller delay propagation etc. Perhaps surprisingly, the HHI on the route doesn't seem to be significantly related to the delay propagation.

In Table 9 we allow for potentially heterogeneous impact of network characteristics in hub and non-hub airports. Most importantly, the competition variables now become significant: an airline operating on a less competitive route seems to suffer from less delay propagation. This could be driven both by its spending more effort to avoid delays such routes or by avoiding delays being simply more costly on more competitive routes.

5.3 Counterfactual 1: Under the hood of the "On-time Machine"

Our estimates for delay propagation allow us to evaluate counterfactuals for which it may be reasonable to assume that the reduced form of the model does not change. In our first such

Table 8: Regressions of Log (delay measures) on Network Characteristics

	sysmins (unwght)	sysmins (passenger-wght)	sysmins (realized delays& pass-wght)
	(1)	(2)	(3)
nhubs	−0.02	−0.04*	−0.06***
	(0.02)	(0.02)	(0.02)
hhig	−1.05	−0.63	0.14
	(0.70)	(0.70)	(0.61)
hubdummy	0.63***	0.62***	0.60***
	(0.09)	(0.09)	(0.08)
hhiairport	−1.13	−1.82**	2.01**
	(0.90)	(0.90)	(0.78)
airportpassN	0.02***	0.02***	0.01***
	(0.002)	(0.002)	(0.002)
nnodes	0.04***	0.04***	0.04***
	(0.01)	(0.01)	(0.01)
avgdistN	31.85***	35.18***	38.59***
	(8.37)	(8.38)	(7.30)
avgdistNsq	−14.50***	−15.60***	−17.37***
	(3.62)	(3.62)	(3.16)
netdensity	19.90***	20.07***	18.36***
	(5.29)	(5.30)	(4.62)
transindex	4.50*	3.09	3.17
	(2.39)	(2.39)	(2.09)
degreedistsd	−0.30***	−0.28***	−0.28***
	(0.07)	(0.07)	(0.06)
Constant	−15.32***	−18.06***	−17.05***
	(5.22)	(5.23)	(4.56)
R^2	0.05	0.05	0.07
Obs.	9,900	9,900	9,900

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9: Regressions of Log (delay measures) with Hub Interactions

	sysmins (unwghtd)	sysmins (passenger-wght)	sysmins (realized delays & pass-wght)
	(7)	(8)	(9)
nhubs	−0.02	−0.04**	−0.06***
	(0.02)	(0.02)	(0.02)
hhig	−0.53	−0.11	0.74
	(0.74)	(0.75)	(0.65)
hhigXhub	−1.60**	−1.63**	−1.70***
	(0.74)	(0.74)	(0.65)
hhiairport	−0.33	−1.13	2.69***
	(0.95)	(0.96)	(0.83)
hhiairportXhub	−3.66	−2.35	−5.18**
	(2.72)	(2.72)	(2.37)
airportpassN	0.04***	0.04***	0.03***
	(0.003)	(0.003)	(0.003)
passNXhub	−0.04***	−0.04***	−0.02***
	(0.004)	(0.004)	(0.003)
nnodes	0.03***	0.04***	0.03***
	(0.01)	(0.01)	(0.01)
nnodesXhub	0.01	0.004	0.01
	(0.01)	(0.01)	(0.01)
avgdistN	32.44***	35.77***	38.68***
	(8.36)	(8.37)	(7.30)
avgdistNsq	−14.79***	−15.90***	−17.42***
	(3.61)	(3.62)	(3.16)
netdensity	14.07**	14.31***	13.18***
	(5.55)	(5.55)	(4.85)
netdensityXhub	12.93***	12.65***	11.77***
	(3.78)	(3.79)	(3.31)
transindex	4.65*	3.08	3.09
	(2.77)	(2.78)	(2.42)
transindexXhub	−1.55	−1.06	−0.84
	(3.94)	(3.95)	(3.45)
degreedistsd	−0.32***	−0.30***	−0.29***
	(0.07)	(0.08)	(0.07)
degreedistsdXhub	0.07	0.07	0.08
	(0.09)	(0.09)	(0.08)
Constant	−15.22***	−17.96***	−16.74***
	(5.21)	(5.22)	(4.55)
R^2	0.07	0.07	0.09
Obs.	9,900	9,900	9,900

Note: *p<0.1; **p<0.05; ***p<0.01

counterfactual, we quantify the contributing factors to the observed on-time performance.

In the 1980s, American Airlines launched a series of TV ads in which they declared themselves "The On-time Machine" of the airline industry. Fast forward thirty-five years to 2015. Delta Air Lines applied for and was awarded the trademark for "The On-Time Machine" and since then promotes itself as such.

There are two competing explanations for the current success of Delta's on-time performance. Some attribute it to a better managed network and "hard work", in general. Our structural model that explanation corresponds to lower costs of effort, An alternative explanation is "pure luck": better weather at Delta's hubs. Indeed, Atlanta, Delta's main hub, has fewer negative weather shocks compared to American's Dallas-Fort Worth (or United's Houston).

To decompose these two effects, we perform the following counterfactual analysis. First, we estimate the distribution of shocks in Atlanta and Dallas-Fort Worth based on the residuals in our regressions for Delta and American, respectively. We then calculate Delta and American's on-time performance using their regression coefficients but replacing Atlanta's distribution of shocks with that of Dallas-Fort Worth and the other way around.

Table 10 compares the counterfactual results with the baseline scenario. The gap in the average on-time performance between Delta and American indeed shrinks. The difference in delays decreases from about 2 minutes in the base scenario to about 1.25 minutes in the counterfactual suggesting that weather ("pure luck") is indeed a contributing factor to Delta' s success and contributes about 38% of the difference. However, this performance gap does not disappear indicating that Delta may indeed have lower costs of effort, or equivalently, Delta is better at managing their operations more efficiently.

Table 10: Average Departure Delay (mins), Q1 2015

Airline	Base	Counterfactual
DL	8.68	9.07
AA	10.68	10.31

6 Structural Form Analysis

The final part of our analysis is to estimate the parameters of the structural form of the model. We use the structural model to achieve two goals. First, we separately identify and contrast the direct and indirect costs of delay. We show that both components are important for the airline's decision whether to delay a flight and by how much. Second, we use the structural form of the model to simulate the remaining two counterfactuals. In these counterfactuals, the reduced form of the model does not remain the same and, therefore, cannot be used for the counterfactual analysis.

6.1 Econometric Framework

To estimate the model more efficiently, we adopt a flexible parametric structure. First, consistent with assumptions adopted by the OR literature for the airline industry, we assume that the direct (marginal) costs of delay take the following form:

$$g(d_i) = g_i \times (1 + d_i)^{\alpha},$$

where g_i is a flight specific fixed effect and $\alpha \in [0, 1]$ is a parameter showing how quickly the marginal costs of delay increase with each additional minute of delay.

Second, we assume that the costs of effort at the origin airport are airport and time-of the-day specific and depend on the aggregate inbound delay realized during the day. Specifically, the costs of effort take the following form:

$$f(z_{as}) = f_{as} \times (1 + \bar{d}_{as})^{\theta},$$

where $\bar{d}_{as} = (\frac{1}{n_{as}} \sum_{i:\bar{a}_i=a,\bar{s}_i=s} d_i^{\gamma})^{1/\gamma}$ is the CES aggregated average inbound delay, n_{as} is the number of inbound flights, f_{as} is the airport–time-of-the-day fixed effect that captures the steady-state level of congestion, and $\theta \in [0, 1]$ is a parameter showing how fast the marginal costs of effort raise with each additional minute.

To be consistent with the parametric representation for the costs of effort, the indirect costs of delay then take the following form:

$$h_{\bar{a}_i\bar{s}_i}(d_i) = f_{\bar{a}_i\bar{s}_i}\theta(1+\bar{d}_{a_is_i})^{\theta-1}\frac{\partial\bar{d}_{a_is_i}}{\partial d_i}.$$

Our moment restrictions naturally follow from the optimality conditions developed in Section 3. The first set of restrictions normalizes the mean of each unobserved shock to zero. The second set of restrictions relies on the assumption that the indirect costs of delay have no unobserved component. However, to account for potential heterogeneity, we use the same identification strategy. Specifically, the delay of flights that depart from the same airport as the incoming flight but to different destination is used as an instrument for the indirect direct costs of delay.

Once these restrictions are set up, we apply the standard two-step GMM with an efficient weighting matrix.

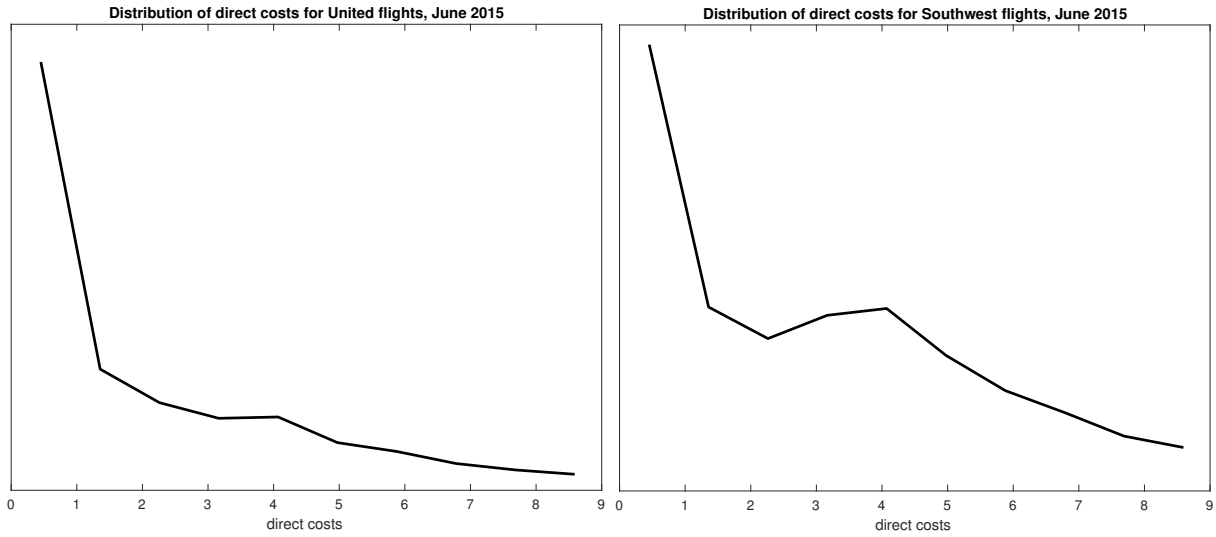
6.2 Estimation Results

Our estimated parameter is multi-dimensional. We estimate fixed effects for each flight and each combination of airport and time-of-the-day observed in the data. Instead of discussing each coefficient separately, we identify several stylized conclusion that can be drawn from these estimates.

First, there is substantial heterogeneity in the direct costs of delay across different flights within the same airline. This heterogeneity is higher for hub-and-spoke carriers. Figure 10 shows the distribution of direct costs of delay for United and Southwest.

Second, different airlines rank the same routes differently. For example, San Francisco – Newark is one of the costlier flights to delay for United, while for American the costs of delay of the San Francisco – JFK flights are relatively low. Generally, transcontinental flights and flights between distant hubs (e.g. Newark – San Francisco for United) have higher direct cost of delay.

Third, more congested airports and hubs have higher costs of effort, as do morning and evening flights. Incidentally, morning and evening flights have roughly the same costs of efforts. This finding implies that the higher observed delays of evening flights are mostly driven by the residual incoming delays rather than idiosyncratic factors like weather or long-term levels of congestion. Figure 11 shows the distribution of costs of effort for United across all airports for morning, early afternoon, and evening departures.



(a) United (b) Southwest

Figure 10: Distribution of direct costs of delay within an airline, June 2015

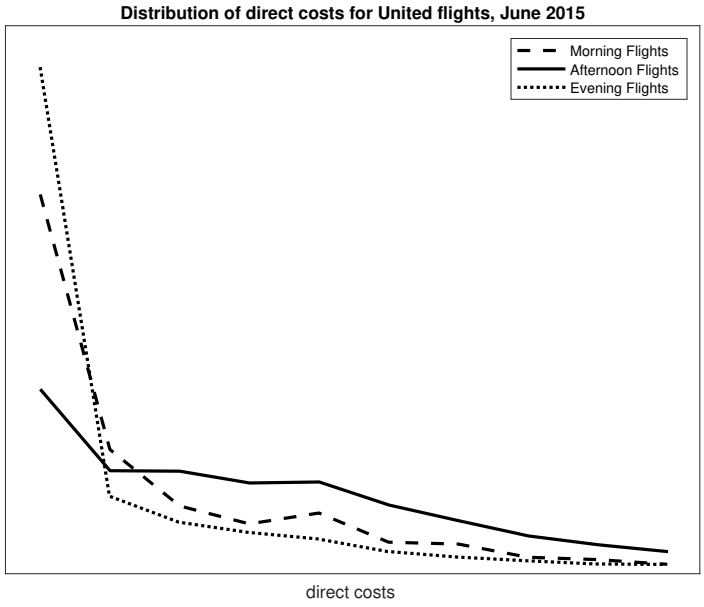


Figure 11: Distribution of direct costs of delay by departure time, United, June 2015

6.3 Counterfactuals

The methods developed in this paper allow us to illustrate the importance of accounting for network externalities for the airline industry. We will consider two counterfactuals. First, we ask the following question: how a common congestion-reducing infrastructure improvement benefits airlines

with different network? Second, we will estimate the network benefits of fleet homogeneity that airlines may pursue post-merger. Both questions are important for the industry and require a careful treatment of network effects.

6.3.1 Counterfactual 2: Local Infrastructure Improvements

Our second counterfactual seeks to evaluate the benefits of a common delay-reducing infrastructure improvement. To have a concrete example in mind, imagine that the manager of Boston Logan Airport considers the implementation of a delay-reducing infrastructure improvement (a new run-away, a set of new gates, an Air-Traffic control improvement). To finance this improvement, the manager needs to figure out how each airline benefits and by how much.

Traditionally, the costs of such projects are financed by the passenger-facility charges (PFC) added to the price of an airline tickets, common to all airlines. As a result, airlines that carry the larger share of passengers from the airport end up paying the larger share. The key advantage of the current system is its simplicity. A potential disadvantage is the possibility that those airlines who benefit the most may end up bearing the smaller share of the cost.

To formalize the question we are after, we assume that the delay-reducing infrastructure improvement reduces the costs of efforts at this airport (in every time slot) by 1%. Under this assumption, we can calculate how much each airline is going to save given this reduction. We then contrast these savings which the airline's share in each airport to see how close the current system of financing that relies on PFC is to the alternative that takes the network effects into account.

Formally, let λ_a be the percentage reduction of the cost of effort in airport a. Then the total cost function will take the following form:

$$C(\lambda_a) = \sum_{i \in \mathcal{I}} c_i(d_i) + \sum_{s=1,\ldots,S} \sum_{a \in \mathcal{A}} (1 - \lambda_a) c_{as} e_{as}$$

Using the envelope theorem, we can calculate by how much an incremental decrease in the cost of effort will reduce the optimal value of the total costs:

$$\frac{dC}{d\lambda_a} = \frac{\partial C}{\partial \lambda_a} = - \sum_{s=1,\ldots,S} c_{as} e_{as}$$

Table 11 presents the results. JetBlue is the largest airline of the Boston airport. Therefore, under the current financing system, JetBlue will end up paying the largest share of public good projects. A delay in the Boston airport, however, has a smaller impact on the overall performance of JetBlue's entire network than on that of American Airlines. The top two premium domestic markets of American Airlines are New York (JFK) – Los Angeles and New York (JFK) – San Francisco. Incidentally, American Airlines assign the same type of aircraft to their Boston – New York (JFK) market. Thus, the indirect cost of delays in Boston for American are significantly more than for JetBlue. It may very well be the case that American benefits significantly more than JetBlue from a delay-reducing infrastructure improvement in Boston. (International traffic may be another, equally important reason, but given the data limitations, there is little we can say about it.)

Airline	Market Share			Gains from a Decrease in Effort Costs
	by flights	by seats	by pax	
JetBlue	23%	24%	24%	21%
American + US Airways	19%	22%	21%	25%
Delta	9%	11%	11%	13%
Southwest	8%	7%	7%	6%
United	7%	9%	9%	11%

Table 11: Local improvement at BOS airport

Thus, somewhat counterintuitively, airlines that have *lower* realized delays and *lower* VAR coefficients are the ones that benefit most from infrastructure improvements. To see that, notice that airlines that *chose* to delay their flight before the improvement effectively reveal that other things equal, they have lower costs of delay and therefore won't gain much if costs of effort become incrementally lower. On the other hand, airlines that work really hard to push their planes on time do so because delay is costly for them. They will receive disproportionately larger gains if delays become less prevalent.

6.3.2 Conterfactual 3: Benefits from Merging Networks

Finally, the third counterfactual quantifies one of the potential benefits of an airline merger. The U.S. airline industry has recently experienced significant consolidation. Over the past 10 years, the number of players in the industry has decreased from ten to six. This trend has attracted increased interest from both the academic community and policymakers. Whenever the global trend on increased market concentration is brought up, the airline industry is the most cited example.

In the third and last counterfactual, we ask the following question: how to evaluate the merger benefits of network integration? To formalize these effects, we compare two scenarios. In the first scenario, the airline scheduler will minimize the total costs of effort over the entire network of the merged airline. In the second scenario, the costs will be minimized over each subnetwork separately, and then added together. Obviously, the sum in the second scenario cannot be lower than the value of the objective function in the first scenario. The percentage difference in the value functions for these two scenarios is a measure of the merger gains associated with the increased fleet homogeneity that airlines can advance as a pro-competitive defense.

We calculated these measures for the last two big mergers in the airline industry: American–US Airways (2015) and Alaska – Virgin (2016). Table 12 shows that the relative benefit from network integration is small. It is larger for the AA–US merger. These results should not be surprising. Neither merger proposal claimed network integration as its procompetitive justification. Pre-merger American and US Airways had little overlap in the types of aircraft they operated, while Alaska and Virgin had no overlap at all. Although the model does predict some benefits from operating a single network, their magnitude is not large enough to make a difference in the overall balance of pro- and anticompetitive effects of these particular mergers.

Table 12: Benefits from Postmerger Integration of Airline Networks

Airline Merger	Relative Decrease in the Overall Costs
American Airlines – US Airways	0.3%
Alaska Airlines – Virgin America	0.07%

7 Conclusion

Many social and economic processes involve network effects. The solvency of a financial institution depends on the network of its partners. The duration of a worker's unemployment depends on the network of her acquaintances. The chance that an adolescent commits a crime depends on the network of her peers. In this paper, we showed that—in the airline industry—answers to many important economic questions depend on the value of the network effects.

To quantify these effects, we developed a new set of econometric tools. Our data generating process is defined by a novel model that rationalizes the decisions of an airline scheduler. We used this model to show that the joint distribution of observed delays can identify the airline's perceived costs associated with delaying a flight. Importantly, both direct and indirect costs affect the decision to delay. The model allowed us to separate causal delay propagation from simple correlations. We saw that topological properties of the networks determine how quickly an airline recovers from shocks in different parts of the system.

As frequently reported, there is some substantial heterogeneity in on-time performance among U.S. airlines. Some experts attribute this heterogeneity to differences in management skills, others claim that the location of key hubs play a bigger role. We showed that these theories are not mutually exclusive and both forces find measurable support in the data.

We saw that the impact of local improvements on the performance of the entire air system crucially depends on the network externalities that these improvements generate. Investments to chronically delayed airports may have little overall impact as chronically delayed flights are typically those that are cheaper to delay. The benefits of a local improvement will naturally affect airlines differently. What determines the magnitude of these benefits, however, is not the airline's size at the airport but the role that this airport plays in the airline's entire network.

Finally, we evaluated the impact of a merger on delay propagation properties of the airline networks. Contrary to often raised claims by merging parties, we saw very limited evidence of these benefits. We conclude that the magnitude of these particular benefits is not large enough to change the overall balance in a merger evaluation.

References

Acemoğlu, Daron, Asuman Ozdaglar, and Alireza Tahbaz-Salehi, "Systemic Risk and Stability in Financial Networks," *American Economic Review*, 2015, *105* (2).

___ , Vasco Carvalho, Asuman Ozdağlar, and Alireza Tahbaz-Salehi, "The Network Origins of Aggregate Fluctuations," *Econometrica*, 2012, *80* (5), pp.1977–2016.

Bloom, Nicholas and John Van Reenen, "Why Do Management Practices Differ across Firms and Countries?," *Journal of Economic Perspectives*, March 2010, *24* (1), 203–24.

Bonaldi, Pietro, Ali Hortaçsu, and Jakub Kastl, "An Empirical Analysis of Funding Costs Spillovers in the EURO-Zone with Application to Systemic Risk," 2013. working paper.

Burzstyn, L., F. Ederer, B. Ferman, and N. Yuchtman, "Understanding Mechanisms Underlying Peer Effects: Evidence From a Field Experiment on Financial Decisions," *Econometrica*, 2014, *82*, pp. 1273–1301.

Carvalho, Vasco M., Makoto Nirei, Yukiko U. Saito, and Alireza Tahbaz-Salehi, "Supply Chain Disruptions: Evidence from the Great East Japan Earthquake," December 2016. working paper.

Chaney, Thomas, "The Network Structure of International Trade," *American Economic Review*, 2014, *104*, pp. 3600–3634.

Chiappori, Pierre-Andre, Denis Kristensen, and Ivana Komunjer, "Nonparametric Identification and Estimation of Transformation Models," *The Journal of Econometrics*, 2015, *188* (1), pp.22–39.

Conley, T. and C. Udry, "Learning About a New Technology: Pineapple in Ghana," *American Economic Review*, 2010, *100*, pp. 35–69.

de Paula, Aureo, "Econometrics of Network Models," in "Advances in Economics and Econometrics: Theory and Applications, Eleventh World Congress" 2017.

—— , **Imran Rasul, and Pedro Souza**, "Recovering Social Networks from Panel Data: Identification, Simulations and an Application," 2018. CeMMAP Working Paper 07/18.

—— , **Seth Richards-Shubik, and Elie Tamer**, "Identifying Preferences in Networks with Bounded Degree," *Econometrica*, 2018, *86* (1), pp.263–288.

Diebold, Francis X. and Kamil Yilmaz, "On the network topology of variance decompositions: Measuring the connectedness of financial firms," *Journal of Econometrics*, 2014, *182* (1), 119 – 134.

Elliot, Matt, Ben Golub, and Matthew Jackson, "Financial Networks and Contagion," *American Economic Review*, 2014, *104* (10), pp.3115–53.

Forbes, Silke and Mara Lederman, "Adaptation and Vertical Integration in the Airline Industry," *American Economic Review*, December 2009, *99* (5), 1831–49.

Graham, B., "An econometric model of network formation with degree heterogeneity," *Econometrica*, 2017.

Graham, Bryan, "Methods of Identification in Social Networks," *The Annual Review of Economics*, 2015, *7*, pp.465–485.

Horowitz, Joel L., "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 1996, *64* (1), 103–137.

Jackson, Matthew and B.W. Rogers, "Relating network structure to diusion properties through stochastic dominance," *B.E. Journal of Theoretical Economics*, 2007, *7* (1).

Jia, Jinzhu and Bin Yu, "ON MODEL SELECTION CONSISTENCY OF THE ELASTIC NET WHEN p $\gg$ n," *Statistica Sinica*, 2010, *20* (2), 595–611.

Manresa, Elena, "Estimating the Structure of Social Interactions Using Panel Data," November 2016. working paper.

Menzel, Konrad, "STRATEGIC NETWORK FORMATION WITH MANY AGENTS," April 2015. working paper.

Syverson, Chad, "What Determines Productivity?," *Journal of Economic Literature*, June 2011, *49* (2), 326–65.

Tibshirani, Robert, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996, *58* (1), pp.267–288.

US Department of Justice, *Horizontal Merger Guidelines*, U.S. Department of Justice, 2010.

Yuan, Ming and Yi Lin, "On the non-negative garrotte estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2007, *69* (2), 143–161.

Zou, Hui and Trevor Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of Royal Statistical Society B*, 2005, *67*, pp.301–320.

Online Appendix – Not for Publication

A Identification Results

In this section, we formally show how the results of Chiappori et al. (2015) establish the non-parametric identification of our model. As in the main text, we start with the reduced form that has the following form:

$$d_{it} = \mathcal{C}(f(z_{ast}) + \varepsilon_{ast} + \epsilon_{it}),$$

where i denotes the flight number, a denotes the origin airport, s denotes the departure time slot, t denotes the observation, which is collected daily, d is the realized departure delay, z is an observable that affects the observed cost of effort, ε is the unobserved post of the costs of effort, and ϵ is the unobserved part of the direct cost of delay, and $\mathcal{C}^{-1}(d) = g(d) + h_{\overline{a}_i \overline{s}_i}(d)$ is the observed part of the total costs of delay (direct and indirect). We define $u_t = \varepsilon_{ast} + \epsilon_{it}$.

To apply the results of Chiappori et al. (2015), we make the following assumptions:

Assumption 1 *For a.e. z from its support $\mathcal{Z}$, the conditional distribution $F_{u|z}$ is absolutely continuous (with respect to the Lebesgue measure on $\mathbb{R}$) with a density that is continuous on its support.*

Assumption 2 *The unobserved error u is independent of the cost shifters z.*

Assumption 3 *The support of d is a connected subset of $\mathbb{R}$) that contains zero.*

Assumption 4 *$\mathcal{C}$ is invertible that is monotonous and continuously differentiable on the support of d.*

Assumption 5 *$f(\cdot)$ is continuously differentiable with respect to z on $\mathcal{Z}$.*

Assumption 6 *The set of points where the partial derivative of the conditional distribution of d with respect to z does not equal to zero, is nonempty.*

Assumption 7 *There exists a set of instruments W such that: $E[u|W] = 0$ almost surely (a.s.).*

Assumption 8 *The conditional distribution of z given W is complete: for every function m such that $E[m(z)]$ exists, $E[m(z)|W] = 0$ a.s. implies $m(z) = 0$ a.s.*

Assumptions 1, 4, and 5 are technical. Assumption 2 requires the innovation to be independent on the factors that affect the inbound delay. Our section on reverse causality discusses the applicability of this assumption in detail. Assumption 3 is easy to verify: we observe flights with no delays and the maximum delay is naturally bounded in our data. Assumption 6 assures that the observed shifters of the inbound delay do vary in the data, which is true for our data.

Assumptions 7 and 8 are standard IV assumptions in the context of nonparametric IV. Intuitively. they state that the instruments are valid (assumption 7) and relevant (assumption 8). These two conditions are discussed in the main text (see Section 3.5).

Under these assumptions, the identification result directly follows from Theorem 1 of Chiappori et al. (2015).

Theorem 1 *Let Assumptions 1–8 hold. Then, $\mathcal{C}$ is globally identified up to a constant, g and $F_{u|z}$ are identified.*

Now that the identification of the reduced form is established, to complete the identification argument, we need to show how to separately identify the direct and indirect costs of delay. Since $\mathcal{C}^{-1}(\cdot)$ is identified, the sum of direct and indirect costs is identified for every flight. To identify these additive cost separately, we make two observations. First, the indirect costs of delay are a function of incoming delay, while the direct costs of delay are not. Second, the indirect costs of delay have to be the same for all flights departing from the same airport at the same time slot. The first condition implies that the partial derivative of $\mathcal{C}^{-1}$ with respect to inbound delay is equal to the partial derivative of the indirect costs of delay with respect to inbound delay almost everywhere. Thus, identification of the indirect costs follows from this differential equation. The direct costs of delay is simply the difference between $\mathcal{C}^{-1}$ and the indirect cost of delay up to a normalization. The second condition imposes this necessary normalization to the direct costs of delay.

B On Consistency of the Elastic Net Estimator

The Elastic Net Estimator was proposed by Zou and Hastie (2005) to combine the strength of the lasso and ridge regression. They demonstrate that the elastic net often enjoys better prediction performance than both the lasso and ridge regression in simulations. The literature on the asymptotic properties of the elastic net estimator is nascent, but rapidly growing. We rely on two main results that has established the path-consistency of the Elastic Net Estimator.

The first set of results establishes the path consistency of the Elastic Net estimator for the case when the number of regressors and the number of regressors with non-zero coefficient does not grow with the sample size (Yuan and Lin (2007)). Since in our asymptotics the number of incoming flight does not grow with each additional observation (i.e. another daily realization of joint delays), this set of assumptions directly applies to our setting. The path consistency result (Theorem 4 in Yuan and Lin (2007)) can be restated for our application in the following way. Let $\mathcal{I}$ denote the subset of incoming flights with non-zero delay propagation θ, $s_{\mathcal{I}}$ be a vector that has $\operatorname{sgn}\theta$ as its elements. Then:

Theorem 2 *A necessary and sufficient condition for the elastic net to be path consistent is*

$$\max_{j\notin\mathcal{I}}(\liminf_{c_1,c_2\to 0^+}[cov(D_j,D_{\mathcal{I}})\{cov(D_{\mathcal{I}})+c_1I\}^{-1}(s_{\mathcal{I}}+\frac{c_1}{c_2}\theta_{\mathcal{I}})])<1.$$

Even though in our setting the number of non-zero coefficients and the number of incoming flights stay the same as the sample size grows, an approximation that allows the number of these coefficient to grow might have better small sample properties. For this setting, Theorem 1 in Jia and Yu (2010) establishes that the Elastic Net Estimator will be path consistent as well.