Quantifying Delay Propagation in Airline Networks *

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

First version: September 2016
This version: December 2018

We propose a model of aircraft scheduler who allocates effort to minimize costs of delay on a network. We further develop a framework for quantifying delay externalities in airline networks and show how the effort cost can be identified. 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 then use these estimates to describe how network topology and other airline network characteristics (such as aircraft fleet heterogeneity) affect the expected delays. We also use the estimated effort cost to evaluate counterfactual scenarios of investments in airport infrastructure.

Keywords: Airline Networks, Delay Externalities, Elastic Net **JEL Classification:**

*PRELIMINARY AND INCOMPLETE, please do not cite or circulate without permission. We thank Eduardo Morales, Bo Honoré, Jan de Loecker, Jeremy Fox, and seminar participants at Harvard, Northwestern, Penn State, Princeton, Rice, Toronto, 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.

[†]Department of Economics, Princeton University
[‡]Department of Economics, Princeton University, NBER and CEPR
[§]Department of Economics, New York University

1 Introduction

Your flight is delayed. You are flying from New York to Los Angeles and your flight is delayed. "Weather," a friendly gate agents informs you. You do not buy this explanation. It is nice and sunny in New York. It is always nice and sunny in Los Angeles. So, where is the weather? The weather is in Boston. The aircraft assigned to your flight is stuck in Boston. Until it gets out of there, you are stuck in New York.

In this paper, we study how delay shocks propagate in airline networks. Our first goal is to understand how exogenous shocks experienced by distinct part 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. In this paper, we view the observed day-to-day realizations of flight delays as an outcome to a (perhaps, very complicated) single-agent optimization problem. This revealed preference approach allows us to develop a simple model of the data generating process that achieves several goals. First, delays of different flights throughout the network may be correlated for multiple reasons. We show what sources of variation in the data identify the causal effect 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 delay of an individual from the observed joint distribution of delays. Flights that are less expensive to delay get delayed longer and more often. We separate delay costs into direct and indirect parts. Delays directly inconvenience passengers on board of the current flight and indirectly make maintaining downline ontime performance harder. We show that direct and indirect costs are separately identified. Finally, we use tools developed for our network analysis to answer several economic questions.

The economic questions that we are after are as follows. First, we explore why some airlines systematically perform better overall than others. In our setting, the overall 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 both factors. Second, we estimate the global network effect of local improvement. We show that the overall network effect of a delay reducing improvement

may qualitatively differ from the local effect. Finally, we apply tools developed in the paper to 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, only quantifiable and verifiable synergies establish a pro-competitive defense of a horizontal merger.

The primary contribution of the paper is a new set of tools that can be applied to study a highly dimensional object. The tools, however, are only as good as the economic questions that they are able to answer. Even though some questions can be directly answered using the observed data, a structural model of the data generating process is oftentimes needed to state the set of underlying assumptions explicitly. In this paper, we show how a structural model of the data generating process complements and supports a descriptive analysis that relies on a simpler reduced form.

The three economic questions that we are after require a counterfactual analysis, although in some cases we do not need to resolve the underlying structural model completely. Instead, we rely on a "reduced form" that does not change for the conterfactual scenario of interest. Even then, however, we have to use the "structural form" to clarify the underlying assumptions of our analysis. The unifying conclusion of the three counterfactuals that we consider is: an analysis without network effects can lead to qualitatively different answers.

The focus of our theoretical and empirical analysis is a conditional distribution $Y|X$, where Y is a joint distribution of realized delays of all flights that an airline is operating during a day, and X is the airline's network characteristics. The main challenge for our analysis is the fact that both objects are highly dimensional. Major U.S. airlines operate thousands domestic flights a day. So, the dimensionality of Y is several thousands. The airline's network characteristics is an object with potentially higher complexity. First, it decodes 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. We have little theoretical guidance on which part of this information may turn out to be crucial. Our goal is to propose a set of tools that would allow us to figure it out. The scope and quality of available data determines which

economic questions we can address with the proposed tools. For example, since the delay shocks are observed in the data, our tools can be used to determine what would happen if shocks were exogenously reduced in one part of the network. 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 have data that with adequate exogenous variation that affects the number of available planes.

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, 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 specific conditions the estimates of descriptives 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 delay. 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 VAR analysis with one important distinction in mind. The asymptotic assumption of the textbook VAR analysis implies that the number of lags grows with the sample size. In out 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 instrumental variables, whose validity and relevance conditions are consistent with 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 and American Airlines and conclude that Delta's advantage can be attributed both to its superior network and to a more favorable distribution of shocks out of its major hubs. In other words, both "luck" and "hard work" are important to Delta's "on-time machine" brand.

Finally, 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 perform two counterfactual scenarios to answer the remaining 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 the local effect. First, although somewhat counterintuitive, it is not necessarily true that airlines experiencing fewer delay benefit from a delay reducing improvement less. 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. Second,

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, the airline that would benefit the most for a delay reducing improvement would be American Airlines. This finding is particularly reassuring since it turns out that American Airlines operates the same aircraft time 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 this premium service. Finally, we quantified the network benefits from United and Continental merger and found that... [TBA].

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 Menzel (2015), Manresa (2016), or Graham (2017) for further examples).

Our paper presents an alternative approach, which is based on ideas in Bonaldi, Hortaçsu and Kastl (2013). Since a researcher typically observes just one realization of shocks on the network (e.g., one realization of the decision to smoke or not to smoke), this literature has to make various assumptions to map the problem into one where some asymptotic arguments can be applied. For example, one needs to assume (analogously to strong mixing conditions in time series econometrics) that there are "islands" in the network that are sufficiently far apart, such that the shocks become asymptotically independent. In those applications, where data on repeated choices are

observed (such as in the case of financial network application studied in Bonaldi et al. (2013)), the problematic assumptions include first the network needs to stay fixed over time and second that the observation-specific shocks be independent. Whenever the first assumption fails, the proposed methods essentially identify an "average" network.

Our estimates of the weighted directed graph of each airline's network of flights allow us to express each flight's "systemicness" for each airline network, in a way similar to Bonaldi et al. (2013) that defines systemicness of banks in a financial network and Diebold and Yilmaz (2014) that defines financial connectedness. The measure we propose can be interpreted as the total number of minutes a one-minute delay to this flight would imply for the whole network. By aggregating this systemicness measure to the airport level, airline level, or studying its development over time, we are able to shed new light on the implicit cost of delays embedded in the commercial aviation in terms of passenger minutes. Using information on capacity utilization on individual routes and on average wages, we are also able to provide some estimates of monetary cost of these delays. Projecting this systemicness measure of airline network characteristics allows us to identify interesting correlation patterns.[1]

One important additional contribution 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

[1]It is important to recognize that there are two different network concepts at play here. First, and our main object of interest, is the network in which a flight represents a node and a link exists whenever delays transmit between two flights. Second, and the usual notion of airline network, is represented by airports being nodes and links existing whenever there is a flight between two airports.

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?

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 disputes. 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.

There are several other industries for which understanding how shocks propagate in networks is crucial (e.g. liquidity shocks in banking). 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.

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 will 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 (Southwest versus 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 a delay 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, the airline chooses which routes to serve. Then it assigns to each route capacity (the total number of available seat) and aircraft type, which determines frequency. Given the schedule, the airline scheduler solves the fleet assignment problem, which determines a sequence of lights 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 to solve it that way.

In fact, there may be different objective functions that the scheduling should optimize (differing for each airline, for example) which makes it virtually impossible for a unique method to solve the whole problem universally. There is a vast literature mostly published in journals devoted to transportation science that applies various methods from operation research (OR) to solve this

complicated scheduling problem.

The literature traditionally assumes, motivated by various industry sources, that the problem can be separated and solved sequentially. In other words, when the airline develops its schedule, it does not take into account how it would affect the subsequent stages. In this paper we will point to some results from the 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.[2] 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.[3] 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 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 the Form 41 report.

To keep the size of the data set manageable, we focus on Jan-Jun 2010-2015 and on 8 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

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

[3] 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.

airlines provides us with nice 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 will be 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. 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, the plane is assigned to two consecutive flights, that 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 amount of time. We, however, will not be using an OR type of model in our analysis for a number of reasons. First, the solution to the problem is likely not be unique. Apart from trivial relabeling, delaying a given aircraft by a minute is likely not 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 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 tha public data. It is possible of course to model these parameters as unobserved and integrate them out. We decided not to follow

it and keep the model as simple as transparent as possible instead.

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 arrive, the airline's irregular operations team adjusts the assignment trying to minimize the overall impact of these shocks on the airline's system. The plan 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 match. 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 shock 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.

When an airline chooses to delay a flight, it faces two types of cost: direct and indirect. The direct cost of a flight f being delayed by τ minutes is $c_f(\tau)$. An airline faces also an indirect cost of a delay of a flight due to the potential propagation of a delay through the network, i.e., through the subsequent scheduled flights. This can be because the airline deliberately holds some flight(s) back in order to accommodate connecting passengers arriving on delayed flights or due to physical constraints of having to wait for a particular aircraft or crew before the next flight can depart.

Airline and Flight Schedule

Let n be the number of flights that are scheduled to be performed during a day. Assume that the day is divided into T 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 $\underline{t}_i$, and scheduled arrival time $\overline{t}_i$.

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

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

$$e_{at} = \sum_{\underline{a}_i=a,\, \underline{t}_i=t} e_i$$

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_{at}.

Delays are costly too. We distinguish between direct and indirect costs of delay. Direct costs are costs that an airline has to incur because its 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 could make the problem of the aircraft scheduling team harder, which, in our model, means that the costs of effort at the destination airport go up. 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.

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:

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

Optimality Conditions

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

$$\underbrace{c_i'(d_i) \times \phi'(e_i)}_{\text{direct costs of delay}} + \underbrace{\frac{\partial c_{\overline{a}_i\overline{t}_i}}{\partial d_i} \times e_{\overline{a}_i\overline{t}_i} \times \phi'(e_i)}_{\text{indirect costs of delay}} + \underbrace{c_{\underline{a}_i\underline{t}_i}}_{\text{costs of effort}} = 0.$$

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 $c_{\underline{a}_i \underline{t}_i}$.

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 $\underline{t}_i = \underline{t}_j$, in equilibrium:

$$\left[c_i'(d_i) + \frac{\partial c_{\overline{a}_i \overline{t}_i}}{\partial d_i} e_{\overline{a}_i \overline{t}_i} \right] \phi'(e_i) = \left[c_j'(d_j) + \frac{\partial c_{\overline{a}_j \overline{t}_j}}{\partial d_j} e_{\overline{a}_j \overline{t}_j} \right] \phi'(e_j).$$

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 decrease its overall costs 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. This redistribution of effort would decrease airline's overall costs until these costs become equal to each other.

In the data, we do not observe the amount of effort exerted by airlines. 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 flight that airline chooses not to delay.

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.

Direct Costs of Delay. We assume that the direct costs of delay are multiplicatively 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 t, the marginal costs of effort are defined as follows:

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

where f is a deterministic function, and ε_{at} is a random mean-zero shock whose realization varies from day to day independently of other shocks. Cost shifters z_{at} include observable characteristics such as: a) realized inbound delay by period t, b) inbound cancellations, c) slack (#spare airplanes on the ground),... For example, if $f(z_{at}; \beta_z) = z_{at}\beta_z$, 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:

$$\frac{\partial c_{\overline{a}_i\overline{t}_i}}{\partial d_i} \times e_{\overline{a}_i\overline{t}_i} \times \phi'(e_i) = \frac{\partial f(z_{\overline{a}_i\overline{t}_i})}{\partial d_i} \times e_{\overline{a}_i\overline{t}_i} \times \phi'(e_i) = h_{\overline{a}_i\overline{t}_i}(d_i),$$

where $h_{\overline{a}_i\overline{t}_i}$ 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. To ease notation, let $p_{f,f'}(\tau, \tau')$ be the probability that a τ-minutes delay to flight f causes a τ'-minutes delay to flight f'.

Assumption 1 *There is no overnight delay effect, i.e., for any two flights f, f' that are scheduled*

on different days, $p_{f,f'}(\tau, \tau') = 0 \ \forall \tau > 0, \tau' > 0$

This assumption imposes 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_{at}. 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_{at}, 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.

Shock Propagation Mechanism

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 $c_{\overline{a}_i \bar{t}_i}$. This increased costs will affect all flights departing from $\overline{a}_i$ in slot $\bar{t}_i$ 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.

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 illustrates that it makes virtually no difference which one we use. 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 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. Facing exogenous factors airlines have to decide which flights to delay and by how much. The costs of these decisions will depend on the entire network of the airline. We observe a lot of variation in airline networks both over time and across different airlines.

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

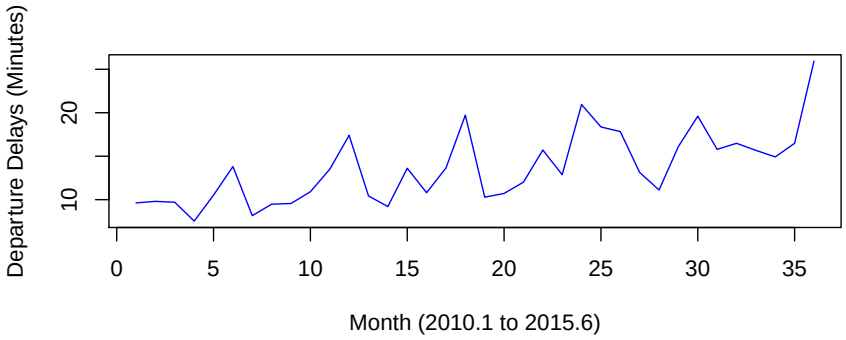


Figure 1: Monthly Average of United Airlines Departure Delays in minutes

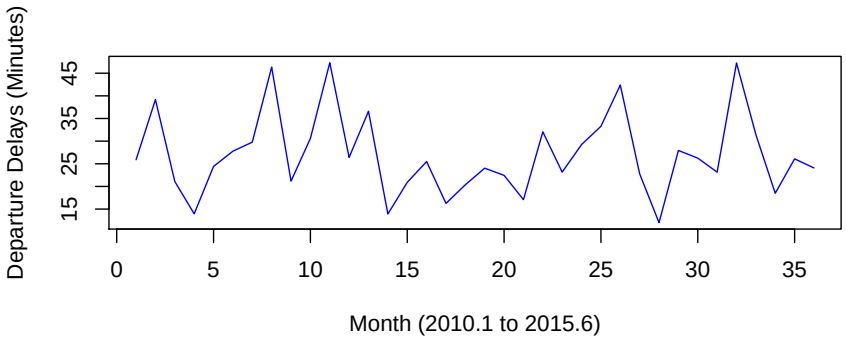


Figure 2: Monthly Average of American Airlines Departure Delays in minutes

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 1-month at a time.

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

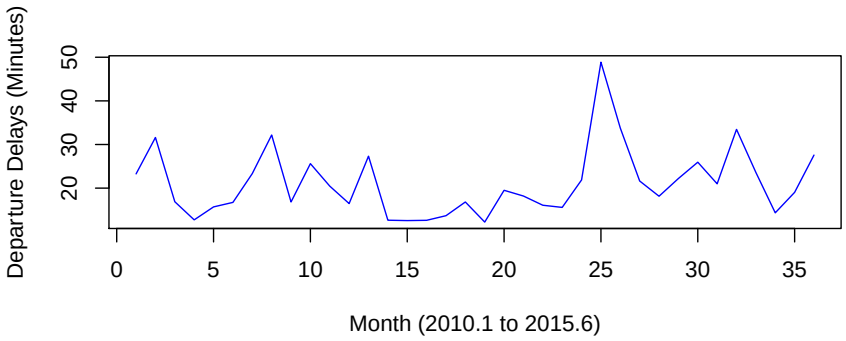
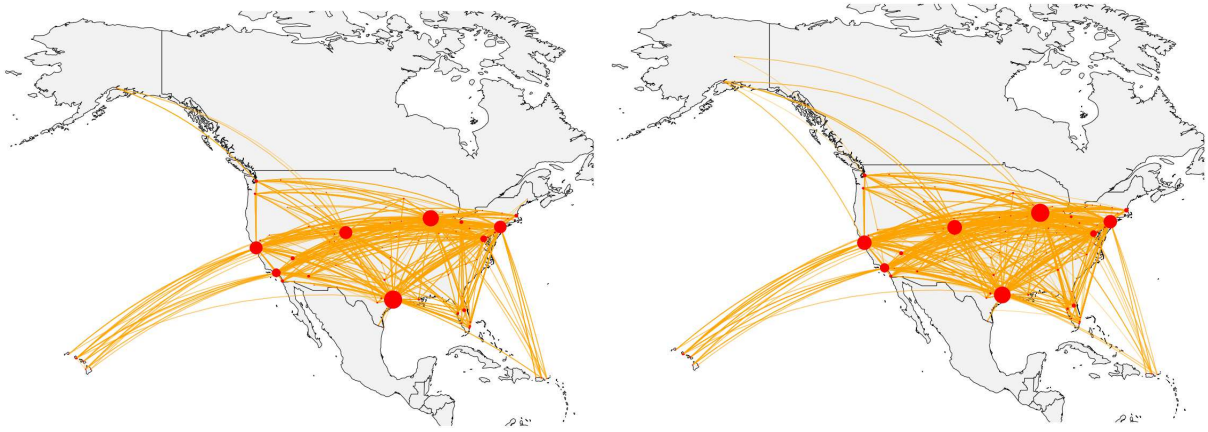


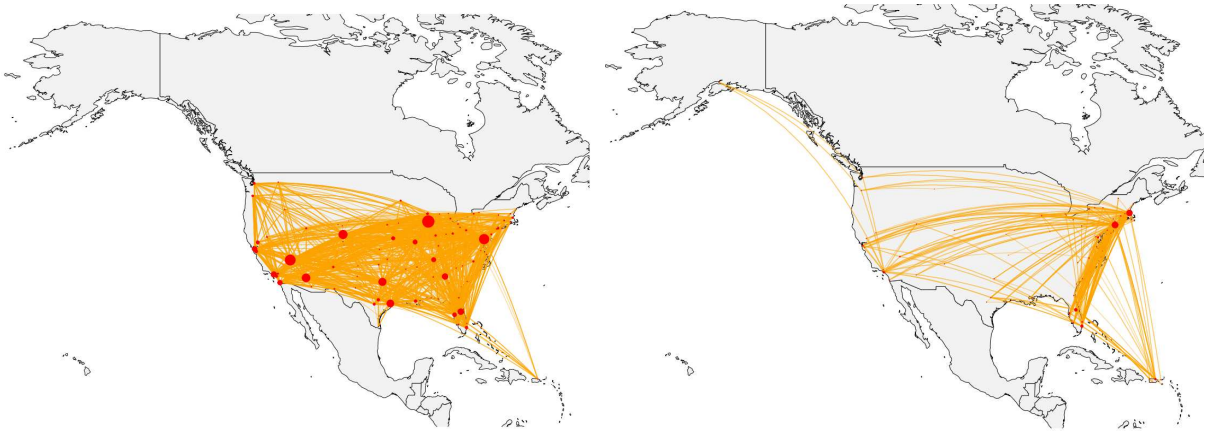
Figure 3: Monthly Average of Southwest Airlines Departure Delays in minutes



(a) January 2012

(b) June 2015

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



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

Figure 5: Networks of Southwest and Jetblue in June 2015

Table 2 reports one of our key airline network characteristics: the degree distribution. These measures are defined formally in Section 5. 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

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' 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

Turning to the data 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.

L-operator

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.

Fix a day d, and to economize on notation we will from now on drop the day-specific index. Consider the collection of all flights on this day, $\mathcal{F} \equiv \{\ldots, f_{ij}, \ldots\}$ in a given order, for instance, by origin, destination, scheduled departure time. Suppose $|\mathcal{F}| = n$. We first define a binary $n \times n$ matrix $L_1(\Delta)$. Whenever an (p, q)- element $L_{1,pq}$ is equal to one, the q-th flight in $\mathcal{F}$ is a "prior flight" of p-th flight in $\mathcal{F}$ and $L_{1,pq} = 0$ otherwise. 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. 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 $sgn(L_1)^2$, where $sgn(\cdot)$ denotes the sign function. The logic is $L_{2,pq} = \sum_{m \in \mathcal{F}} L_{1,pm} L_{1,mq} = 0$ if and only if there does not exist a flight as m-th element of $\mathcal{F}$ such that $L_{1,pm} = 1$ and $L_{1,mq} = 1$. Thus, as long as q is a prior flight of a prior flight of p, $L_{2,pq} \neq 0$ 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.

Delay VAR

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 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. Let the column vector of delays for all flights in $\mathcal{F}$ be denoted by D. Recall that $|\mathcal{F}| = n$ and hence $len(D) = n$. 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 $\mathcal{F}$.[4] We

[4] In our estimation, we will impose $K = 4$ due to computational constraints for most airlines and $K = 2$ for Southwest.

now specify the following statistical model for delays on a network by an airline:

$$D = c + \begin{pmatrix} \sum_{m\in\mathcal{F}} \beta_{1,1m} L_{1,1m} D_m \\ \vdots \\ \sum_{m\in\mathcal{F}} \beta_{1,nm} L_{1,nm} D_m \end{pmatrix} + \cdots + \begin{pmatrix} \sum_{m\in\mathcal{F}} \beta_{K,1m} L_{K,1m} D_m \\ \vdots \\ \sum_{m\in\mathcal{F}} \beta_{K,nm} L_{K,nm} D_m \end{pmatrix} + \varepsilon \tag{1}$$

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

where c is a vector of constants with length n, β_l is a $n \times n$ matrix for $l = 1, \ldots, K$. $\beta_{k,pq}$ denotes the k-lag delay effect of flight q on flight p if q is a prior flight of p and there is a delay in flight q. ε denotes exogenous delay shocks to each flight in $\mathcal{F}$. Notation "$\circ$" denotes element-wise matrix product (Hadamard product).

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

$$D = c + X\beta + \varepsilon, \tag{3}$$

where

$$X = (X_1, X_2, \ldots, X_K)_{n\times Kn^2}$$

and

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

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

$L_{l,1.}$ denotes the first row of L_l. $vec()$ 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 adaptive elastic-net regression (Zou and Hastie 2005, Zou 2006), 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 Assumption 1. The elastic net estimator is then simply a solution to:

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

where $Z = \begin{bmatrix} 1 & X \end{bmatrix}, \theta = \begin{bmatrix} c & \beta \end{bmatrix}$, 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.[5] 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.

Inference. A valid inference procedure for the LASSO-type estimators is a well-known problem.[6] We follow the method described in Chatterjee and Lahiri (2011) based on a modified residual bootstrap. It works roughly as follows: First, use the estimated model to construct residuals and re-center them. Construct a bootstrap sample of re-centered residuals. Recreate the LHS variable using the estimated model to create a bootstrap sample on which the model is re-estimated. However, the components of the estimator are thresholded (hence "modified" bootstrap)[7]: whenever a particular estimate is below a threshold, it is set to zero. Chatterjee and Lahiri (2011) show in Theorem 3.1 that this procedure is strongly consistent (pointwise) under some regularity conditions. They also provide a method for the appropriate choice of the thresholding parameter and illustrate the performance of their procedure in various Monte Carlos.

[5]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 use a modification of the *glmnet* package in R, which we modify to allow for non-negativity constraints on the parameters.

[6]See e.g., Knight and Fu (2000).

[7]Chatterjee and Lahiri (2010) show that a standard residual bootstrap (i.e., one without the thresholding) is actually inconsistent.

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

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

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_{pq} can be interpreted as the long run effect of a minute delay shock to flight q on flight p. Then $k_q = \frac{1}{n} \sum_{p \in \mathcal{F}} K_{pq}$ can be used to measure average effect a minute delay in flight q on the rest of flights in $\mathcal{F}$. It can be shown that under certain regularity conditions, some desirable properties of the adaptive elastic-net estimator (e.g. consistency in model selection) translate also into k_q. Note that this matrix 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).

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), D_{ij}: $\binom{N}{2}^{-1} \sum_{i=1}^{N} \sum_{j<i} D_{ij}$. 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.

To illustrate the variation in the network characteristics that is available to us, 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).

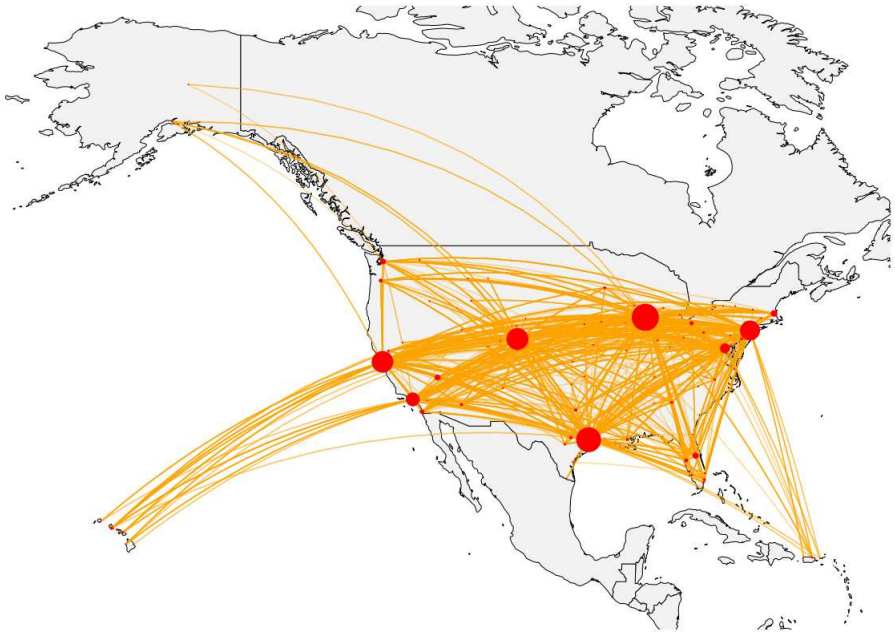


Figure 6: United Airlines, June 2015

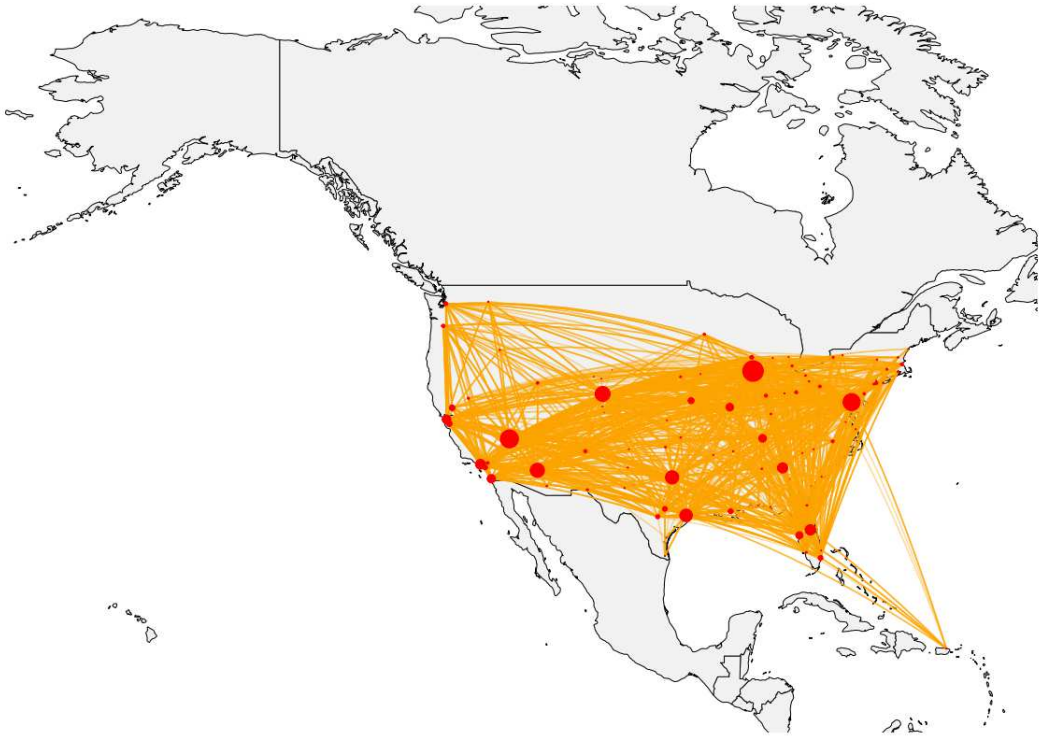


Figure 7: Southwest Airlines, June 2015

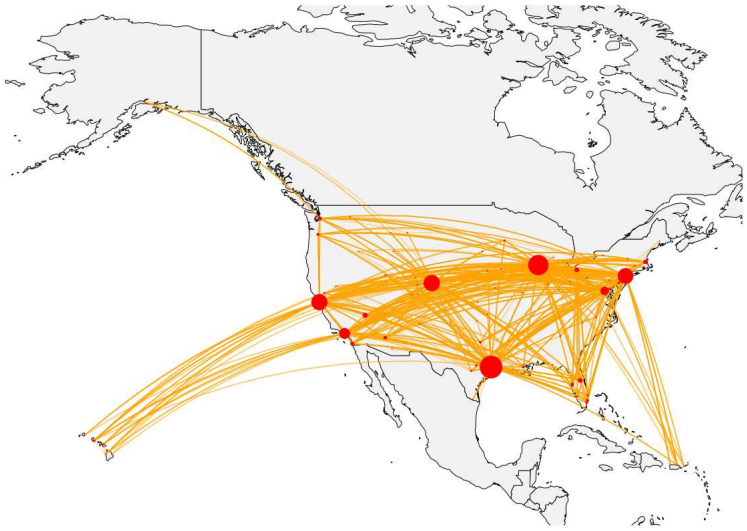


Figure 8: United Airlines, January 2012

5.1 Identification Challenges and Reverse Causality

To understand what the descriptive regressions may uncover, we go back to the model defined in the previous section and investigate the reduced form for observed delays. Recall that the optimality

conditions imply:

$$d_i = T(f(z_{at}) + \varepsilon_{at} + \epsilon_i).$$

Thus, the realized delay is a function of 1) observed costs of effort at the origin airport, 2) observed (total) costs of delay, 3) unobserved costs of effort at the origin airport, 4) unobserved (direct) costs of delay. The shock propagation mechanism defined in this paper postulates that the delay of incoming flight will affect the (observed) costs of effort at the destination airport and, through them, the delay of outgoing flights. However, this is not the only mechanism that creates correlation between the incoming and outgoing flights.

The indirect costs of delay depend on the effort and, therefore, delay, of outgoing flights at the destination. If an outgoing flight received a negative shock to the direct costs of its delay, the airline's scheduler will apply less effort to this flight and delay it longer. That will reduce the total effort at the destination and therefore the indirect costs of delay for incoming flight. To reduce overall costs, the scheduler will delay the incoming flight as well. This is a classic situation of reverse causality. The incoming flight is causing the delay of the outgoing flight. On the contrary, the incoming flight is delayed *because* the outgoing flight is delayed.

From the econometric standpoint, the presence of reverse causality will make OLS estimates of the VAR regressions inconsistent due to endogeneity. To be consistent, the innovation in the delay of outgoing flight has to be uncorrelated to the delay of the incoming flight, which is not the case in the presence of reverse causality as argued in the previous paragraph.

One approach to dealing with this problem is simply to assume this effect away. Indeed, if the costs of delay for outgoing flights are not known to the scheduler when the decisions about incoming flights are made, then the reverse causality effect should be absent in the data. Under this assumption, the delay of incoming flights will be treated as exogenous, making OLS estimates consistent.

This assumption, however, is falsifiable and can be tested in the available data.

5.2 Statistical Test for Endogeneity

To construct a test for the presence of reverse causality effect described in the previous section, consider the following setting. Out 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 null (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 I delay the incoming flight *because* I need to delay the outgoing flight, I will delay the lag-two incoming flight as well.

Thus, under the null hypothesis, the coefficients before the higher order lags of delay should not be statistically significant. Under the alternative, the presence of reverse causality will make all lags mechanically correlated even when the delay of incoming flights does not cause the delay of outgoing flights.

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

5.3 IV Estimation

5.4 Interpretation of the Coefficients

The model developed in the previous section allows us to explicitly state conditions under which the coefficients of the descriptive VAR 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 costs shocks of flight c_i. An example of such variation was discussed in Section **??**.

As long as we are willing to assume away this endogeneity issue, we can interpret the VAR coefficients using the language of the structural model.

Recall that the optimality conditions imply:

$$d_i = T(f(z_{at}) + \varepsilon_{at} + \epsilon_i).$$

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

$$\frac{\partial d_i}{\partial d_j} = T'(f(z_{at}) + \varepsilon_{at} + \epsilon_i)\frac{\partial f(z_{at})}{\partial d_j}.$$

Under these assumptions, the VAR coefficients approximate the local average value of the left-hand side of this equation.

Other things equal, we should expect higher VAR 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 VAR coefficients will be equal to the inverse of the ratio of the corresponding total costs of delay. For example, if, according to the VAR estimates, 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.

5.5 Estimation Results

We implement the estimation method described above on the sample of realized delays. Essentially, our approach can be viewed as asking the following question: Given the mechanical constraints, given the airline making optimal decisions conditional on their relevant information and given the observed delay realizations, what is the realized average spillover effect of a minute delay of a flight on other flights on that airline's network? The effect that we estimate is thus a combination of the direct (or mechanical) and indirect (arising from airline's optimal choices) effects. The former effect can be viewed as coming from an airplane or its crew being scheduled to continue elsewhere on different flight and hence a delay will "cause" a delay on this scheduled flight mechanically. The latter effect can be described as an airline deciding based on the current situation to delay a flight operated by a different crew and served by a different aircraft. While we will not be able to decompose these two effects, we will still provide some more detailed analysis making use of the fact that at least the graph of aircraft schedules is observed (unlike the graph of crew schedules).

By implementing our estimation procedure separately for each airline/month 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 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.

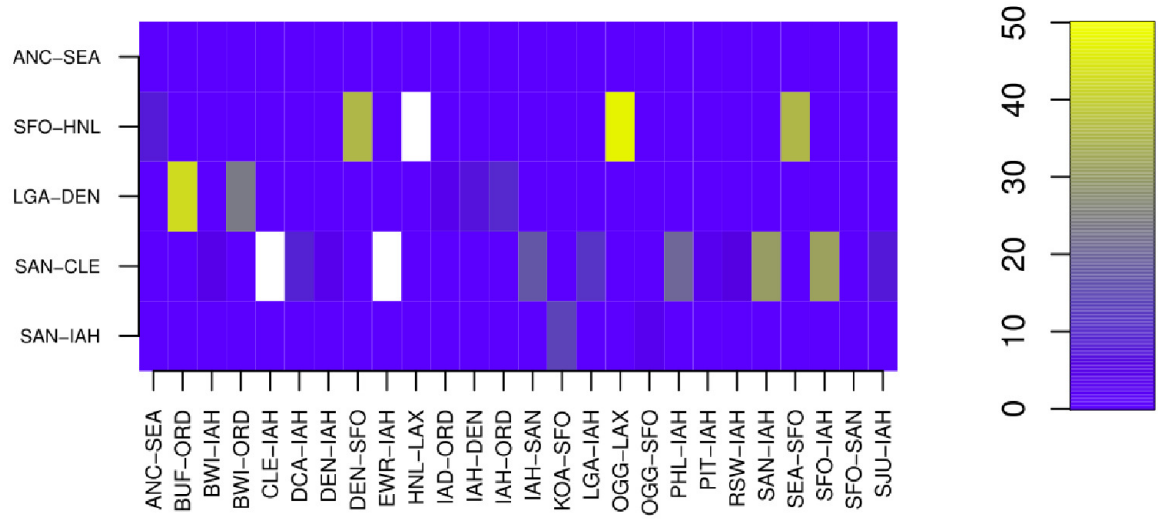


Figure 9: Overall effects: Jan 2012, United Airlines

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

Table 5 reports 10 airports with 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.[8] 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 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.

[8]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

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

5.7 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. In Table 8 we present a projection of centrality measures on these various network characteristics. Most of the coefficients are qualitatively similar with 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.

5.8 Decomposition of Direct and Indirect Costs

Due to network propagation, a flight delayed in the morning may have a qualitatively different impact on the entire airline's performance compared to a flight delayed closer to the end of the day. To quantify this effect systematically, we.... [TO BE COMPLETED]

5.9 Estimated versus Observed Network

As previously mentioned, we can trace out partially how the delays should be mechanically transmitted by utilizing the information on tail numbers. When we define an unweighted directed graph where flights are still nodes, but links between two flights equals one if and only if the same aircraft with the same tail number is operating both flights and one flight is an immediate predecessor of the other. We then apply a sign function to our estimated weighted directed graph of spillovers. We convert both matrices into vectors and ask how these two vectors are "correlated." Depending on airline/month, we get correlations in the There are several sources of variation that we will exploit

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

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. Facing exogenous factors airlines have to decide which flights to delay and by how much. The costs of these decisions will depend on the entire network of the airline. We observe a lot of variation in airline networks both over time and across different airlines. neighborhood of 0.5. While this clearly cannot be interpreted as evidence that the proposed procedure works and that 50% of the spillovers should be attributed to the "mechanical" spillovers and 50% to the unobserved part (i.e., crews scheduling, holding back flights due to connecting passengers etc), we believe that it is suggestive that our elastic net procedure is picking up at least some important links.

5.10 Counterfactual 1: The "On-time Machine"

The VAR estimates for delay propagation allows 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 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 compated 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 the VAR's regressions for Delta and American, respectively. We then calculate Delta and American's on-time performance using their VAR 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 decreases in the counterfactual suggesting that weather ("pure luck") is indeed a contributing factor to Delta' s success. However, this gap does not disappear indicating that Delta may indeed have lower costs of effort.

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

Airline	Base	Counterfactual
DL	8.68	9.07
AA	10.68	10.31

6 Structural Analysis

6.1 Estimation Method

Identification Strategy

We observe the joint distribution of realized delays d_i together with covariates that affect $f(\cdot)$, $g(\cdot)$, and $h_{at}(\cdot)$. The optimality conditions derived in the previous section rationalize the observed delay of each flight i:

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

The observed delay, d_i, is thus an (unknown but deterministic) function of 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 have the following reduced form:

$$d_i = T(f(z_{at}) + \varepsilon_{at} + \epsilon_i),$$

where T is an unknown transformation. This class of econometric models known as "regression models with an unknown transformation of the dependent variable" was pioneered by Horowitz

(1996). Chiappori et al (2015) further extends the analysis of these models by providing a set of sufficient conditions that guarantee that the the unknown functions $T(\cdot)$ and $f(\cdot)$, together with the distributions of the unobserved shocks are non-parametrically identified. We will not restate these conditions explicitly. 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_{at}. The first condition ("relevance") ensures that the observable part of the costs of effort at the origin airport, $f(z_{at})$, varies from observation to observation. Without such variation we cannot identify the function $f(\cdot)$. The second condition ('validity") requires the unobservable part, $\varepsilon_{at} + \epsilon_i$, to be independent of the cost shifter, z_{at}. 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).

Identifying Sources of Variation.

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_{at}), observed delays to other flights that leave from the same 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_{at} 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.

Separating Direct and Indirect Costs of Delay. So far we have established the identification of the total costs of delay. To identify the direct and indirect costs of delays separately, 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.

6.2 Estimation Results

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 implications.

6.4 Local Infrastructure Improvements

Our first 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 implementation of a delay-reducing infrastructure improvement (a new runaway, 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 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.

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, may have a smaller impact on the overall performance of JetBlue's entire network that 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.)

To see whether or not this conjecture is true, we will formalize the question we are after as follows. 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:

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

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:

$$\frac{dC}{d\lambda_a} = \frac{\partial C}{\partial \lambda_a} = - \sum_{t=1,\ldots,T} c_{at} e_{at}$$

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 thing 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 gain if delays become less prevalent.

[Numerical results to be added]

6.5 Fleet Homogeneity

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.

Every time two airlines merge, antitrust authorities scrutinize the potential effects of such consolidation. The negative short-term effects of a horizontal merger are straightforward: fewer players imply more market power, which leads to an increase of the market price. To alleviate these antitrust concerns, airlines traditionally advance two arguments: new entry and cost-saving efficiencies. If post-merger entry is likely, timely, and sufficient, then the long-run number of active firms in each market is not affected by the merger. Thus, any post-merger increase of market

power (if any) will be just temporary. To demonstrate cost-saving efficiencies, the merging airlines have to establish that these efficiencies cannot be achieved without a merger. When two airlines merge, they effectively merge their networks. Even though newly merged airlines may initially operate their legacy subnetworks separately, eventually they achieve full integration and decrease their fleet heterogeneity.

In this 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.

[Numerical results to be added]

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 implications.

6.5.1 Local Infrastructure Improvements

Our first 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 implementation of a delay-reducing infrastructure improvement (a new runaway, 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 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.

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, may have a smaller impact on the overall performance of JetBlue's entire network that 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.)

To see whether or not this conjecture is true, we will formalize the question we are after as follows. 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:

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

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:

$$\frac{dC}{d\lambda_a} = \frac{\partial C}{\partial \lambda_a} = -\sum_{t=1,\ldots,T} c_{at} e_{at}$$

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 thing 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 gain if delays become less prevalent.

[Numerical results to be added]

6.5.2 Fleet Homogeneity

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.

Every time two airlines merge, antitrust authorities scrutinize the potential effects of such consolidation. The negative short-term effects of a horizontal merger are straightforward: fewer players imply more market power, which leads to an increase of the market price. To alleviate these antitrust concerns, airlines traditionally advance two arguments: new entry and cost-saving efficiencies. If post-merger entry is likely, timely, and sufficient, then the long-run number of active firms in each market is not affected by the merger. Thus, any post-merger increase of market power (if any) will be just temporary. To demonstrate cost-saving efficiencies, the merging airlines have to establish that these efficiencies cannot be achieved without a merger. When two airlines merge, they effectively merge their networks. Even though newly merged airlines may initially operate their legacy subnetworks separately, eventually they achieve full integration and decrease their fleet heterogeneity.

In this 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.

[Numerical results to be added]

7 Conclusion

We present a model of an aircraft scheduler who optimally allocates effort to minimize cost of delays. We show that using observational data on delays of individual flights one can identify the underlying cost of effort. To this end, we propose a method that allows us to estimate the "network effects" between flights (i.e., spillovers of delays across flights) by utilizing a model selection algorithm to reduce the dimensionality of the problem. Using our results we evaluate the effectiveness of cost sharing in airport infrastructure investment.

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.

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.

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.

Chatterjee, A. and S. N. Lahiri, "Asymptotic properties of the residual bootstrap for Lasso estimators," *Proceedings of the American Mathematical Society*, 2010, *138*, 4497–4509.

___ and ___ , "Bootstrapping Lasso Estimators," *Journal of the American Statistical Association*, 2011, *106* (494), 608–625.

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

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.

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.

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

Knight, K. and W. Fu, "Asymptotics for lasso-type estimators," *The Annals of Statistics*, 2000, *28*, 1356–1378.

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.

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

Zou, Hui, "The Adaptive Lasso And Its Oracle Properties," *Journal of the American Statistical Association*, 2006, *101* (476), pp.1418–1429.

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