Pricing of Complements in the U.S. freight railroads: Cournot versus Coase

Alexei Alexandrov and Russell Pittman and Olga Ukhaneva

Antitrust Division, U.S. Department of Justice

18 April 2018

Pricing of Complements in the U.S. freight railroads: Cournot versus Coase[*]

Alexei Alexandrov[†] Russell Pittman[‡] Olga Ukhaneva[§]

April 3, 2018

Abstract

Monopolists selling complementary products charge a higher price in a static equilibrium than a single multiproduct monopolist would, reducing both the industry profits and consumer surplus. However, firms could instead reach a Pareto improvement by lowering prices to the single monopolist level. We analyze administrative nationally-representative pricing data of railroad coal shipping in the U.S. We compare a coal producer that needs to ship from A to C, with the route passing through B, in two cases: (1) the same railroad owning AB and BC and (2) different railroads owning AB and BC. We do not find that price in case (2) is higher than price in case (1), suggesting that the complementary monopolist pricing inefficiency is absent in this market. For our main analysis, we use a specification consistent with the previous literature; however, our findings are robust to propensity score blocking and machine learning algorithms. Finally, we perform a difference-in-differences analysis to gauge the impact of a merger that made two routes wholly-owned (switched from case 2 to case 1), and these results are also consistent with our main findings. Our results have implications for vertical mergers, tragedy of the anticommons, mergers of firms selling complements, and royalty stacking and patent thickets.

[*] Keywords: vertical merger, double marginalization, mergers of complements, antitrust, tragedy of the anticommons, royalty stacking, patent thickets, intellectual property. JEL codes: D23, D43, K21, L40, L92, O34. The authors thank the participants in the International Industrial Organization Conference (and especially our discussant Mihkel Tombak), the participants in the American Law and Economics Association Conference, the participants in the Georgetown Rail Colloquium, the participants in the 25th anniversary conference of the New Economic School, and Gerard McCullough and Wesley Wilson for extensive comments.

[†] The views of the authors do not represent the views of Alexei's past or current employers. Email: alexei01@gmail.com.

[‡] U.S. Department of Justice. The views of the authors do not represent the views of the U.S. Department of Justice. Email: russell.pittman@usdoj.gov.

[§] Navigant and Georgetown University. The views of the authors do not represent the views of Navigant. Email: ukhaneva.olga@gmail.com.

1 Introduction

One of the oldest issues in economics is the pricing of complements, with formal treatment dating back to Augustin Cournot's 1838 treatise (Cournot (1897)). One needs both copper and zinc to make brass. Suppose that firm A is a monopolist selling copper and firm B is a monopolist selling zinc to brass producers. Cournot showed that the sum of the prices of A and B is more than what a monopolist selling the combination would charge. Prices are strategic substitutes, and each firm has an incentive to raise its price higher than the single-monopolist level, since the other firm has an incentive to lower its price in response, in a sense, to subsidize the other firm's price increase. In the resulting equilibrium, these complementary monopolists are worse for society than a single monopoly: the prices are higher than a single monopolist charges, hurting both the consumers and the firms themselves.[1]

However, there is a clear potential Pareto improvement: the firms could lower prices to the single-monopolist level. Researchers, for example, Coase (1960), have postulated that, absent transaction costs, agents with a potential Pareto improvement should be able to arrive at a Pareto efficient outcome. In our setting, there could be several mechanisms of arriving to a Pareto improvement: Schumpeter (1928) suggests, effectively, a collusion between the two monopolists, while Spulber (2017) shows that a Pareto optimal solution can be achieved through negotiations of monopolists with producers.

Thus, the question is an empirical one: do the firms act as static complementary monopolists and arrive at a Pareto suboptimal outcome, as predicted by Cournot, or do they figure out a way not to leave money on the table, as predicted by Coase and other researchers?[2] We analyze this question in the context of the U.S. freight railroads shipping coal. Controlling for the relevant route characteristics, we do not find that the price of shipping through two firms is higher than the price of shipping through one. Thus, we argue that in the case of the U.S. freight railroads shipping coal, the Coasian prediction fits the data better. From the historical perspective, it is notable that in 1839 – a year after Cournot published his work in France – Charles Ellet, an American engineer, published an analysis of railroad pricing in the United States, using similar calculus-based methods and making many points similar to Cournot's 1838 work, see Calsoyas (1950) for a review.

Aside from other applications, our results should be of interest for the railroad industry: coal, together with intermodal freight, are the two largest rail commodities by revenue, with coal accounting for about 17% of freight railroads' revenue in 2015, see Association of American Railroads (2016). Conversely, about 70% of coal is shipped via freight rail. Our results suggest that an end-to-end railroad merger would not be needed to fix a Cournot or double-marginalization-like pricing issue.[3]

[1] Sonnenschein (1968) formalized Cournot's arguments and showed that Cournot duopoly is the dual of this complementary monopoly.

[2] We refer to this possibility of Coasian prediction as a shorthand, without trying to disentangle whether the underlying mechanism is closer to the one suggested by Schumpeter (1928) or the one suggested by Spulber (2017).

[3] Nonetheless, we do not analyze other, more traditional merger efficiencies, and thus do not address the question of whether a vertical railroad merger could indeed result in lower prices to consumers for reasons of other efficiencies

We describe other applications below, including royalty stacking in intellectual property, vertical double marginalization (and mergers), and the tragedy of the anticommons. We also describe our method and why we believe that the railroad data set presents a unique opportunity to study this question.

While making brass is important, arguably the most contested recent application of this analysis is in the area of intellectual property. Hundreds of different complementary patents might be needed to make a modern device, for example a cell phone. The concepts of royalty stacking or patent thickets (see Shapiro (2001) and Lemley and Shapiro (2006)), are effectively Cournot's analysis applied to intellectual property. The policy implications are clear: having many different patent holders that are all needed to make a product would result in an inefficient outcome with individual patent license rates (prices) that are too high. Other researchers, soon after, suggested that there are mechanisms in the market that prevent this Pareto-inefficient outcome (see, for example, Geradin, Layne-Farrar, and Padilla (2007), Elhauge (2008), and Sidak (2008)). Even more relevant to our paper, see Spulber (2016) and Spulber (2017) for formal models of how bargaining between complementary input monopolists and producers can lead to avoidance of the inefficient Cournot outcome.

We discuss below some of the empirical research relating to patents and royalty stacking; however, the common thread is significant data limitations. In short, there is no representative database where one could observe royalties that each manufacturer pays to each patent holder, as these negotiations are highly confidential and situation-specific (see Hagiu and Yoffie (2013)). Even non-representative databases are non-existent as far as we know.[4] Thus, the existing empirical research had to rely on secondary indicators, for example the size of the patent portfolio, the proclivity to patent, and the firms' market value.

In contrast, we use years of nationally-representative administrative data on pricing of railroad freight, allowing a direct test of the theory. Consider a shipper that wants to transport goods from A to C, a route that passes through B. We estimate whether the price paid from A to C is the same when the same railroad owns both AB and BC tracks as it is when one railroad owns AB and another owns BC. We estimate difference in prices by comparing otherwise similar routes, with the difference being whether the route is wholly-owned. We control for available characteristics of the route, including competitiveness of the railroads, and use a comparatively homogeneous product – coal.

For identification, we believe that our treatment variable – whether a route is wholly-owned by the same railroad – is nearly as good as random conditional on observables. The specific junctions – points where railroads meet – are outcomes of factors that often date back to the mid 19th century when railroads in the U.S. were expanding. Those factors determined which railroad had the resources to extend its tracks further, and thus whether a particular route ended up wholly-owned. While coal was, and continues to be, an important commodity for railroads, the locations where coal

such as economies of scope. For such analysis see Ivaldi and Mccullough (2010).

[4] A notable exception is some historical data on patent pools, for example, see Lampe and Moser (2010) and Lampe and Moser (2013) analyzing 19th century patent pools for sewing machine patents.

is mined have changed dramatically since the 19th century. In particular, while currently Wyoming produces around half of the coal in the U.S., in the 19th century the leader was Pennsylvania, with virtually no production in Wyoming. The current coal production in Pennsylvania is around one tenth of that in Wyoming. A potential weak point in this position is that endogenous mergers since that time have greatly changed the ownership patterns of once separate rail lines; however, we do not believe that Cournot-like inefficiencies related to coal transportation were among the important drivers of these mergers.

In addition to our main specification – a fixed effects pricing regression that was used in several existing studies both by academics and by regulators – we use other methods to check whether our results are robust. As one alternative, we use machine learning methods, as outlined by Belloni and Chernozhukov (2013), Belloni, Chernozhukov, and Hansen (2014), and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, et al. (2016). The machine learning approach provides a way to check the robustness of our results without relying on the exact functional specification used in our main analysis. As another alternative, we construct propensity scores for the treatment, split observations in smaller blocks based on propensity scores, run the same regression as in the fixed effects specification within each block, and then analyze the weighted average of the resulting coefficients. The method is described in Imbens and Rubin (2015), who recommend this method over other propensity score techniques (for example, propensity score matching) or a simple OLS regression. Both the machine learning analysis and regressions in the propensity score blocks alleviate potential concerns over the exact functional form in the main analysis and variable selection.

In addition to several methodological approaches producing the same answer, we also provide an identification robustness check by using a difference-in-differences approach to analyze prices following a merger that made two routes wholly-owned. To the extent that the premerger behavior was Cournot, a merger should decrease the price for traffic carried over the newly wholly-owned route. Also, given that this was a merger of two major railroads, but only two routes were affected, it is doubtful that the merger was undertaken specifically to change prices on these two routes. We also could not find any references to these routes in news stories around the time of the merger. Any, even in our view implausible, concerns with the identification strategy above should be alleviated by this analysis. While our estimate for this merger analysis is not as precise, our results are consistent with the main specification.

This theoretical analysis had been applied in many other settings. Vertical double marginalization – a monopolist manufacturer selling through a monopolist retailer – is one of the main examples, e.g., Spengler (1950). Vertical mergers and mergers of firms producing complementary products are generally viewed considerably more benign than horizontal (substitutes) mergers, in part because of this very idea being that such a merger might alleviate the double marginalization/complementary monopoly concern.[5] Complementary monopolies have also received plenty of attention from the

[5] This branch of the literature is also related to foreclosure and tying. See, for example, Posner (1979), Moresi and Salop (2013), and Tirole et al. (2015); see also Burton and Wilson (2006) on vertical exclusion in rail markets. For a discussion of mergers of firms producing complements, see Anderson, Loertscher, and Schneider (2010).

strategy literature, with a prominent example of Microsoft (Windows) and Intel in the 1990s and 2000s producing complementary inputs to a personal computer, see Casadesus-Masanell and Yoffie (2007) and Casadesus-Masanell, Nalebuff, and Yoffie (2012).

Similar discussions of whether inefficiency survives market forces appear in other settings as well. For example, in the vertical setting, a solution to this problem is well-known: a manufacturer could offer a two-part tariff, see Oi (1971). At least some of the empirical literature seems to agree that this is occurring in retail (Villas-Boas (2007)), although there are severe data limitations in that stream of research as well, due to the manufacturers' marginal cost not being directly observed. For more recent empirical work, see Crawford, Lee, Whinston, and Yurukoglu (2015) for multichannel television markets, Gayle (2013) for airline markets, and references therein.[6]

The same question appears in the literature on the tragedy of the anticommons, see, for example, Heller (1998), Heller and Eisenberg (1998), and Buchanan and Yoon (2000), with some of the work being specifically about patents. This literature also struggles with the Cournot-like issues of many owners of, to use an analogy from property law, various sticks of property rights for the same property, where someone needs to get everyone's approval to get anything done. Similarly, the literature also mentions the Coasian possibility of negotiating to an efficient outcome.

There are multiple empirical studies of patents that try to shed light on this issue, and on the effect of patents on innovation in general, for example, see Murray and Stern (2007), Cockburn and MacGarvie (2009), Gupta (2014), Cohen, Gurun, and Kominers (2015), Kiebzak, Rafert, and Tucker (2016), and Hegde and Luo (2017).[7] Between these studies, one can find estimates that would support either conclusion. However, as noted above, none of the estimates that we are aware of have pricing data to test the theory directly.

The intellectual property literature on royalty stacking influenced the ideas of fair and reasonable non-discriminatory (FRAND) license terms and the discussion on patent assertion entities (PAEs, also sometimes referred to as patent trolls) and patent pools. See Chiao, Lerner, and Tirole (2007), Layne-Farrar, Padilla, and Schmalensee (2007), and Lemley (2007). Theoretical literature and models continued to develop, again pointing to an empirical question of whether the resulting equilibrium is closer to the work of Cournot or to the work of Coase, see for example, Llanes and Trento (2012), Lemley and Shapiro (2013), Lerner and Tirole (2015), Rey and Salant (2012), and Spulber (2013).

2 Data

Our data – the Waybill Sample – come from the U.S. Surface Transportation Board (STB), the regulator of freight railroads in the U.S. There exists a version of the Sample for public use; however, that version has much information aggregated or otherwise masked for competitive reasons. Since

[6] In particular, see Brueckner (2003) and Bamberger, Carlton, and Neumann (2004) showing that the effects in the airline industry might be different than those that we are finding in the railroad industry.

[7] See also review articles, for example Boldrin and Levine (2013), Graham and Vishnubhakat (2013), and Khan and Sokoloff (2001).

this would not be as helpful for the purposes of our analysis, we went through a procedure outlined in 49 C.F.R. 1244.9(c)(1) to request the full sample for research use. The procedure is somewhat similar to a FOIA process that might be familiar to many researchers. The main difference is that the information request has to be published in the Federal Register, and any interested parties can comment on the request. In order to limit any competitive concerns or concerns regarding the impact of our findings on current railroad practices, we requested the data only up to and including 2003. As noted above, our identification robustness check is a merger; however there were no major mergers since 2003 (for that matter since 1999), thus we did not believe that the more current years would have been particularly helpful.

The sample includes a multitude of railroads. However, from 2001 to 2003 – the years that we use in our main analysis – there are four major railroads that comprise of about 80% of the overall volume. We use only these railroads in our main analysis. We use only the last three years because previous years had several mergers that could have affected prices in many ways. We discuss these mergers below and use one of them as a robustness check for our estimates.

The sample is weighted by the STB to ensure national representation. A datapoint is a shipment. For that shipment we observe railroad(s) providing the service, origination point, termination point, any junctions where the shipment changed railroads, total price charged, the distance of the route, the weight of the shipment, the commodity shipped, and well over a hundred other characteristics. The industry standard for measuring price is RPTM – revenue per ton mile. Upon computing RPTM we found considerable variation. In order to eliminate any potential confounding effects, we focus on the most homogeneous commodity that is frequently shipped by rail – coal (STCC code 1121290). Up until the explosion of intermodal freight (standard-size containers that can be stacked on ships, rail, or trucks), coal was the commodity contributing the most revenue to freight railroads. After we focus on coal, the outliers in the RPTM were not as far from the rest of the distribution as in the whole sample, however, we still observed occasional shipments with RPTM of more than 100 times the median. We eliminate the top 5% and the bottom 5% of RPTM data from our sample. Eliminating the top and the bottom 1% does not change our results qualitatively.

The variables in Table 1 are the same variables as we use in our analyses below. The variables are log of the distance of the route, log of total weight of this shipment, log of weight per loaded railcar, volume on that route (to account for possible economies of scale/scope), whether the shipper owns the rail car, HHI at the origin of the shipment, HHI at the destination (both HHIs at the county level), a binary variable indicating whether the railroad is a monopolist at the origin, same variable for the destination, the log of the shipment-specific variable cost, and whether the rate is masked in the public sample.[8] Note that the variable indicating whether the rate is masked (CalcRate) is, effectively, also the variable indicating that there was negotiation over the rate, and allowing us to control for any potential selection.

[8] As for almost any cost measurement, there is a debate whether the cost measured is actually a marginal cost, see Wilson and Wolak (2016) arguing that it is not. We simply use this as a noisy proxy for the actual cost, and do not take a stand on whether this is the proper cost to use for any regulatory reasons.

Table 1: Summary statistics.

	Mean	p25	Median	p75
lnRPTM	-4.002	-4.676	-4.069	-3.478
Treatment	0.0805	0	0	0
lnMiles	6.143	5.733	6.415	6.980
lnTons	9.156	9.200	9.456	9.569
lnTonsCar	4.702	4.635	4.723	4.777
lnVolTons	14.432	13.485	14.650	15.592
DOwn	0.637	0	1	1
HHI_{origin}	0.706	0.501	0.505	1
HHI_{term}	0.905	.927	1	1
DM_{origin}	0.315	0	0	1
DM_{term}	0.585	0	1	1
lnCosts	10.734	10.206	11.117	11.572
CalcRate	0.604	0	1	1

3 Main analysis

Our estimation strategy is as follows. Consider a shipper that wants to ship coal from A to C, a route that passes through B. We estimate whether the price (RPTM) paid from A to C is the same when the same railroad owns both AB and BC tracks as it is when one railroad owns AB and another owns BC. We estimate difference in prices by comparing otherwise similar routes, with the difference being whether the route is wholly-owned.

We use a pricing specification used by previous railroad-specific research to get our estimates. The specification, effectively a hedonic pricing regression, was used by Christensen Associates (2010) in a report prepared for the STB – in other words a report by a specialized consulting firm, with the industry's regulator as the customer. The report tweaks a previously existing specification from the academic literature, Mac Donald (1989), with the article looking at an unrelated deregulatory question. In addition to the exact specification used in the report for the STB, we also include shipment-specific costs that have a highly significant coefficient.[9] We do not use proximity of water ports that the report used – it is not statistically significant in our estimates and does not change the estimate of interest regardless of whether it is included.

We also include the treatment variable that we are interested in – whether the shipment is served by two railroads or by one. None of the shipments that we observe are served by more than two railroads.

A particular industry practice somewhat complicates our analysis. The railroads are permitted to rebill shipments that are served by two railroads. In other words, we often observe a shipment

[9] The cost variable inherently has a significant measurement error, since measuring marginal costs is typically a hard problem in a specific case, let alone across industry for all shipments. However, given that we are utilizing a hedonic pricing regression, are not interested in the coefficient on the cost variable in and of itself, and expect that costs matter for pricing, we felt that we should include this variable. If the measurement error is overwhelming, then we should get a not statistically significant coefficient. We assume that the measurement error in marginal costs is not correlated with our coefficient of interest.

served by one railroad with the flag indicating that it was rebilled: instead of observing route AC, showing that AB was served by railroad 1 and BC was served by railroad 2, we see AB served by railroad 1 with a rebilling flag. There are considerably more rebills in the data than there are routes with junctions. If we were to have the universe of shipments, as opposed to a sample, we could have connected the missing pieces since BC served by railroad 2 with a rebilling flag would have also been in our sample. However, this type of a match is not possible given that we only have a sample. Thus, our treatment variable is 1 if we observe *either* a junction or a rebill. It is 0 otherwise.[10]

Thus, our specification is

$$\begin{aligned} lnRPTM = {} & \beta_{interest} treatment + \\ & + \overline{\beta} \times [lnMiles + lnTons + lnTonsCar + lnVolTons + DOwn + HHI_{origin} + HHI_{term} \\ & + DM_{origin} + DM_{term} + lnCosts + CalcRate] + FE_{origin} + FE_{term} \\ & + FE_{railroadorigin} + FE_{railroadterm} + FE_{quarter}. \end{aligned} \tag{1}$$

In addition to the variables described above, we also include fixed effects for origin, termination point, railroad serving the origin, railroad serving the termination point (different if there is a junction), and quarter-year (for example, Q1 of 2001). We cluster our standard errors at the level of origin-termination-quarter-year.

We present the estimation results in Table 2. We estimate seven models in total. First, we estimate equation (1) on the whole sample without fixed effects. These results are presented in column (1). Next, we estimate the same model with fixed effects. The results are shown in column (2). Because our results can be affected by the extreme values of the dependent variable, we re-estimate models (1) and (2) on the sample that excludes top and bottom 1% of the PRTM distribution and on the sample that excludes top and bottom 5% of the PRTM distribution. These results are presented in columns (3)-(6), respectively. Finally, we eliminate top and bottom 5% of the RPTM distribution and observations with missing rebill variable[11] and re-estimate equation (1). These results are presented in column (7).

The results are very consistent across all models. The effect of treatment is economically small and statistically insignificant. The coefficients of other covariates are also pretty consistent across the models, except the weight per loaded railcar, HHI at origin and termination point, and indicators of monopoly at origin and termination points.

While the estimates of the treatment effect are consistent across the models, one might worry that the results might be affected by trimming the sample and excluding treated observations with high revenue per ton mile. Therefore, we present the number of excluded observations by

[10] This measurement issue was also noted by McCullough and Thompson (2013).

[11] The rebill variable is poorly recorded in years 2001 and 2002 with 40% of observations missing. Starting 2003 the reporting of this variable has significantly improved and there are no missing values of rebill in 2003. While, there are many observations missing in the earlier years, it seems that most of them migrated in 'no rebill' category in 2003 – the percentage of observations in 'no rebill' category increased from 60% to 85%. While the percentage of observations in 'rebill' category increased from 0.4 percent to 15 %.

treatment group in Table 3. The number of treated observations in top 1% and top 5% of the RPTM distribution is very small – 2 and 21 observations, respectively.

Table 2: Effect of either a junction or a rebill on price

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	lnRPTM	lnRPTM	lnRPTM	lnRPTM	lnRPTM	lnRPTM	lnRPTM	RPTM
	Whole	Whole	Trim at 1 %	Trim at 1 %	Trim at 5%	Trim at 5%	Trim at 5%	Trim at 5%
							exclude	exclude
							missing rebill	missing rebill
Treatment	-0.007	-0.013	0.001	-0.014	0.004	-0.006	-0.002	0.0002
	(0.023)	(0.008)	(0.023)	(0.008)	(0.024)	(0.008)	(0.009)	(0.0002)
lnMiles	-0.766***	-0.919***	-0.775***	-0.829***	-0.833***	-0.779***	-0.726***	-0.019***
	(0.023)	(0.024)	(0.024)	(0.021)	(0.028)	(0.022)	(0.023)	(0.001)
lnTons	-0.344***	-0.174***	-0.373***	-0.180***	-0.460***	-0.174***	-0.131***	-0.001**
	(0.028)	(0.026)	(0.028)	(0.018)	(0.030)	(0.016)	(0.017)	(0.0002)
lnTonscar	-0.432***	-0.303***	-0.290***	-0.063	-0.221**	-0.041	-0.126*	-0.004***
	(0.076)	(0.089)	(0.077)	(0.045)	(0.082)	(0.039)	(0.054)	(0.001)
lnVoltons	-0.072***	-0.038***	-0.077***	-0.031***	-0.069***	-0.026***	-0.025***	-0.001***
	(0.004)	(0.004)	(0.004)	(0.003)	(0.004)	(0.002)	(0.002)	(0.000)
DOwn	-0.081***	-0.020**	-0.082***	-0.025***	-0.059***	-0.028***	-0.045***	-0.002***
	(0.010)	(0.007)	(0.010)	(0.006)	(0.010)	(0.005)	(0.006)	(0.0001)
HHI_{origin}	0.354***	-0.179*	0.373***	-0.132	0.387***	-0.102	-0.194**	-0.007***
	(0.043)	(0.075)	(0.042)	(0.070)	(0.042)	(0.060)	(0.064)	(0.003)
HHI_{term}	0.063	0.042	0.036	0.049	0.076*	0.092	0.107	0.002
	(0.034)	(0.069)	(0.032)	(0.069)	(0.030)	(0.066)	(0.071)	(0.001)
DM_{origin}	-0.047*	0.005	-0.061**	0.003	-0.057**	0.007	0.022***	0.001***
	(0.020)	(0.006)	(0.019)	(0.006)	(0.019)	(0.006)	(0.006)	(0.0002)
DM_{term}	0.099***	-0.004	0.102***	-0.001	0.077***	0.002	-0.003	0.00002
	(0.014)	(0.007)	(0.013)	(0.006)	(0.011)	(0.006)	(0.007)	(0.0001)
lnCosts	0.370***	0.165***	0.406***	0.174***	0.500***	0.167***	0.122***	
	(0.033)	(0.030)	(0.033)	(0.021)	(0.035)	(0.019)	(0.020)	
Costs								-0.000**
								(0.000)
CalcRate	-0.024**	-0.047***	-0.031***	-0.047***	-0.025**	-0.046***	-0.051***	-0.001***
	(0.009)	(0.004)	(0.009)	(0.004)	(0.009)	(0.004)	(0.004)	(0.000)
Fixed Effects	–	✓	–	✓	–	✓	✓	✓
N	78,629	78,569	77,057	77,007	70,769	70,728	52,447	52,447
Adj. R2	0.86	0.96	0.86	0.97	0.83	0.96	0.97	0.96

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

All models with fixed effects include fixed effects for country of origin, destination county, railroad serving the origin, railroad serving the termination point, and quarter-year.

Standard errors are clustered by country of origin, destination county, and quarter-year.

As in any hedonic regression, the coefficients should not be interpreted as causal. In particular, the coefficients on HHI are not causal – HHI is endogenous, and we do not use a valid instrument

Table 3: PRTM Extreme Values by Treatment Group

	Top 1%	Bottom 1%	Top 5%	Bottom 5%
Controls	784	564	3,910	3,2345
Treated	2	222	21	686
All	786	786	3,931	3,931

to address this endogeneity problem. Nonetheless, the coefficients on our control variables are generally of the expected sign, statistically significant, and broadly similar to those obtained by both Mac Donald (1989) and Christensen Associates (2010). In particular, the economies associated with longer hauls (Miles), larger shipments (Tons), larger loads per car (Tonscar), and higher annual volumes on the route (Voltons) are all reflected in negative and statistically significant coefficients in almost all specifications. In terms of magnitudes, our estimated coefficients are closer to those of the previous studies for Miles and Voltons, not as close for Tons and Tonscar; however, as Christensen Associates (2010) point out, these latter two variables are not independent of each other, so that their separated coefficients must be interpreted with greater caution.

Like Christensen Associates (2010) but unlike Mac Donald (1989), we observe unstable and often counterintuitive signs on the coefficients for competition at origin (especially) and destination points. Christensen Associates conjecture that a positive coefficient on the number of railroads serving the origin may reflect competition in aspects of service quality that are unobservable in the waybill data.

4 Methodological robustness checks

4.1 Propensity score blocking

We follow Imbens and Rubin (2015) and estimate causal effect based on subclassification (blocking) on the estimated propensity score. First, we estimate the propensity score, or probability of being treated, as a function of the variables available in the Waybill sample. Next, we partition the sample into blocks based on the values of the estimated propensity score, so that within a block, the estimated propensity scores are approximately the same. Then, within each block, we estimate causal effect using the fixed effects regression outlined earlier. Finally, the average treatment effect for the whole sample is calculated as an average of the within-block estimated treatment effects weighted by block sizes.

Imbens and Rubin (2015) show that within blocks with the same estimated propensity score, the super-population distribution of covariates is identical in the treated and control groups. This property of the propensity score implies that splitting sample into blocks with approximately constant propensity score eliminates systematic biases associated with differences in observed covariates between treated and control groups, and thus leads to more precise estimates. However, blocking alone typically does not eliminate all biases that arise because of the differences in the covariates between control and treatment groups, because often even when data are split into smaller groups the estimated propensity score is not constant within blocks. Therefore, we run a regression to estimate the effect within each block to further reduce bias of estimates.

The key in the blocking approach is to construct comparable control and treatment groups within each block. Therefore, the first step is to refine the sample and eliminate outliers in the same way as in the previous section – we eliminate observations in the 5 bottom and 5 top percent of the RPTM distribution. Additionally, we identify counties with all originating/terminating shipments in either control or treatment group and eliminate shipments that originate/terminate in such counties. Ideally, we would use county dummies in our propensity score regression; however, there are over a thousand of such dummies, and doing this is not practical. Eliminating counties that perfectly predict treatment is effectively doing what a logit propensity score would do if we could run the estimation with all the county dummies.

Finally, we eliminate all shipments for which rebilling flag variable is missing. Initial sample includes 72,300 control group observations and 6,331 treatment group observations. We discard 61,588 observations in total, the vast majority of which is due to the perfect predictor counties mentioned above. The final sample includes 16,043 observations, among them 12,860 observations are in the control group (18% of the original control group) and 3,183 are in the treatment group (66% of the original treatment group).[12]

Next, we estimate the propensity score using the following logit specification:

$$\begin{aligned}Treatment = \overline{\gamma} \times [&lnMiles + lnTons + lnTonsCar + lnVolTons + DOwn + HHI_{origin} + HHI_{term} \\ &+ DM_{origin} + DM_{term} + lnCosts + CalcRate + Share\_treated_{origin} + Share\_treated_{term}] \\ &+ FE_{railroadorigin} + FE_{railroadterm} + FE_{quarter}.\end{aligned} \tag{2}$$

The control variables are as in equation (1); additionally, $Share\_treated_{origin}$ is a share of shipments served by more than one railroad originating in a county of shipment origin, and $Share\_treated_{term}$ is a share of shipments served by more than one railroad terminating in a county of shipment final destination (this variable serves as another proxy for having county-level dummies in the specification). The results of this estimation are shown in Table A2.

Next, we discard observations with the estimated propensity score too close to zero or one to eliminate units from either control or treatment group that do not a have good counterpart in treatment or control group, respectively. Specifically, we drop observations with propensity scores above 0.9375 and below 0.0269. The top threshold cuts off the top 0.25% of the untreated observations and about 24% of the treated. The bottom threshold cuts off about 4% of the treated observations and about 60% of the untreated. Clearly, the data above the top threshold is highly skewed towards being treated and the data below the bottom threshold is highly skewed towards

[12] As Imbens and Rubin (2015) argue, this approach sacrifices some external validity – the final estimates of the average treatment effect for the trimmed sample are less likely to be valid for the original sample. However, the advantage of this approach is the internal validity, i.e., the estimates of the treatment effect for the trimmed sample are more accurate than the estimates of the average treatment effect in the original sample.

being untreated. Table 4 displays the subsample sizes by treatment group and propensity score value.

Table 4: Sample Sizes for Trimming Based on Estimated Propensity Score

	$\hat{e}(X_i) < 0.0269$	$0.0269 < \hat{e}(X_i) < 0.9375$	$\hat{e}(X_i) > 0.9375$
Controls	7,716	5,112	32
Treated	164	3,027	992
All	7,880	8,139	1,024

Next, we split the sample into twenty blocks. Table A1 shows the details for these twenty blocks including the cut off values for the propensity score, the number of units by treatment status in each group, and the standardized differences in control variables and propensity scores for the whole trimmed sample and within each block.[13] The idea is to keep splitting the blocks until either the covariates look balanced or until there aren't enough treated or untreated observations relative to the number of controls in the regression that we run inside each block. Our regression specification has over a dozen of observables and even more fixed effects, thus we do not split blocks further if we have 30 or fewer treated or 30 or fewer untreated observations. Each individual block is much more balanced comparing to the whole sample – the normalized differences between covariates are much smaller within the blocks than in the whole sample. Finally, we estimate treatment effect within each block using model specified in equation (1). As we show in Table A1 in the Appendix, there is sufficient difference in the covariate distributions within the blocks and thus regression helps to further adjust for these differences.

We present results for the parameter estimates from the regressions for the twenty blocks in Table 5. The overall average treatment effect (ATE) is calculated in the following way:

$$ATE = \sum_j q_j \times \hat{\tau}(j), \tag{3}$$

where $j = 1, ..., 20$ corresponds to the block number; $q(j) = N(j)/N$, where $N(j)$ is the number of observations in block j and N is the total number of observations in the sample; finally, $\hat{\tau}(j)$ is the within-block least squares estimate of the treatment effect for block j.

The variance of the overall ATE is calculated in the similar manner:

$$V(ATE) = \sum_j q_j^2 \times \hat{V}(\hat{\tau}(j)), \tag{4}$$

where $\hat{V}(\hat{\tau}(j))$ is the estimated variance of the treatment effect within block j.

The results indicate that the ATE equals -0.05 with the standard deviation of 0.02, which

[13] The standardized difference between two samples for a variable is calculated using the following formula, e.g., for lnMiles:

$$z = \frac{\overline{lnMiles}_{treated} - \overline{lnMiles}_{control}}{\sqrt{s^2_{treated}/N_{treated} + s^2_{control}/N_{control}}}$$

confirms our main results that shipments served by several railroads are not higher priced than shipments served by one railroad. The magnitude of the coefficient is small in the absolute value, and is close to the coefficient that we obtain in the main specification.

Table 5: Independent Least Squares Regressions within Blocks

Block	N	Est.	S.E.
1	2,069	0.06	0.04
2	1,201	0.00	0.01
3	358	-0.09	0.03
4	151	0.05	0.04
5	455	0.01	0.02
6	90	0.01	0.01
7	255	0.00	0.02
8	251	-0.23	0.03
9	58	-0.14	0.07
10	118	0.00	0.00
11	78	0.40	0.05
12	78	0.10	0.08
13	97	-0.02	0.11
14	147	-0.30	0.06
15	120	-0.09	0.08
16	50	0.11	0.04
17	167	0.08	0.07
18	1,088	-0.06	0.06
19	838	-0.40	0.10
20	314	-0.01	0.02
ATE		-0.05*	0.02

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.2 Double Machine Learning

We refer readers who are not familiar with the standard machine learning techniques, such as neural nets, to the Appendix. Even though ML estimators cannot be used for causal inference directly, the estimation techniques that combine regression and machine learning methods are able to provide valid estimates of causal effects. Specifically, we implement double machine learning (DML) estimator developed by Chernozhukov et al. (2016). In combination with cross-fitting, the estimator is efficient and approximately unbiased and normal. The estimation proceeds as follows.

First, we model the outcome variable as the following partially linear model:

$$Y = D\theta_0 + g_0(Z) + U, \tag{5}$$

$$D = m_0(Z) + V, \tag{6}$$

where $E[U|Z, D] = 0$ and $E[V|Z] = 0$. Y is the outcome variable, D is the treatment variable, Z is a vector of covariates listed in equation (2), and U and V are disturbances. The first equation is the main equation that we would like to estimate with the parameter of interest θ_0. The second equation keeps track of confounding, or dependence of treatment variable on covariates. A set of control variables Z impacts outcome variable and treatment variable via the functions $g_0(Z)$ and

$m_0(Z)$ respectively.

DML estimator is obtained by partialing out the effect of Z from both Y and D and estimating the regression model implied by equations (5) - (6):

$$W = V\theta_0 + U, \tag{7}$$

where $V = D - m_0(Z)$ and $W = Y - l_0(Z)$, where $l_0(Z) = E[Y|Z] = m_0(Z)\theta_0 + g_0(Z)$. We estimate functions m_0 and l_0 using neural nets. We chose neural nets because this is one of the machine learning techniques that is most often used by machine learning researchers due to out-of-sample predictive success. In addition, when choosing which machine learning method to use, we were driven by our desire to deal flexibly with any potential nonlinearities and interactions. We believe that neural nets accomplish this objective better than, for example, post-lasso.

First, we split the data into two equal subsamples – the training sample and the test sample. Next, we obtain parameter estimates using neural nets and the training sample and construct estimates of $\hat{l}_0$ and $\hat{m}_0$ using obtained parameters and the test sample. Finally, we use these estimates to form $\hat{W} = Y - \hat{l}_0(Z)$ and $\hat{V} = D - \hat{m}_0(Z)$ and then obtain "double" ML estimator:

$$\hat{\theta}_0 = \Big(\frac{1}{n}\sum_{i=1}^{n}\hat{V}_i^2\Big)^{-1}\frac{1}{n}\sum_{i=1}^{n}\hat{V}_i\hat{W}_i. \tag{8}$$

Chernozhukov et al. (2016) prove that this estimator is root-n consistent and approximately Gaussian under a very mild set of conditions. As Chernozhukov et al. (2017) note, the specific sample partitioning has no impact on estimation results asymptotically but may be important in finite samples. In other words, when estimating using a finite sample, the value of the estimator depends on a specific split of the sample. Hence, to get asymptotically valid estimates, we repeat estimation procedure S times each time partitioning sample in halves. We then report estimates that incorporate information from the distribution of the estimates obtained from the different data partitions. As a result, we report mean estimate based on S obtained estimates of the parameter of interest:

$$\hat{\theta}_0^{Mean} = \frac{1}{S}\sum_{s=1}^{S}\hat{\theta}_0^s, \tag{9}$$

where $\hat{\theta}_0^s$ is a point estimate obtained in each of S estimations. Finally, we calculate the standard error of the $\hat{\theta}_0^{mean}$ that incorporates additional variation due to different data splits:

$$\hat{\sigma}^{Mean} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\Big(\hat{\sigma}_s^2 + (\hat{\theta}_0^s - \hat{\theta}_0^{Mean})^2\Big)}, \tag{10}$$

where $\hat{\sigma}_s = (EV^2)^{-1}EU^2V^2(EV^2)^{-1}$.

We present the results in Table 6. We trim the top and the bottom 5% of the RPTM distribution as before. We implement neural net estimator using a 2-fold cross-fitting (splitting sample in halves) with 10 hidden layers and a decay parameter of 0.01, which prevents over-fitting.[14] We repeat the main procedure 100 times repartitioning data in each replication. The result is again very similar to the result in the the main specification. While the results show a statistically significant negative coefficient, the coefficient is small in absolute terms, thus we do not attempt to explain the unexpected sign.

Table 6: Effect of either a junction or a rebill on price, double machine learning.

ATE	-0.06***
	(0.016)
N	70,728

Standard errors in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5 Identification robustness check: BNSF merger

On December 31st, 1996 two major railroads merged, ATSF and BN, creating the largest railroad in the U.S. – BNSF. Although the ATSF and BN were primarily parallel rather than interconnecting railroads – for years they joined the Southern Pacific and Union Pacific as the four major railroads serving the western US – there were some routes on which they connected to provide end-to-end service. Cournot analysis, or the presence of double marginalization, would suggest that if any route went from being served as a connecting route by ATSF and BN to being served by a single company BNSF, the price on that route should decrease. However, so would economies of scale or scope coming from the merger. Following previous academic work suggesting that it takes a couple of years for any economies of scale or scope to materialize, we use years 1995 – 1999 to test the hypothesis of whether price decreased on previously separately-owned routes following the merger.[15]

It turns out that only three coal routes were affected in this way (served by ATSF and BN as connecting carriers before the merger), all originating in Wyoming. As a part of getting the merger approved by the STB, BNSF agreed to give trackage rights to another railroad (UP) for one of the routes. That could, of course, lead to an immediate price drop due to competition on this route that is now served by both the combined ATSF and, through trackage rights, the UP; thus, we do not use this route. Therefore, we have two routes left.

We use a difference-in-differences specification mirroring our setup in the main analysis. Thus, instead of the treatment variable in the main analysis, we have a variable for affected routes, another variable for post-merger, and finally yet another variable for the interaction (difference-

[14] We tried different number of hidden layers, from 1 to 12, and the results are very similar to the results we get with 10 hidden layers. Adding hidden layers marginally improves mean standard error that measures model fit to the data and significantly increases time of the calculation. The estimate of average treatment effect practically stays the same.

[15] See Ivaldi and Mccullough (2010). See also Berndt, Friedlaender, Chiang, and Vellturo (1993).

in-differences) term. The coefficient of interest is the interaction term. Thus, our specification is

$$\begin{aligned} lnRPTM = {} & \beta_{interest} Post \times AffectedRoutes + \beta_{post} Post + \beta_{affected} AffectedRoutes \\ & + \overline{\beta} \times [lnMiles + lnTons + lnTonsCar + lnVolTons + DOwn + HHI_{origin} + HHI_{term} \\ & + DM_{origin} + DM_{term} + lnCosts + CalcRate] + FE_{origin} + FE_{term} + FE_{railroadorigin} \\ & + FE_{railroadterm} + FE_{quarter}. \end{aligned} \tag{11}$$

The variable names and clustering are the same as in the main analysis. In particular, we use the same controls and the same fixed effects.[16]

We are also not sure when exactly the price change would have occurred post-merger if it indeed occurred. In other words, it might have taken some time for BNSF to synchronize the pricing systems. Thus, we run specification with the first quarter, first two quarters, and first year post-merger thrown out. Our coefficient of interest does not change significantly. The results are in Table 7 and, while noisy, are consistent with the results that we presented earlier.

As with any difference-in-differences analysis, a parallel trend graph helps with convincing ourselves that we have the correct identification strategy. We present the graph below, with the graph not corrected for any controls. We demean the RPTM for affected routes to preserve pricing anonymity. The price drop in the affected routes after month ‘460’ is consistent with the previous empirical literature that suggested that it takes close to two years for economies of scale to materialize for railroad mergers.

6 Conclusion

We found that, in the U.S. freight rail, prices for shipping coal are not consistent with a Cournot-like complementary monopoly outcome. Instead, we find evidence consistent with an equilibrium, where complementary monopolists on routes AB and BC do not charge more than a single monopolist would charge if she were to own the whole route AC.

How are the railroad companies able to accomplish this? In discussions with industry experts, we have learned that coal shipment contracts that involve two interconnecting railroads often include discussions and negotiations between the railroads concerning both the joint rate and the divisions of the rate, and that these discussions may be motivated/incentivized by coal customers, such as power plants soliciting joint rate bids for coal supplies. In such circumstances it seems not at all surprising that the two railroads seeking to win a joint bid can avoid the double marginalization characteristic of independent price setting of complements.[17]

[16] Note that the term $\beta_{post} Post$ is unnecessary given the quarter-year fixed effects. We still have it in the regression simply to make the point that we are using the standard difference-in-differences setup. The downside is the large magnitude and no statistical significance on coefficient β_{post}.

[17] This anecdotal evidence suggests that the negotiation between monopolists and producers, along the lines of

Table 7: Effect of merger on price of affected routes

	(1)	(2)	(3)	(4)
	lnRPTM	lnRPTM	lnRPTM	lnRPTM
		(without 1997Q1)	(without 1997Q1 & Q2)	(without 1997)
Post ×	0.00355	0.0125	0.0214	0.00978
Affected Routes	(0.0359)	(0.0397)	(0.0399)	(0.0426)
Post	-7605.9	-1972.3	1409.2	-242.3
	(5118458.0)	(2691777.1)	(3536224.9)	(2711872.8)
Affected Routes	-0.239*	-0.257*	-0.568***	-0.595***
	(0.114)	(0.118)	(0.0664)	(0.0695)
lnMiles	-0.570***	-0.571***	-0.576***	-0.586***
	(0.0129)	(0.0131)	(0.0134)	(0.0142)
lnTons	0.00953	0.00862	0.00542	-0.00684
	(0.00978)	(0.0100)	(0.0104)	(0.0108)
lnTonsCar	-0.169***	-0.166***	-0.162***	-0.150***
	(0.0194)	(0.0199)	(0.0205)	(0.0222)
lnVolTons	-0.0203***	-0.0197***	-0.0191***	-0.0189***
	(0.00204)	(0.00209)	(0.00210)	(0.00217)
DOwn	-0.0856***	-0.0858***	-0.0851***	-0.0807***
	(0.00419)	(0.00429)	(0.00441)	(0.00460)
HHI$_{origin}$	0.0348	0.0334	0.0314	0.0265
	(0.0213)	(0.0224)	(0.0228)	(0.0246)
HHI$_{term}$	-0.0461*	-0.0487*	-0.0567*	-0.0456
	(0.0191)	(0.0206)	(0.0221)	(0.0267)
DM$_{origin}$	0.00524	0.00339	0.00107	-0.00943
	(0.00555)	(0.00565)	(0.00563)	(0.00544)
DM$_{term}$	0.000639	0.00140	0.00326	0.00430
	(0.00504)	(0.00520)	(0.00534)	(0.00596)
lnCosts	-0.0444***	-0.0429***	-0.0389**	-0.0233
	(0.0119)	(0.0122)	(0.0125)	(0.0131)
CalcRate	-0.183***	-0.180***	-0.174***	-0.163***
	(0.00566)	(0.00577)	(0.00584)	(0.00595)
N	191,565	183,418	174,784	157,296

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

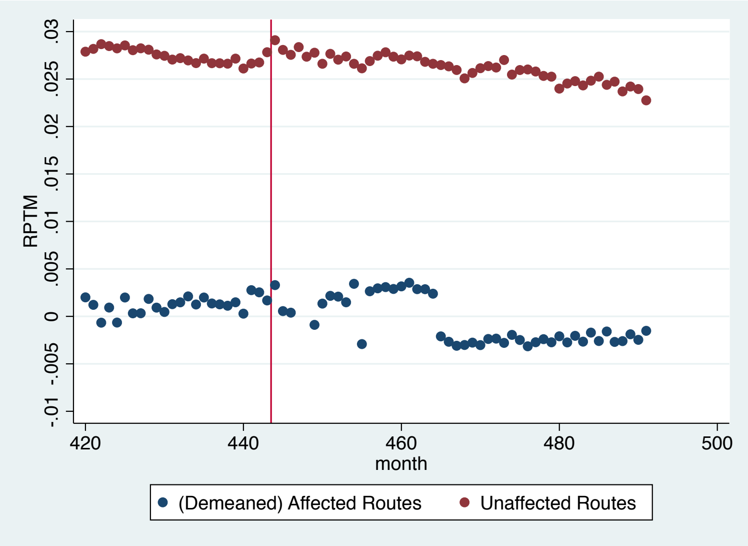

Figure 1: Pricing trends for affected and unaffected routes, with the merger occurring on December 31st, 1996.

Our results are directly applicable to the U.S. freight railroad market: for example, we show that an end-to-end railroad merger would not be needed to fix a Cournot or double-marginalization-like pricing issue in this market. However, one should be cautious when extrapolating from our results to other settings.

In particular, while railroads and products requiring multiple patents share the potential for a Cournot-like outcome, there are many important differences. First, products like cell phones oftentimes require hundreds of patents, instead of any coal company being able to ship using only two railroads in our setting. It is possible that a much higher number of complementary monopolists results in a Cournot-like outcome. Second, in the intellectual property realm, at least some of the often-discussed issues are around unscrupulous entities that use 'deceptive sales claims and phony legal threats' in order to attempt to collect royalties on invalid patents, effectively threatening a strike suit.[18] This is, of course, not possible for railroads to do. Third, while another concern in the intellectual property realm is that occasionally manufacturers do not even realize that their product impinges on patents, coal companies know for sure that they will have to use railroads to transport coal before they open a coal mine. On the other hand, fourth, while many manufacturers are also patent holders, so that oftentimes patents might be used defensively, freight railroads typically do not mine coal themselves.

We hope that our study inspires further work in this area, and in particular more direct analyses of whether the Cournot hypothesis holds in particular markets. Ideally, economists, legal scholars, and other interested researchers will analyze similar data from the industries and countries where

Spulber (2017), might be a better fit for this market than direct contact between monopolists as in Schumpeter (1928).

[18] See, for example, the FTC's MPHJ settlement, https://www.ftc.gov/news-events/press-releases/2014/11/ftc-settlement-bars-patent-assertion-entity-using-deceptive.

such data is available. Then, upon having multiple such studies, we could make educated hypotheses about more industries and, in particular, the degree to which, if any, the four differences above change our conclusions.

References

Anderson, S. P., S. Loertscher, and Y. Schneider (2010). The ABC of complementary products mergers. *Economics Letters 106*(3), 212–215.

Association of American Railroads (2016). Railroads and coal.

Bamberger, G. E., D. W. Carlton, and L. R. Neumann (2004). An empirical investigation of the competitive effects of domestic airline alliances. *The Journal of Law and Economics 47*(1), 195–222.

Belloni, A. and V. Chernozhukov (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli 19*(2), 521–547.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Berndt, E. R., A. F. Friedlaender, J. S.-E. W. Chiang, and C. A. Vellturo (1993). Cost effects of mergers and deregulation in the us rail industry. In *Productivity Issues in Services at the Micro Level*, pp. 123–140. Springer.

Boldrin, M. and D. K. Levine (2013). The case against patents. *The Journal of Economic Perspectives 27*(1), 3–22.

Brueckner, J. K. (2003). International airfares in the age of alliances: The effects of codesharing and antitrust immunity. *The Review of Economics and Statistics 85*(1), 105–118.

Buchanan, J. M. and Y. J. Yoon (2000). Symmetric tragedies: Commons and anticommons. *Journal of Law and Economics 43*, 1.

Burton, M. and W. W. Wilson (2006). Network pricing: Service differentials, scale economies, and vertical exclusion in railroad markets. *Journal of Transport Economics and Policy (JTEP) 40*(2), 255–277.

Calsoyas, C. (1950). The mathematical theory of monopoly in 1839: Charles Ellet, Jr. *Journal of Political Economy 58*(2), 162–170.

Casadesus-Masanell, R., B. Nalebuff, and D. Yoffie (2012). Competing complements.

Casadesus-Masanell, R. and D. B. Yoffie (2007). Wintel: Cooperation and conflict. *Management Science 53*(4), 584–598.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, et al. (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/debiased/neyman machine learning of treatment effects. *arXiv preprint arXiv:1701.08687*.

Chiao, B., J. Lerner, and J. Tirole (2007). The rules of standard-setting organizations: An empirical analysis. *The RAND Journal of Economics 38*(4), 905–930.

Christensen Associates (2010). An update to the study of competition in the us freight railroad industry: Final report. Christensen Associates, Inc. A report prepared for The Surface Transportation Board (STB).

Coase, R. H. (1960). The problem of social cost. In *Classic Papers in Natural Resource Economics*, pp. 87–137. Springer.

Cockburn, I. M. and M. J. MacGarvie (2009). Patents, thickets and the financing of early-stage firms: Evidence from the software industry. *Journal of Economics & Management Strategy 18*(3), 729–773.

Cohen, L., U. G. Gurun, and S. D. Kominers (2015). Patent trolls: Evidence from targeted firms. *Harvard Business School Finance Working Paper* (15-002).

Cournot, A. A. (1897). *Researches into the Mathematical Principles of the Theory of Wealth*. Macmillan Co.

Crawford, G. S., R. S. Lee, M. D. Whinston, and A. Yurukoglu (2015). The welfare effects of vertical integration in multichannel television markets. *National Bureau of Economic Research Working Paper*.

Elhauge, E. (2008). Do patent holdup and royalty stacking lead to systematically excessive royalties? *Journal of Competition Law and Economics 4*(3), 535–570.

Gayle, P. G. (2013). On the efficiency of codeshare contracts between airlines: Is double marginalization eliminated? *American Economic Journal: Microeconomics 5*(4), 244–273.

Geradin, D., A. Layne-Farrar, and J. Padilla (2007). Royalty stacking in high tech industries: Separating myth from reality.

Graham, S. and S. Vishnubhakat (2013). Of smart phone wars and software patents. *The Journal of Economic Perspectives 27*(1), 67–85.

Gupta, K. (2014). Technology standards and competition in the mobile wireless industry. *Geo. Mason L. Rev. 22*, 865.

Hagiu, A. and D. B. Yoffie (2013). The new patent intermediaries: Platforms, defensive aggregators, and super-aggregators. *The Journal of Economic Perspectives 27*(1), 45–65.

Hastie, T., R. Tibshirani, and J. Friedman (2009). The elements of statistical learning.

Hegde, D. and H. Luo (2017). Patent publication and the market for ideas. *Management Science*.

Heller, M. A. (1998). The tragedy of the anticommons: Property in the transition from marx to markets. *Harvard law review*, 621–688.

Heller, M. A. and R. S. Eisenberg (1998). Can patents deter innovation? The anticommons in biomedical research. *Science 280*(5364), 698–701.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

Ivaldi, M. and G. Mccullough (2010). Welfare tradeoffs in us rail mergers. *IDEI Working Paper*.

Khan, B. Z. and K. L. Sokoloff (2001). History lessons: The early development of intellectual property institutions in the united states. *The Journal of Economic Perspectives 15*(3), 233–246.

Kiebzak, S., G. Rafert, and C. E. Tucker (2016). The effect of patent litigation and patent assertion entities on entrepreneurial activity. *Research Policy 45*(1), 218–231.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *The American economic review 105*(5), 491–495.

Lampe, R. and P. Moser (2010). Do patent pools encourage innovation? Evidence from the nineteenth-century sewing machine industry. *The Journal of Economic History 70*(04), 898–920.

Lampe, R. and P. Moser (2013). Patent pools and innovation in substitute technologies—evidence from the 19th-century sewing machine industry. *The RAND Journal of Economics 44*(4), 757–778.

Layne-Farrar, A., A. J. Padilla, and R. Schmalensee (2007). Pricing patents for licensing in standard-setting organizations: Making sense of frand commitments. *Antitrust Law Journal 74*(3), 671–706.

Lemley, M. A. (2007). Are universities patent trolls. *Fordham Intell. Prop. Media & Ent. LJ 18*, 611.

Lemley, M. A. and C. Shapiro (2006). Patent holdup and royalty stacking. *Tex. L. Rev. 85*, 1991.

Lemley, M. A. and C. Shapiro (2013). A simple approach to setting reasonable royalties for standard-essential patents. *Berkeley Tech. LJ 28*, 1135.

Lerner, J. and J. Tirole (2015). Standard-essential patents. *Journal of Political Economy 123*(3).

Llanes, G. and S. Trento (2012). Patent policy, patent pools, and the accumulation of claims in sequential innovation. *Economic Theory 50*(3), 703–725.

Mac Donald, J. M. (1989). Railroad deregulation, innovation, and competition: Effects of the staggers act on grain transportation. *Journal of Law and Economics*, 63–95.

McCullough, G. J. and L. S. Thompson (2013). A further look at the staggers rail act: Mining the available data. *Research in Transportation Business & Management 6*, 3–10.

Moresi, S. and S. C. Salop (2013). vGUPPI: Scoring unilateral pricing incentives in vertical mergers. *Antitrust Law Journal 79*(1), 185.

Murray, F. and S. Stern (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization 63*(4), 648–687.

Oi, W. Y. (1971). A Disneyland dilemma: Two-part tariffs for a Mickey Mouse monopoly. *The Quarterly Journal of Economics*, 77–96.

Posner, R. A. (1979). The chicago school of antitrust analysis. *University of Pennsylvania Law Review 127*(4), 925–948.

Qi, M. (1999). Nonlinear predictability of stock returns using financial and economic variables. *Journal of Business & Economic Statistics 17*(4), 419–429.

Rey, P. and D. Salant (2012). Abuse of dominance and licensing of intellectual property. *International Journal of Industrial Organization 30*(6), 518–527.

Schumpeter, J. (1928). The instability of capitalism. *The economic journal 38*(151), 361–386.

Shapiro, C. (2001). Navigating the patent thicket: Cross licenses, patent pools, and standard setting. In *Innovation Policy and the Economy, Volume 1*, pp. 119–150. MIT press.

Sidak, J. G. (2008). Holdup, royalty stacking, and the presumption of injunctive relief for patent infringement: A reply to lemley and shapiro. *Minnesota Law Review 92*(3), 714–748.

Sonnenschein, H. (1968). The dual of duopoly is complementary monopoly: Or, two of cournot's theories are one. *The Journal of Political Economy*, 316–318.

Spengler, J. J. (1950). Vertical integration and antitrust policy. *The Journal of Political Economy*, 347–352.

Spulber, D. F. (2013). How do competitive pressures affect incentives to innovate when there is a market for inventions? *Journal of Political Economy 121*(6), 1007–1054.

Spulber, D. F. (2016). Patent licensing and bargaining with innovative complements and substitutes. *Research in Economics*.

Spulber, D. F. (2017). Complementary monopolies and bargaining. *The Journal of Law and Economics 60*(1), 29–74.

Swanson, N. R. and H. White (1997). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics 79*(4), 540–550.

Tirole, J. et al. (2015). Market failures and public policy. *American Economic Review 105*(6), 1665–82.

Villas-Boas, S. B. (2007). Vertical relationships between manufacturers and retailers: Inference with limited data. *The Review of Economic Studies 74*(2), 625–652.

White, H. (1988). Economic prediction using neural networks: The case of IBM daily stock returns.

Wilson, W. W. and F. A. Wolak (2016). Freight rail costing and regulation: The uniform rail costing system. *Review of Industrial Organization 49*(2), 229–261.

A Fitting of Neural Networks

Background on machine learning

Readers familiar with the standard machine learning techniques and intuition can safely skip this section.

Supervised machine learning is used to build a prediction model that relates a set of inputs X and an outcome variable Y. Machine learning (ML) in this case is called *supervised* because the outcome variable guides the learning process. ML models are designed to optimize prediction, and therefore are concerned with overfitting and are not explicitly concerned with unbiased estimates. Consequently, ML is tailored for applications when there are many attributes of a unit relative to the number of observations and when one wants to allow flexible functional form between the inputs and an output, e.g., when non-linearity might be hard to capture using conventional reduced form models.

While ML techniques excel at prediction, they are not necessarily great for *causal* inference. The focus of ML methods, prediction and improved prediction, is sometimes achieved by using biased estimates (for example, by placing zero coefficients on some covariates to simplify the model). The intuition is explained in Kleinberg et al. (2015) and has to do with *variance-bias tradeoff.* Suppose a training dataset T of n data points (x_i, y_i) is used to pick a function f to predict y using x. Now, consider a mean squared error at a new point x_0, $MSE(x_0)$:

$$\begin{aligned} MSE(x_0) &= E_T[f(x_0) - \hat{y_0}]^2 \\ &= E_T[\hat{y}_0 - E_T(\hat{y}_0)]^2 + [E_T(\hat{y}_0) - f(x_0)]^2 \\ &= Var_T(\hat{y}_0) + Bias^2(\hat{y}_0) \end{aligned} \tag{12}$$

This *bias-variance decomposition* of the MSE shows that there is a tradeoff between variance and bias of the estimate. More generally, as the model complexity increases, the variance tends to increase and the squared bias tends to decrease.[19]

Ordinary least squares (OLS) estimate is unbiased under some assumptions. However, unbiasedness comes at the cost of higher variance. Gauss-Markov theorem states that the least squares estimate has the smallest variance among all unbiased linear estimates, but there exist biased estimates with smaller variance. For example, setting to zero some of the least squares estimates might result in a small bias and a significant reduction in variance and thus a better prediction. ML methods optimize this balance between bias and variance and therefore while potentially outperforming OLS in prediction, ML estimates are likely biased.

ML methods select a subset of predictors to produce a model that is interpretable and has possibly lower prediction error than the full model. In particular, ML techniques minimize:

$$\hat{f}_{ML} = arg \min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda R(f), \tag{13}$$

[19] See Hastie et al. (2009) pp. 9-42 for more detail.

where $R(f)$ is a *regularizer* that penalizes model complexity. $\lambda \geq 0$ is a complexity parameter that controls the amount of *shrinkage* (i.e., how many coefficients are shrunk toward zero). The larger the λ is, the more parsimonious is the model. Regularizer function may take various forms, for example, for linear models $R(f_\beta) = \|\beta\|^d$, where $d = 1$ corresponds to the lasso estimator and $d = 2$ corresponds to the ridge and neural networks estimators (in neural networks λ is known as *weight decay*).[20]

Another important technique used in ML to ensure the quality of prediction is *cross-validation.* The data sample is split into training and test sample. First, the model is fitted using the training sample. Then the performance of the obtained model is tested on the test sample. This procedure allows to avoid overfitting and derives the optimal level of model complexity.

The ML method that we use in this estimation is *neural net.* Neural nets have been previously used in economics and finance literature (e.g., White (1988), Swanson and White (1997), Qi (1999)). Neural nets consist of a number of simple neuron-like processing units, organized in layers. Every unit in a layer connected to all the units in a previous layer. These connections are not equal: each connection may have a different strength or *weight.* Data enters at the inputs and passes through the network, layer by layer, until it arrives at the output. Layers between the inputs and an output are called *hidden layers* as they are not directly observed (latent). In the network that we use for our analysis there is no feedback between layers, and thus it is called *feed-forward* network. Figure 1 shows a network with several inputs, one output, and one hidden layer.

Neural nets can be seen as a two-stage non-linear regression. The hidden units Z_m are created from linear combinations of the inputs, and output variable Y_k is modeled as a function of linear combinations of Z_m:

$$\begin{aligned} Z_m &= \sigma(\alpha_{0m} + \alpha_m' X), \quad m = 1, ..., M, \\ f(X) &= \beta_0 + \beta' Z, \end{aligned} \tag{14}$$

The activation function $\sigma(v) = \frac{1}{1+e^{-v}}$ is known as *sigmoid.* The unknown parameters in neural network are called *weights*: $\theta = \{\{\alpha_{0m}, \alpha_m : m = 1, 2, ..., M\}, \{\beta_0, \beta\}\}$. Weights are found from the training data by fitting the model.

We use the sum-of-squared errors as a measure of fit:

$$R(\theta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda J(\theta), \tag{15}$$

where $J(\theta)$ is a *weight decay* and

$$J(\theta) = \sum_m \beta_m^2 + \sum_{ml} \alpha_{ml}^2 \tag{16}$$

The decay parameter penalizes large weights in the neural network and prevents overfitting. The generic approach is to minimize $R(\theta) + \lambda J(\theta)$ by gradient descent described directly below.

[20] $\|\beta\|^1 = \sum_{i=1}^k |\beta_k|$, $\|\beta\|^2 = \sum_{j=1}^k \beta_k^2$, where k is the number of controls in a subset chosen by the ML algorithm.

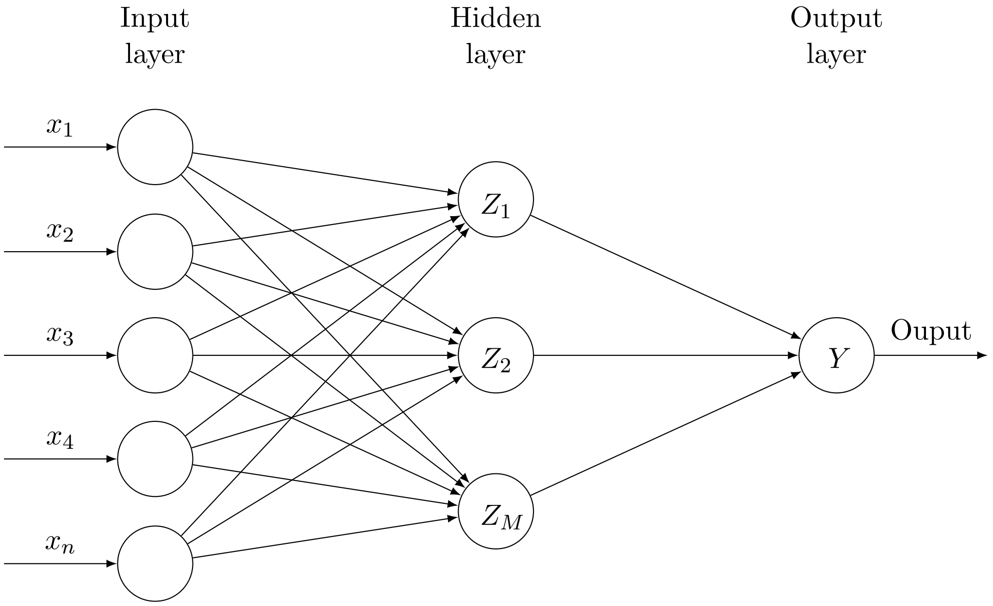

Figure 2: Neural Net with One Hidden Layer

Gradient descent algorithm

To find the set of weights θ that yield the best prediction in the neural network we minimize the sum of sum-of-squared errors $R(\theta)$ and *weight decay* $J(\theta)$[21]

$$\begin{aligned} R(\theta) + \lambda J(\theta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda J(\theta) = \\ \sum_{i=1}^{N} (y_i - f(x_i))^2 + \lambda(\sum_{m=1}^{M} \beta_m^2 + \sum_{m=1}^{M} \sum_{l=1}^{N} \alpha_{ml}^2). \end{aligned} \tag{17}$$

Conventionally this function is minimized by gradient descent. Given a function defined by a set of parameters, gradient descent starts with an initial set of parameter values and iteratively moves toward a set of parameters that minimize the function. This iterative minimization is achieved by taking steps in the negative direction of the function gradient. In this case, the gradient is easily derived using the chain rule:

$$\begin{aligned} \frac{\partial R_i}{\partial \beta_m} &= -2(y_i - f(x_i)) z_{mi} + 2\lambda \beta_m, \\ \frac{\partial R_i}{\partial \alpha_{ml}} &= -2(y_i - f(x_i)) \beta_m \sigma'(\alpha_{mi} x_i) x_{il} + 2\lambda \alpha_{ml}. \end{aligned} \tag{18}$$

Given these derivatives, the gradient descent update at the $(r+1)$st iteration has the form

[21] This algorithm is thoroughly described in Hastie et al. (2009), pp. 395-396.

$$
\begin{aligned}
\beta_m^{(r+1)} &= \beta_m^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \beta_m^{(r)}}, \\
\alpha_{ml}^{(r+1)} &= \alpha_{ml}^{(r)} - \gamma_r \sum_{i=1}^{N} \frac{\partial R_i}{\partial \alpha_{ml}^{(r)}},
\end{aligned}
\tag{19}
$$

where γ is called a *learning rate.*

Now, equation (18) can be rewritten as

$$
\begin{aligned}
\frac{\partial R_i}{\partial \beta_m} &= \delta_i z_{mi}, \\
\frac{\partial R_i}{\partial \alpha_{ml}} &= s_{mi} x_{il},
\end{aligned}
\tag{20}
$$

The quantities δ_i and s_{mi} are called "errors" from the current model at the output and hidden layers, respectively. From equations (18) and (20)

$$
s_{mi} = \sigma'(\alpha_m^T x_i)\beta_m \delta_i, \tag{21}
$$

which is known as a *back-propagation equations.* Then the gradient in (19) is updated using a two-pass algorithm. First, in the *forward pass*, the current weights are fixed and the predicted values $\hat{f}(x_i)$ are computed using equation (14). Next in the *backward pass*, the errors δ_i are computed, and then back-propagated via (21) to obtain the errors s_{mi}. Finally, δ_i and s_{mi} are used to update the gradient in (19).

There are certain guidelines that are recommended to successfully use neural networks. First, the starting values are chosen to be random values near zero. Second, it is recommended to use weight decay to avoid overfitting. It might be useful to scale inputs to have zero mean and standard deviation one – it ensures that inputs are treated equally in the regularization process and gives a higher quality prediction. Finally, it is better to have many hidden layers than too few to allow for model flexibility. Usually the number of hidden layers varies from 5 to 100. Hastie, Tibshirani, and Friedman (2009) provide a discussion of these guidelines. Finally, we use the *R* package *nnet* to train our neural network.

B Additional Tables

Table A1: Normalized Differences in the Covariates after Subclassification for Trimmed Sample

		Block									
	Whole Sample	1	2	3	4	5	6	7	8	9	10
lnMiles	11.06	3.17	1.83	3.21	3.32	0.07	-1.71	-5.35	2.88	2.40	3.64
lnTons	8.21	-3.37	-3.93	-1.87	-0.62	-1.79	0.33	-1.16	-3.82	0.51	-1.24
lnVolTons	14.25	-10.80	-6.07	-4.55	0.76	-2.87	-2.54	-2.40	-4.72	0.18	-0.92
Down	9.04	-5.69	1.87	-2.35	1.29	-3.41	0.93	2.23	-0.29	0.42	-2.00
lnCosts	10.81	2.77	-0.12	2.16	2.43	-0.45	-1.29	-5.60	0.94	1.72	2.80
lnTonsCar	10.40	-0.82	-1.21	-1.20	-1.27	-3.25	-0.38	4.04	-2.23	-1.95	-0.20
HHI$_{origin}$	-7.20	9.32	4.90	1.79	-1.47	0.98	2.47	2.83	0.44	-0.04	2.96
HHI$_{term}$	-9.70	2.39	-2.28	0.42	1.79	-6.31	2.75	2.65	-1.72	0.23	4.39
DM$_{origin}$	-2.61	11.37	4.18	1.97	-1.16	2.10	3.48	2.92	-1.16	0.30	1.25
DM$_{term}$	-6.71	1.92	-2.01	-4.66	-2.69	-0.68	3.50	1.72	-1.19	1.13	2.84
CalcRate	-3.67	2.07	3.11	3.72	4.10	-0.24	1.51	1.83	-1.14	0.21	-0.59
P-Score	100.79	0.18	4.39	1.22	-0.04	-4.24	2.16	-3.36	4.16	1.52	-1.58
Min P-Score	0.027	0.027	0.069	0.129	0.152	0.164	0.232	0.262	0.365	0.419	0.455
Max P-Score	0.938	0.069	0.129	0.152	0.164	0.232	0.262	0.365	0.419	0.455	0.510
# Controls	12,860	2,026	1,174	324	128	436	71	216	210	34	96
# Treated	4,183	45	43	42	31	31	31	47	47	32	31

	Block									
	11	12	13	14	15	16	17	18	19	20
lnMiles	6.90	0.33	2.86	5.95	-0.71	-0.83	-3.92	0.09	-0.68	-0.21
lnTons	-0.49	0.91	1.59	2.26	2.86	1.36	3.15	2.83	2.94	9.17
lnVolTons	1.06	0.61	2.45	4.25	1.70	0.16	1.30	3.45	3.27	1.32
Down	-2.08	-0.77	-1.72	-2.62	1.27	2.57	6.42	3.16	2.11	0.50
lnCosts	4.48	0.25	2.41	6.54	-0.03	-0.06	-2.61	0.35	-0.22	1.41
lnTonsCar	-1.15	1.46	1.35	1.68	4.30	1.72	0.28	2.05	1.94	0.13
DM$_{origin}$	0.27	1.75	0.20	-1.63	-1.75	-1.00	-3.96	-2.77	-3.23	-1.17
HHI$_{term}$	-2.04	-0.28	3.78	5.35	6.31	2.03	1.59	-0.88	0.16	4.71
DM$_{origin}$	-2.65	0.85	0.36	-1.62	-1.74	-0.70	-3.02	-2.46	-3.23	-1.22
DM$_{term}$	-2.46	-0.06	3.90	4.10	6.21	2.00	0.74	-1.03	-0.45	0.16
CalcRate	-2.94	-0.79	-2.03	-1.47	1.42	1.57	2.85	0.60	0.33	-0.72
P-Score	1.02	1.16	0.17	-0.22	-1.51	1.13	0.89	1.04	0.44	-2.15
Min P-Score	0.510	0.547	0.565	0.592	0.612	0.641	0.672	0.758	0.865	0.922
Max P-Score	0.547	0.565	0.592	0.612	0.641	0.672	0.758	0.865	0.922	0.938
# Controls	54	52	42	27	31	31	47	52	30	31
# Treated	30	33	63	126	95	30	126	1,044	813	287

Table A2: Propensity score estimation.

	(1)		
	Treatment		
lnMiles	-2.594***	DM_{origin}	1.297***
	(0.319)		(0.166)
lnTons	-3.061***	DM_{term}	-0.230*
	(0.334)		(0.094)
lnTonscar	3.448***	lnCosts	3.649***
	(0.640)		(0.421)
lnVolTons	-0.0646	CalcRate	-0.00227
	(0.038)		(0.083)
DOwn	0.616***	$Share\_treated_{origin}$	14.18***
	(0.112)		(0.598)
HHI_{origin}	-3.346***	$Share\_treated_{term}$	12.48***
	(0.350)		(0.291)
HHI_{term}	-0.790***	Const	-16.75***
	(0.234)		(2.731)
N	17,043		
Pseudo R^2	0.64		

Standard errors in parentheses. The model also includes originating railroad fixed effects, terminating railroad fixed effects, and quarter fixed effects.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$