Bill of Lading Data in International Trade Research with an Application to the COVID-19 Pandemic[*][†]

Aaron Flaaen Flora Haberkorn Logan Lewis Anderson Monken
Justin Pierce Rosemary Rhodes Madeleine Yi

Federal Reserve Board

September 2021

Abstract

We evaluate high-frequency bill of lading data for its suitability in international trade research. These data offer many advantages over both other publicly accessible official trade data and confidential datasets, but they also have clear drawbacks. We provide a comprehensive overview for potential researchers to understand these strengths and weaknesses as these data become more widely available. Drawing on the strengths of the data, we analyze three aspects of trade during the COVID-19 pandemic. First, we show how the high-frequency data capture features of the within-month collapse of trade between the United States and India that are not observable in official monthly data. Second, we demonstrate how U.S. buyers shifted their purchases across suppliers over time during the recovery. And third, we show how the data can be used to measure vessel delivery bottlenecks in near real time.

JEL classifications: F14, F17, C81

Keywords: Bill of lading, COVID-19 trade, Firm-level trade

[*]Corresponding author: aaron.b.flaaen@frb.gov. The views expressed here should not be interpreted as reflecting the views the Federal Reserve Board of Governors or any other person associated with the Federal Reserve System.

[†]Link to GitHub repository: `https://github.com/maddieky/panjiva-code`.

1 Introduction

Researchers, policymakers, and firms increasingly turn to nontraditional, administrative, or other so-called "big data" to measure economic activity. These data are often available more quickly and offer a finer level of disaggregation than official statistics, but they can also pose new challenges. Without a proper understanding of features such as conceptual definitions, representativeness, and reporting details, such data can result in improper inference, biased forecasts, or non-replicable results. This paper provides the first detailed analysis of the utility of a major source of nontraditional administrative data related to international trade: the shipment-level bill of lading (BoL) data collected by U.S. ports.

BoL data provide a number of advantages and disadvantages relative to other publicly accessible official data and confidential datasets. In this paper, we use S&P Panjiva as our source of BoL data, as they provide both the raw data and also a number of useful derivative variables, including identifiers that allow researchers to longitudinally track firms engaged in international trade. In Section 2, we describe the data in detail, while in Section 3, we explore the advantages and disadvantages of the data.

Advantages of the data include detail, timeliness, and the data's unrestricted nature. Data are available at the shipment level, often with company names for both the shipper (exporter or freight forwarder) and consignee (the importer, person, or firm taking final delivery of the merchandise). They are also available to researchers within weeks, rather than months or years in the case of some detailed confidential data. The ability to access the data outside of restricted environments allows easier merging with other datasets as well as diving into specific case studies that can help illuminate how these shipments work in practice for both researcher and reader. A list of the top 10 U.S. consignees and shippers in these data is simple and illustrative to show (see tables 4 and 5, below), but doing the same with public U.S. data would be impossible, and with confidential U.S. data, explicitly prohibited.

As with all datasets, BoL data have disadvantages as well. U.S. law restricts public access to bill of lading records to only those shipped via vessel; some countries have broader access, but in this paper, we will largely focus on the strengths and weaknesses of U.S. data. In addition, shipment values are missing from BoL data. Quantity measures and descriptions are included, but mapping these to commonly used product classifications and estimating values could introduce measurement error. Companies also have the right in U.S. law to redact their name from the records, which can hamper efforts to track supply chains comprehensively.

One of the most novel aspects of these data is information on the shippers (exporters) and consignees (importers) for each shipment. This is unique among publicly available datasets, especially for the United States, where access to and disclosures from the U.S. Census Bureau's confidential Longitudinal Firm Trade Transactions Database (LFTTD) are highly restricted.

In Section 4, we dive into the characteristics of these shippers and consignees to better understand global supply chains. We show how over 60 percent of U.S. consignees have only one foreign shipper, but that these consignees represent less than 20 percent of import volumes. We also find that most shipper-consignee pairs ship in only three or fewer months in a year, with a surprisingly small number of pairs shipping every month, though these monthly shippers represent over 50 percent of trade. We also show how the number of shippers per consignee dropped dramatically in 2020 and remained volatile before recovering in mid-2021.

Finally, we turn to ways in which these data are well-positioned to analyze the striking effects of the COVID-19 pandemic on U.S. trade. The daily frequency of the data show how exports from India to the United States fell within just a few weeks of the start of the pandemic, and given shipping lags, how that collapse took 5-10 weeks to show up in U.S. import data.[1]

We can also analyze the margins on which imports collapsed and subsequently recovered: the intensive margin of changes for a given consignee-shipper pair, the net extensive margin of entry and exit of consignees, and the switching by a given consignee to a different shipper, a different country, or both. We begin to analyze these margins by focusing on an industry with particularly interesting trade patterns during this period: furniture. After initially plunging in the first half of 2020, demand for durable goods, such as furniture, skyrocketed, and furniture's weight and size tends to preclude shipping by air, making it an ideal case to analyze with BoL data.

We find that, during the initial collapse in trade, the extensive margin accounted for much of the plunge in trade volumes. Then, during the extraordinary rebound in U.S. imports, the intensive margin was most important for the first few quarters, with the extensive margin and switching margin slowly growing by the third quarter of the recovery. We find that the intensive margin is similarly important in the first few quarters of the recovery in *total* U.S. imports. These results provide important lessons on the limitations of supply chain flexibility in the very short run and the time required to source products from new shippers, or for new importers to enter the market.

The field of international trade is particularly well-situated to benefit from new sources of nontraditional data, such as the port data we examine. Since the pioneering work of Bernard et al. (1995), trade economists have focused on firm-level participation in inter-

[1]We examine exports from India to the U.S. because India instituted a particularly stringent lockdown at the onset of the COVID-19 pandemic, and because China—another natural candidate country—stopped making its BoL data available as tensions with the U.S. rose in 2018 and 2019.

national trade. Subsequent work by Monarch (2021) and Heise et al. (2019) has exploited information on the timing and frequency of trade transactions. This research has been conducted almost entirely using the confidential data available from the U.S. Census Bureau, which comes with strict access and disclosure restrictions. The administrative data collected by ports offer an alternative data source for firm- and transaction-level research, without these restrictions.

Though little-used in international economics, this paper is not the first to make use of BoL data. The recent availability of this processed BoL data is enabling researchers to conduct more detailed studies of global trade flows, supply chains, and firm operations. Ganapati et al. (2021) use an extract of BoL data (also from Panjiva) and pair it with with vessel location data derived from transponders used for navigational safety purposes. They use these combined data to present new stylized facts on the shipping network of global trade flows, with corresponding implications for trade costs. Bonfiglioli et al. (2020) uses BoL data from Piers to show that richer countries have higher average sales per firm from two sources of heterogeneity. In related work, Bonfiglioli et al. (2021a) shows that market concentration in international trade has fallen overall.[2] In addition, Feenstra and Weinstein (2017) use BoL data to estimate the concentration of exporters to the United States from markets outside of Canada and Mexico.

In addition to the trade literature, BoL data have been combined with financial datasets to yield new insights on the behavior of firms that operate internationally. Jain et al. (2014) construct a novel dataset by combining BoL data with publicly available country-year-level data on business regulations and firm-quarter-level accounting data to evaluate the participation of different firms and sectors in global trade. Jain and Wu (2020) use BoL data to examine the sourcing of different categories of imported goods by firms with global supply chains, exploring the relationship between firms' global sourcing strategy and future profitability. Bruno and Shin (2020) match BoL data from Mexico to financial data from Capital IQ (both available from S&P). They show that when the U.S. dollar appreciates, dollar wholesale-funded banks pare back credit to Mexican exporters, hampering their exports.[3]

[2]Bonfiglioli et al. (2021b) review the literature on heterogeneous firms in trade with additional results derived from BoL data.

[3]They supplement the Panjiva data with estimates from PIERS to fill out the dollar value of imported goods, as this variable is largely missing via Panjiva. These papers note the particular challenges that comes with working with BoL, specifically widespread spelling inconsistencies, as well as the various use of trade names and subsidiaries.

2 Data Description

Bill of lading data from S&P Panjiva-—the data provider we use for this analysis-—contain over one billion transaction-level records of goods traded across borders, with information including consignees and shippers, product descriptions, quantity, and, in limited cases, estimated values of shipment transactions (in USD). The data provide trade flows across 17-country-level datasets, including Bolivia, Brazil, Chile, China, Colombia, Costa Rica, Ecuador, India, Mexico, Panama, Pakistan, Paraguay, Peru, Sri Lanka, Uruguay, the United States, and Venezuela. For each of these countries, data users are able to observe both imports and exports of goods for all trading partners.

We focus our analysis specifically on U.S. import data and, to a lesser extent, U.S. export data. Panjiva provides transactions since 2007 for imports and since 2009 for exports. U.S. import data are updated several times per week, but U.S. export data updates are typically delayed by a 23-day lag for regulatory reasons (Panjiva).

Panjiva acquires these data by collecting bills of lading from U.S. Customs and Border Protection (CBP), which are freely available under the Freedom of Information Act of 1966 (FOIA). A BoL is a legal document that serves as a record that a shipment has been transported from its origin to its final destination. It also details the contract between the shipper and consignee. Each BoL requires companies to fill out various fields, including shipper/consignee name and address, description of the goods, vessel name, transport company name, ports of lading (loading) and unlading (unloading), weight, quantity, and container information. (See Appendix Figures 18 and 19 for the CBP inward (import) and outward (export) cargo declaration forms.)

In addition to providing the raw information collected on Bills of Lading, Panjiva generates additional variables that may be of use to researchers. First, Panjiva imputes a standard measure of volume, twenty-foot equivalent units (TEUs), based on existing container information and other shipment characteristics. Second, while BoL forms require product descriptions, they do not collect data on Harmonized System (HS) product codes. Panjiva attempts to assign HS codes to each shipment by searching product descriptions for HS codes that may have been optionally included by shippers and by using a text processing algorithm to translate descriptions to HS codes. Third, Panjiva attempts to provide an estimate of the value of a transaction since this information is not required in a BoL. As discussed below, these values, which are based on publicly-available average unit values, are only estimates and they are also currently unavailable for most transactions. Fourth, Panjiva also includes a unique company ID variable that can be used to link the trade transactions of some shippers and consignees to their associated companies in other S&P Global datasets, such as S&P Capital IQ. One limitation is that this company

ID linking variable only exists for 10-15 percent of shippers and consignees in U.S. import data at this time, so not all transactions can be linked to S&P's broader ecosystem of data.

Table 1 lists variable names and descriptions for some of the key variables contained in the Panjiva BoL data, with the top panel reporting raw BoL data variables and the bottom panel reporting variables that are imputed by Panjiva. Next, we compare aggregated BoL data against official U.S. Census data.

Table 1: U.S. import data description for select variables

Raw variable	**Description**
`arrivaldate`	Arrival date of shipment
`shpname`	Entity Resolved name of the shipper
`conname`	The party to take final delivery of the merchandise
`shpmtorigin`	Location from which shipment left for the U.S
`portoflading`	Port of lading
`portofunlading`	Port of unlading
`weightkg`	Shipment weight in kilograms
`vessel`	Name of the vessel that transported the goods
Imputed variable	**Description**
`panjivarecordid`	Unique Panjiva ID for shipment record
`shppanjivaid`	Unique Panjiva ID for party acting as shipper
`conpanjivaid`	Unique Panjiva ID for party acting as consignee
`volumeteu`	Volume of shipment in TEU
`valueofgoodsUSD`	Value of goods in USD
`hscode`	Harmonized Item Description and Coding System (HS)
`companyid`	Capital IQ company ID

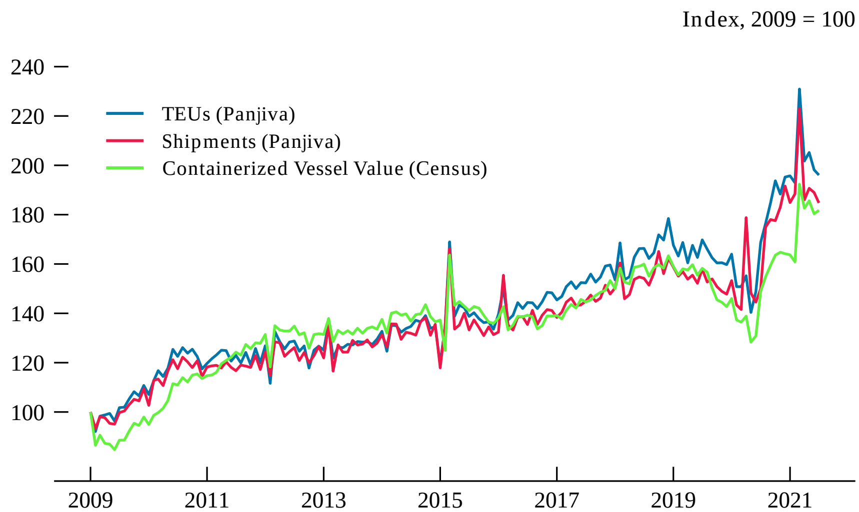
The massive size of these datasets combined with continuous updating makes data management a particular challenge. In Appendix B, we describe some key features of the system we've created at the Federal Reserve Board of Governors to update, store, and query the complete raw BoL data files.

How Bill of Lading Data Compare to Census Data

Here, we evaluate how well BoL aggregates align with official public trade aggregates. Figure 1 shows two measures of trade volume from BoL data: containers, measured by twenty-foot equivalent units, or TEUs (in blue) and shipments (in red), both normalized so that 2009 = 100. A shipment is the cargo, regardless of size, recorded in a single bill of lading.[4] That TEUs and shipments track one another closely implicitly highlights the

[4] Around four percent of shipments in the raw U.S. import data from U.S. Customs and Border Protection share a BoL number with at least one other shipment. These shipments may represent duplicate observations, though in at least some cases the arrival date is different for the two shipments while other fields are the same. Researchers should be aware of these potential duplicate shipments and consider whether their research questions require further actions to address them.

Figure 1: Comparison of bill of lading data and Census containerized vessel value for U.S. imports



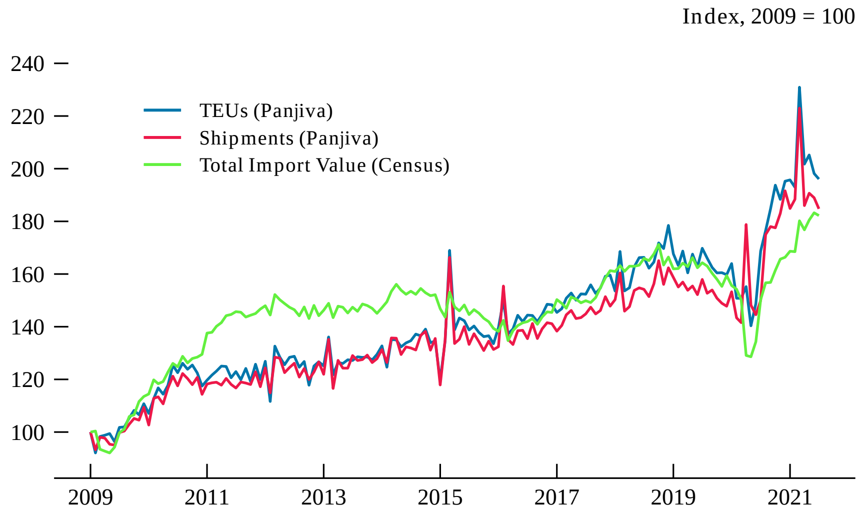
Source: S&P Global Market Intelligence, U.S. Census, and authors' calculations.
Notes: Seasonally adjusted.

stability in the average number of TEUs per shipment. In order to exclude transshipments that ultimately end up in a different country, we limit this analysis (as well as all our further analysis of U.S. import BoL data) to shipments where the consignee country is either listed as the United States or is missing. In addition, while BoL data contain non-containerized vessel trade, in particular oil imports, these do not have corresponding TEU values and represent relatively few shipments of large value. Therefore, the most relevant publicly available measure of trade flows to compare to our BoL measures is the containerized vessel import value available from the Census Bureau. Importantly, Figure 1 indicates that this nominal measure aligns quite closely with the BoL volume measures over time.[5]

Figure 2 shows the comparison between BoL aggregates and total U.S. goods import value. Here, we see that BoL data still capture the broad pattern of trade growth, as well as the dramatic trade collapse and recovery during 2020. Relative to Figure 1, the Census Total Import Value in Figure 2 includes non-maritime trade as well as vessel trade not via containers, notably including oil imports, which leads to some modest differences with the Panjiva trade measures. Oil prices were elevated in 2011-2014, for example, contributing to the Census total value line being above the lines from BoL data. As we

[5]Import price inflation in these goods is near zero: for nonpetroleum goods, BEA national accounts data show annualized import price inflation of −0.35% over the 11-year span of 2010Q1 to 2021Q1.

Figure 2: Comparison of bill of lading data and Census total value for U.S. goods imports



Source: S&P Global Market Intelligence, U.S. Census, and authors' calculations.
Notes: Seasonally adjusted.

discuss in Section 3.2, the limitation to only maritime trade with U.S. BoL data should be carefully considered in the context of each research question. For example, as we examine the 2020 trade collapse and recovery, we focus on categories like furniture rather than medical equipment or semiconductors, as the latter two categories are more likely to use air shipping.

3 Advantages and Limitations of Bill of Lading Data

3.1 Advantages of Bill of Lading Data

Bill of lading data have a variety of advantages relative to official trade statistics, making them a valuable resource for both researchers and policymakers.

The first benefit of these data stems from the fact that shipments are associated with specific firms on both the shipper (exporter) and consignee (importer) side of the transaction. The combination of these data allows consideration of firm characteristics such as the frequency of shipments per consignee, which provides important information about the nature of firms' procurement systems (Heise et al. 2019). We discuss interesting stylized facts based on exploiting the shipper and consignee identifiers in Section 4.

A second benefit of the data is their high frequency. Official trade data are available at a monthly frequency, but BoL data track shipments arriving or departing the U.S.

at a daily frequency. This higher frequency is important in many contexts, with one prominent example being an analysis of the timing of the collapse in trade associated with COVID-19. Our examination in Section 5.1 of U.S. imports from India during the initial days of the global pandemic lockdown reveals intra-month shifts in trade that are simply not observable at the monthly level.

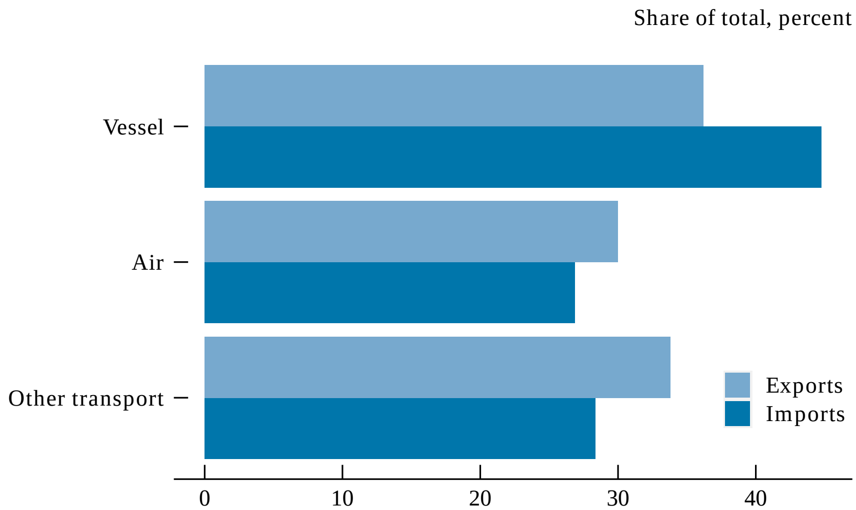
A third benefit of the BoL data is the timeliness of their release relative to official data. Official monthly U.S. trade data typically lag the close of a month by more than 30 days. By contrast, and as summarized in Appendix Figure 15, BoL data are updated nearly continuously; data for a particular day are reasonably complete within 10-14 days. This timeliness allows for observation of supply chain disruptions, such as those arising from COVID-19 or the blockage of Suez Canal in essentially real-time.

A final and intriguing benefit of BoL data is the potential of combining transaction-level data from multiple trading partners. Combining data in this manner opens the possibility of linking shippers and consignees across multiple countries' trade data, allowing for a level of detail on firms' global supply chains that is not available elsewhere, even in confidential transaction-level trade data from the U.S. Census Bureau. Linking multiple countries' data also holds the potential of observing trade networks (see e.g. Bernard and Moxnes (2018) and Dhyne et al. (2021)) and the propagation of supply chain shocks across firms and borders (Boehm et al. 2019).

3.2 Limitations of Bill of Lading Data

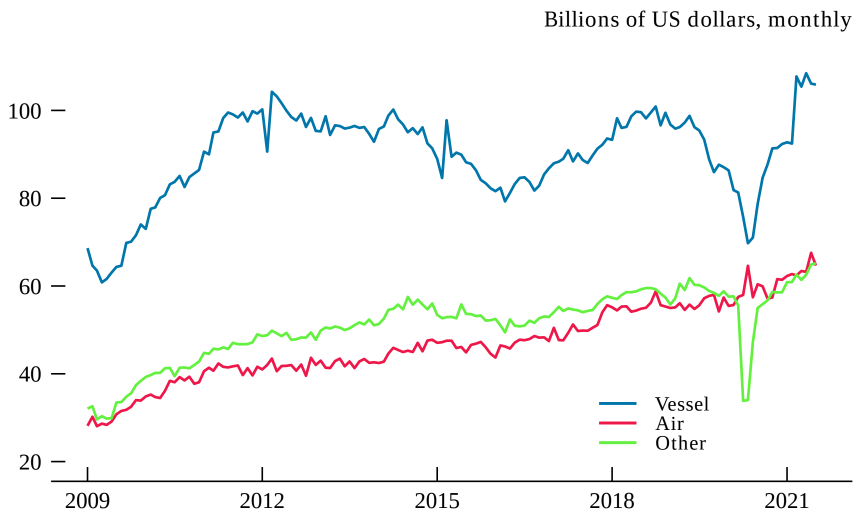
While the advantages for these data relative to publicly available sources can be substantial, there are also some limitations about which researchers should be aware. These limitations include missing or redacted data, as well as a general lack of non-imputed data on transaction values.

Figure 3: U.S. trade shares by mode of transport, 2019



Source: U.S. Census and authors' calculations.
Notes: Other includes rail, vehicle, pipeline, etc.

Figure 4: U.S. imports by mode of transport



Source: U.S. Census and authors' calculations.
Notes: Other includes rail, vehicle, pipeline, etc. Seasonally adjusted.

One of the key limitations of bill of lading data is its lack of information on non-maritime trade for the United States. As indicated in Figure 3, maritime trade—i.e. trade transported by vessel—is the largest mode of transport by value, accounting for nearly 50 percent of the value of U.S. imports and nearly 40 percent of the value of U.S. exports in 2019. Nonetheless, the remaining value of U.S. trade, which is split between air and land-based transport like trucks, railroads, and pipelines, is not available in U.S. bill of lading data. Moreover, as shown in Figure 4, the relevance of this exclusion has also grown somewhat over time, with land and air increasing in importance as modes of transportation.

Table 2: Trade Shares by Mode of Transport, 2019

	Imports			**Exports**			**Total Value***
	Vessel	**Air**	**Other**	**Vessel**	**Air**	**Other**	
Mexico	9.39	1.98	88.63	12.37	3.50	84.13	608.43
Canada	5.34	4.61	90.05	4.79	6.25	88.96	607.85
China	63.61	28.96	7.43	49.24	43.00	7.76	557.81
Japan	71.80	24.57	3.63	51.79	40.09	8.13	217.77
Germany	51.62	40.39	7.99	33.94	57.75	8.32	187.00

Source: U.S. Census and authors' calculations.

Notes: Includes top 5 U.S. trading partners by value. *In billions.

The exclusion of air and land-based trade also leads to substantial differences in coverage across major U.S. trading partners. As shown in Table 2, trade with Mexico and Canada—two of the largest trading partners of the United States—is conducted almost entirely via land-based modes of transportation. Bilateral U.S. trade with those countries, therefore, is largely excluded from BoL data. However, the vessel share of trade is, unsurprisingly, much higher for other important U.S. trading partners outside North America. Trade by vessel accounts for 64 percent of the value of U.S. trade with China, 72 percent of the value of trade with Japan, and 52 percent of the value of trade with Germany.

Missing Data

Most big data sources suffer from missing information in some observations, and BoL data from Panjiva is no exception. There are two primary sources of missing data in Panjiva: fields for which a firm requests that the U.S. Customs and Border Protection (CBP) redact their identity in the shipper or consignee field, and fields like TEU, HS code, and value that Panjiva imputes from other information that is not always available. Generally speaking, fields that are directly filled in on CBP form 1302 (see appendix Figure 18) are available for the vast majority of observations.

Table 3 reports the share of U.S. import observations for which particular key variables

Table 3: Missing U.S. Import Data by Variable (Percent)

	Shipper ID	Consignee ID	HS Code	TEU	Value
2007	19.9	16.7	4.6	3.6	100.0
2008	22.6	22.6	4.9	3.8	95.2
2009	30.2	27.6	3.9	3.4	68.8
2010	33.8	31.5	5.0	3.1	69.9
2011	34.9	35.1	4.9	3.3	70.4
2012	33.5	31.0	4.4	3.4	69.8
2013	25.2	8.5	3.5	3.1	67.9
2014	23.7	8.0	4.2	3.0	65.7
2015	23.7	8.0	3.7	3.0	63.5
2016	28.4	10.6	3.8	2.7	63.8
2017	32.1	12.9	3.9	2.8	64.6
2018	32.6	14.2	3.7	2.7	68.3
2019	33.5	14.0	6.6	2.6	67.0
2020	31.4	12.3	7.3	1.8	67.5
2021	33.9	16.8	4.8	1.8	67.7

Source: S&P Global Market Intelligence and authors' calculations.

are missing. As shown in the table, the variables for the shipper/consignee IDs and value have the highest probability of being missing, while the HS code and twenty-foot equivalent unit (TEU) fields are missing in a much lower share of observations.[6] A few key variables, such as weight and shipment origin country, are not included in the table since they have nearly zero missing observations in U.S. import data. Importantly, the share of observations with missing data for particular variables can vary fairly substantially over years. For example, across the years 2007 to 2021, the share of observations with missing shipper (consignee) IDs ranges from 19.9 percent (8.0 percent) to 34.9 percent (35.1 percent).

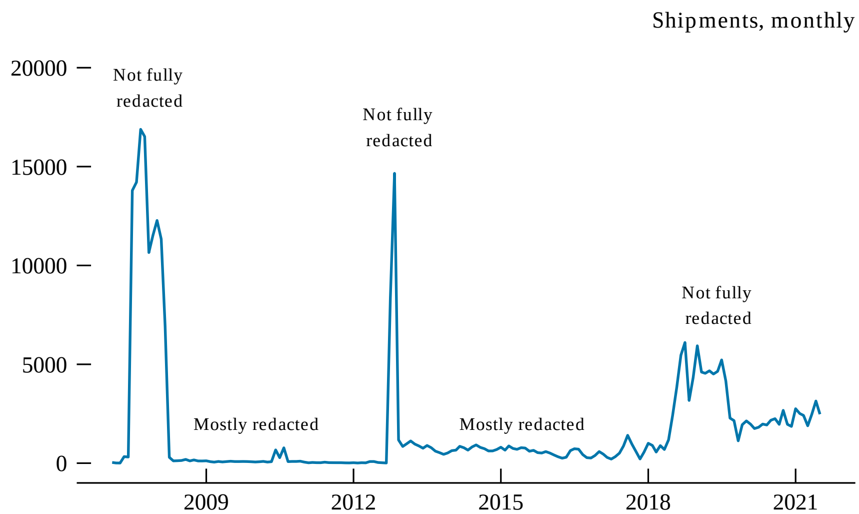
Firms' requests for redactions of shipper and consignee information contribute to variation in the share of missing data over time. After a firm requests redaction, this request is fulfilled for two years before requiring renewal. When a request expires, a firm's transactions from that point forward are no longer redacted. These redaction requests must be made for a specific firm name, so firms that use multiple names on bills of lading must submit a request for each entity. Given that one feature Panjiva adds to the raw data is the matching of firm names (including likely typos) to a corporate entity in their overall data framework, this can lead to firms having some but not all of their shipments represented in the database.

One important illustration of this phenomenon is Walmart, which appears to redact its

[6]Missing TEU values can simply reflect shipments that are not containerized, such as oil imports.

information imperfectly. As shown in Figure 5, Walmart's monthly shipments generally hover around zero but spike briefly in 2007, 2012, and sporadically between 2017 and 2021, which suggests that some of Walmart's redaction requests may have briefly expired before they were subsequently renewed.

Figure 5: Walmart, Inc.'s monthly shipments



Source: S&P Global Market Intelligence and authors' calculations.

Limited data on trade values

While BoL data consistently report transaction weights, they typically lack data on the value of trade associated with each transaction. A small share of transactions include data on trade value pulled from the transaction description in the BoL. However, for the majority of other observations the shipment value is either missing or imputed by applying average unit values from public trade data to the BoL weights. As a result, over 60 percent of observations have missing data for shipment value.

Product descriptions versus product codes

As described above, CBP forms require shippers to report product descriptions, but not HS Product Codes. The HS codes provided in the data, therefore, are not official HS codes, but rather are assigned based on Panjiva's proprietary algorithm. The assignment is actually quite comprehensive: as as indicated in Table 3, the imputed HS code variable is generally well-populated, with five percent or fewer of observations missing for this variable in all but two years. Nevertheless, it is important to emphasize that BoL records are based on shipments, and therefore an individual record (and hence unit of quantity) could be comprised of more than one (and often many) individual products. This feature can make disaggregation by product an imperfect exercise.

4 Characteristics of Shippers and Consignees

One of the most novel aspects of BoL data is the detailed, shipment-level information on shippers and consignees. Subject to the firm-level redactions described above, researchers can track company-specific details over time, including a company's trading partners, its frequency and weight of shipments, its ports of lading and unlading, and even its contact information. In addition, Panjiva assigns unique ID codes to all shippers and consignees after collecting and parsing firm names from bill of lading data, which makes it easier for users to identify and track specific companies as well as merge BoL data with other datasets.

Table 4: Top consignees by total TEU, 2020

Consignee name	**Total TEU**	**TEU** (%)	**Shipments** (%)
Expeditors International	1,150,675	5.23	6.30
Ups Supply Chain Solutions	779,248	3.54	2.79
Dole Fresh Fruit Co.	236,312	1.07	0.52
Samsung Electronics	187,707	0.85	0.55
Chiquita Fresh North America Llc	171,205	0.78	0.10
Maersk Line	170,783	0.78	0.01
Fedex Trade Networks Transport	162,425	0.74	0.88
Seaboard Marine Ltd.	138,075	0.63	0.02
Geodis USA Inc.	123,801	0.56	0.36
Carmichael International Service	113,359	0.52	0.37

Source: S&P Global Market Intelligence and authors' calculations.

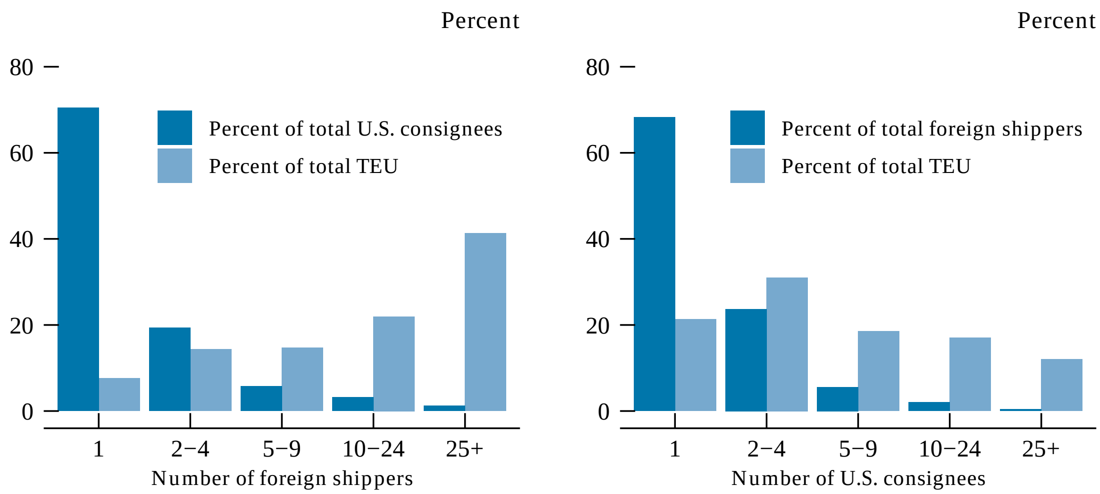
With these data, researchers can analyze certain industries or countries by reporting top suppliers and buyers. Tables 4 and 5, for example, report the top 10 U.S. consignees and foreign shippers, respectively, in U.S. import data. Table 4 reveals that eight of the top 10 consignees are freight and logistics companies, highlighting the importance of intermediaries in the actual execution of international trade. Table 5 shows that the top 10 foreign shippers to the United States are a mixture of these transportation companies, electronics and agricultural producers, and, improbably, Red Bull. As users of confidential Census Bureau data are well aware, revealing this type of information with those datasets is impossible.

Table 5: Top shippers by total TEU, 2020

Shipper name	Country	TEU	TEU (%)	Shipments (%)
Thor Joergensen A S	Denmark	170,351	1.04	0.01
Chiquita Brands International SARL	Switzerland	153,933	0.94	0.10
Sm Line Corp.	United States	75,207	0.46	0.00
Thai Samsung Electronics Co., Ltd.	Thailand	60,748	0.37	0.20
Samsung Electronics Co., Ltd.	South Korea	56,431	0.34	0.32
Lg Electronics Inc.	South Korea	56,181	0.34	0.18
Samsung Electronics Digital	Mexico	44,968	0.27	0.13
Red Bull GmbH	Austria	43,309	0.26	0.03
Union De Bananeros Ecuatorianos S.A. Ubesa	Ecuador	36,198	0.22	0.20
Seadom Units	Dom. Republic	35,766	0.22	0.00

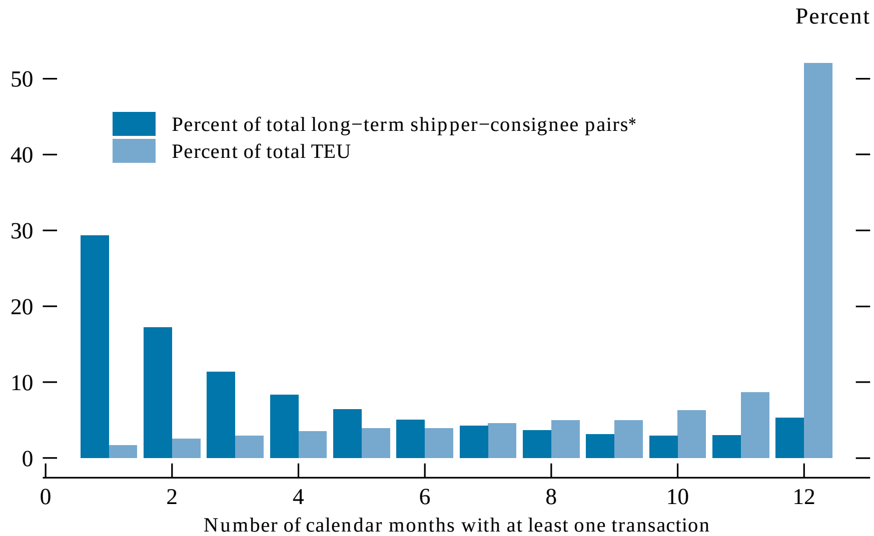
Source: S&P Global Market Intelligence and authors' calculations.

Figure 6: Shippers and consignees by TEU, 2019



Source: S&P Global Market Intelligence and authors' calculations.

Figure 7: Frequency of transactions by shipper-consignee pair, 2019



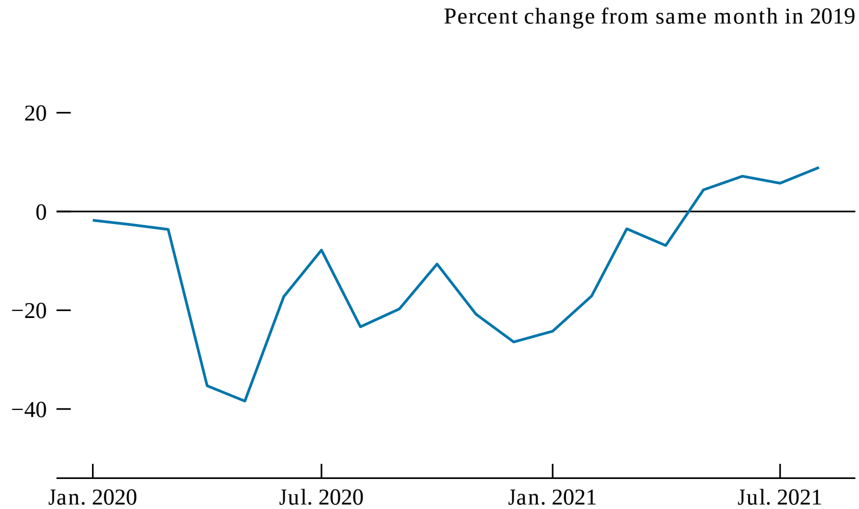
Source: S&P Global Market Intelligence and authors' calculations.

*Includes shipper-consignee pairs that traded at least once in the previous year (2018).

Bill of lading data offer further information on firm-level trade that are unobservable in public official data. As shown in the left panel of Figure 6, the majority of U.S. importers have a single foreign trading partner, but these firms account for a disproportionately small share of total U.S. imports by TEU. By contrast, only a small handful of U.S. importers have many trading partners (over 1000 partners, in some cases), but this small number of firms accounts for a disproportionately large share of imports by TEU. Moreover, the number of shippers and total TEU per consignee are positively correlated. These patterns are largely the same when we switch attention to the number of U.S. consignees per foreign shipper (left panel of Figure 6, and taken together, they highlight the significance of large firms in international trade.[7] In addition, the majority of shipper-consignee pairs interact infrequently in a given year, which emphasizes the lumpiness of trade by pair. For example, in 2019, only 5% of all long-term shipper-consignee pairs traded at least once each month, while 47% of all pairs only traded in one or two months of the year (Figure 7).

[7]For example, the largest consignee, Expeditors International, accounted for 5.2% of total U.S. imports by TEU and 6.3% of total shipments in 2019.

Figure 8: Change in shippers per consignee



Source: S&P Global Market Intelligence and authors' calculations.

This shipper-consignee data can also be used to track how disruptions such as recent Covid-related lockdowns affect these relationships. In Figure 8, we plot monthly data on the percent change in the number of shippers per consignee relative to the previous year. As shown in the figure, the number of shippers per U.S. consignee dropped by over 35% in April 2020 relative to April 2019, as importers only managed to maintain a small share of typical trading relationships. This pattern of reduced shippers per consignee continued through much of 2020..

5 Trade and the COVID-19 Pandemic

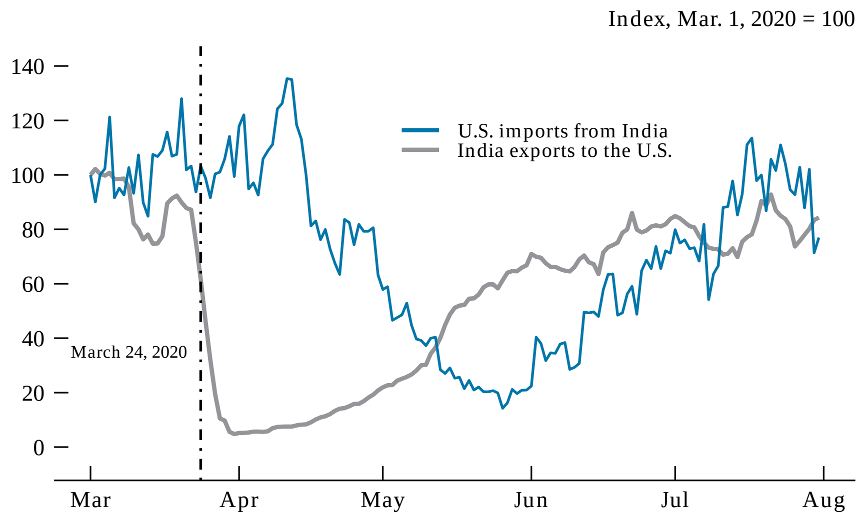
As was just mentioned, the timeliness and granularity of BoL data are especially valuable in understanding the enormous changes to international trade patterns brought on by the COVID-19 pandemic. This section details several insights from these data about the collapse and resurgence of trade during 2020-2021.

5.1 The precise timing and effects of country-level lockdowns

Unlike official statistics, the daily frequency of the BoL transaction-level data allow the observation of intra-month patterns of trade. This feature is particularly useful in evaluating the impact of shocks to trade, with perhaps the largest and most abrupt in the modern era coming from the various country-level lockdowns associated with the early stages of

COVID-19. We leverage the multiple sources of information coming from BoL data to highlight the transmission of the trade shock from the March 2020 national lockdown in India to U.S. imports.

Figure 9: U.S.-India shipments during the COVID-19 lockdowns



Source: S&P Global Market Intelligence and authors' calculations.
Notes: This figure plots the 7-day moving average of shipments of U.S. imports from India and India exports to the United States, with each indexed to equal 100 on March 1, 2020.

We focus on the specific case of India because that country instituted a particularly strict COVID-19 lockdown, because pandemic-era U.S.-India trade has been relatively unstudied, and because bill of lading data are available for Indian exports to the U.S.[8] As shown in Figure 9, the national lockdown announced by the Indian government on March 24, 2020 is evident in the immediate decrease in India's exports to the United States and then subsequently in the delayed drop in U.S. imports from India several weeks later. The high-frequency BoL data reveal a much sharper drop in Indian exports to the U.S. than would be visible with monthly-frequency publicly available data. Moreover, the patterns in Figure 9 reveal important information on the translation of this shock into U.S. imports: The drop in U.S. imports from India is considerably less steep than the drop in Indian exports and lagged by 4 weeks. More broadly, Figure 9 indicates that BoL data can help researchers learn how the timing of such transmission of trade shocks varies across trading partner based on distance, shipping routes (such as the use of entrepôt trading hubs), and the particular characteristics of the shock.

[8]BoL data on exports from China to the U.S. are not available after March 2018.

5.2 Decomposing the collapse and subsequent surge in U.S. imports

The enormous drop in trade in the first quarter of 2020 was followed by a remarkable recovery, such that U.S. import volumes surpassed typical levels by the middle of 2020. Given the surprising speed of the resurgence in trade, a natural question is how importers and exporters managed to increase shipments so dramatically. For one useful perspective on both the collapse and subsequent surge in U.S. imports, we decompose the import changes based on the following margins at a quarterly frequency:

- Entry/Exit of Consignees Margin: The changes in imports due to the net entry and exit of consignees across two quarters.[9]
- Add/Drop Shipper or Country Margin: The changes in imports across two quarters from a given consignee that changes either the shipper or the country associated with the import transaction.
- Intensive Margin: The changes in imports from a given consignee—shipper—country pair across two quarters.
- Redacted: The changes in imports coming from changes in the pool of redacted consignees across two quarters.

Apart from the complicating feature of redactions, the decomposition outlined above is similar in spirit to the work of Bernard et al. (2009), which uses confidential, firm-level Census data. By contrast, with official public data, researchers are forced to define the extensive margin as something like an HS10 code coming from a particular country. That level of aggregation, however, would not capture the changes in relationships associated with entry/exit of consignees or switching among suppliers by continuing consignees. BoL data allow for the ability to track relationships defined at the consignee$\times$shipper$\times$country level.

To focus attention on the dynamics introduced by COVID-19, we fix the baseline period to be the fourth quarter of 2019, and then track the change along each margin in subsequent quarters. We begin with a decomposition of furniture imports (Chapter 94 in the HS classification system) due to the dramatic changes in demand experienced by this product group during our period of study. In addition, unlike some other categories, furniture is unlikely to be moved by air. Finally, to account for the significant seasonal patterns and trends in the data, we calculate the identical decomposition for each of the previous three years (baseline quarters of 2016Q4, 2017Q4, and 2018Q4), and for each

[9] A consignee is considered to have exited in a particular quarter if it has no imports during that quarter. A consignee is considered an entrant in a particular quarter if it had imports during that quarter but had no imports in 2019Q4.

margin of adjustment and then subtract out the average change across each time horizon from the COVID-19 period. The results are displayed in panel (a) of Figure 10.

The black line shows the overall change in U.S. furniture imports, relative to 2019Q4. Imports fell modestly in the first quarter of 2020 and then more significantly in the second quarter.[10] The surge in imports for product categories such as furniture is evident in subsequent quarters, with imports up over 65 percent (seasonally adjusted) by 2021Q1 from pre-pandemic levels.

We derive several useful lessons from decomposing these overall changes into the margins of adjustment outlined above, which are illustrated by the colored bars in Figure 10a. First, the drop in U.S. imports during the initial lockdowns of COVID-19 in 2020Q1 were driven largely by net consignee exit (the red bars), a feature that continued into the second quarter of 2020. Second, although the intensive margin (light blue) accounts for the largest individual share of the increase at the end of our sample period, when we combine the two extensive margins—i.e. the net consignee exit (in red) and add/drop shipper or country margin (in dark blue)—their contribution is slightly larger than the intensive margin.[11] Hence, by 2021-Q1, roughly half of the growth in furniture imports (a nearly 30 percentage point increase relative to 2019-Q4) came from trading relationships that did not exist in 2019-Q4. Finally, increases in consignee redactions (in gray) are also an important component in the overall increase in imports; without the consignee redaction, we would have been able to allocate these transactions into one of the other margins of adjustment.

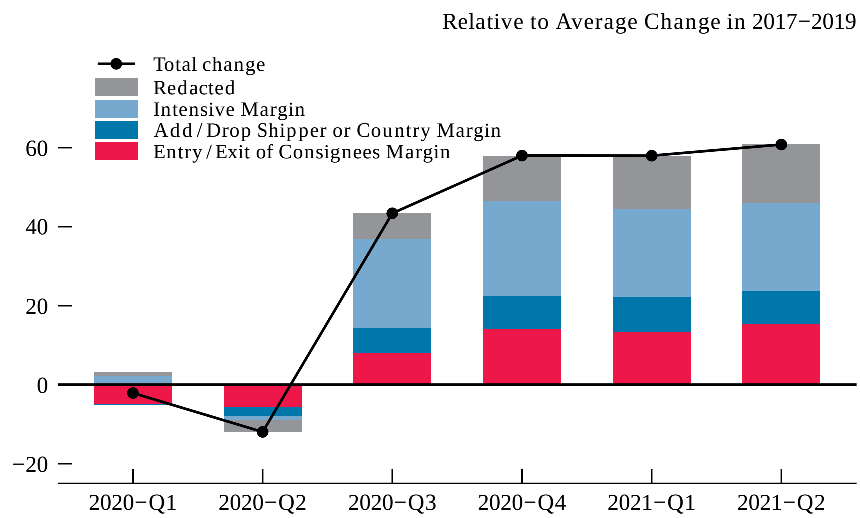
Panel (b) of Figure 10 provides a contrasting perspective by decomposing the growth in overall BoL imports during this time period. The most obvious difference relative to the decomposition for furniture imports in Figure 10a is the smaller and more gradual increase following the 2020-Q2 nadir: overall imports were up 15 percent in early 2021 relative to the baseline compared with the roughly 60 percent increase for furniture imports. A second and more surprising result comes from the component margins of this increase revealed by the decomposition. The contributions from the two extensive margins—the entry/exit of consignees margin along with the add/drop shipper/country margin—provide a notably smaller contribution to the overall increase in imports than is the case for furniture imports shown in Figure 10a. The extensive margin also accounted for a *larger* fraction of the decrease in 2020-Q2. The clear differences in the role of the extensive margin suggests there are likely important variations across products and in-

[10] Imports fell considerably more in the first quarter of 2020 on a non-seasonally adjusted basis. However, furniture imports tend to peak in the fourth quarter each year, and then fall substantially in the first quarter.

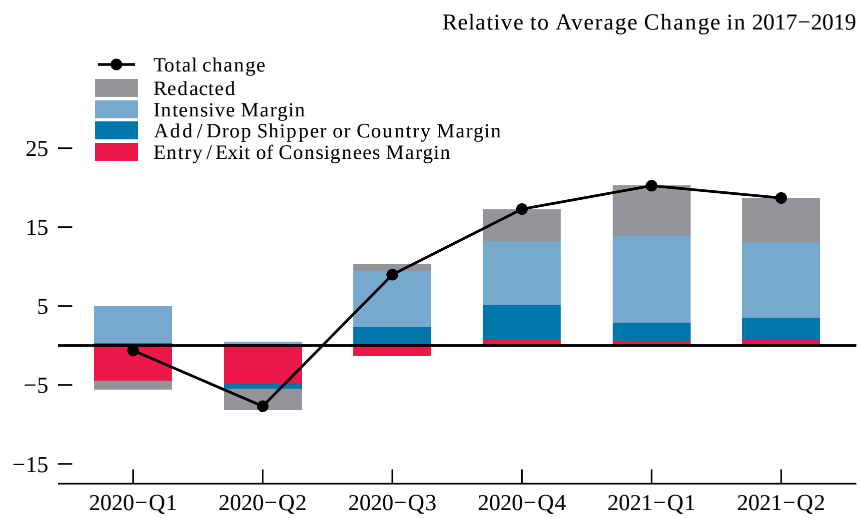
[11] As first discussed in Bernard et al. (2009) the extensive margin becomes more important as the horizon lengthens. In our case, the switching product/country margin of adjustment is predominantly composed of cases where the consignee switches suppliers but maintains the same source country.

Figure 10: Decomposing percent change in imports of (by TEU) relative to 2019Q4

(a) Furniture Imports (HS 94)



(b) Total Imports



Source: S&P Global Market Intelligence and authors' calculations.
Notes: This figure plots the quarterly change in U.S. imports (by TEU) relative to 2019-Q4 along four margins described in the text. The quarterly change for each margin is net of the average change during the equivalent quarter during 2017-2019 to account for seasonal variation and trend growth. Panel (a) restricts to imports of furniture (HS Chapter 94) whereas Panel (b) reports the decomposition for total imports.

dustry that warrant further study. Some of these differences could result from 1) the mix of supply and demand factors during the heart of COVID-19 lockdowns, 2) idiosyncratic factors affecting the ease with which trading relationships form and break, and 3) capacity constraints in existing foreign suppliers.

In summary, the BoL data allow researchers to understand the mechanisms underlying the extraordinary growth in imports during the onset of the COVID-19 pandemic. These decompositions would have been invisible using traditional, publicly available datasets.

5.3 Real-time measures of shipping bottlenecks during the COVID-19 trade recovery

The dramatic resurgence of trade in the second half of 2020 led to some much-discussed bottlenecks across many transportation modes. In this section, we show how the BoL data can be used to examine characteristics of vessel shipping that shed light on the prevalence and effects of bottlenecks in oceanic vessel shipping in nearly real time.

The use of BoL data to study the shipping network is highlighted by Ganapati et al. (2021) when used in conjunction with newly available vessel transponder data (otherwise known as Automatic Identification System (AIS) data) that tracks vessel ship movements.[12] While BoL data alone can identify the presence of indirect shipping—the primary topic of interest in Ganapati et al. (2021)—based on shipments coming from many different ports of lading on a given vessel-port of unlading combination, the key drawback is a lack of date associated with a shipment's foreign departure. The time stamp on AIS vessel movements enable researchers to track the precise route of a vessel through multiple ports of call. However, the key limitation of AIS data is a lack of any easily quantifiable measure of trade volume associated with each vessel. The analysis below leverages the vessel and ports of unlading variables that are typically reported in the BoL data, and focuses attention on the vessel congestion centered in the ports of Los Angeles / Long Beach in late 2020 and into 2021.

We take several steps to convert the raw BoL data into a dataset useful for tracking vessel arrivals at U.S. ports. First, we clean and standardize vessel name and a corresponding vessel identifier to account for inconsistencies in these variables.[13] Second, for many analyses at a vessel-port level it is helpful to restrict attention to container vessels. While external lists can identify vessels based on vessel type, for our purposes, we classified container vessels based on a measure of observed capacity: whether the maximum

[12]See Heiland et al. (2021), Cerdeiro et al. (2020) and Cerdeiro and Komaromi (2020) for examples of recent papers using AIS transponder data.

[13]We provide detail for this process in Appendix C.

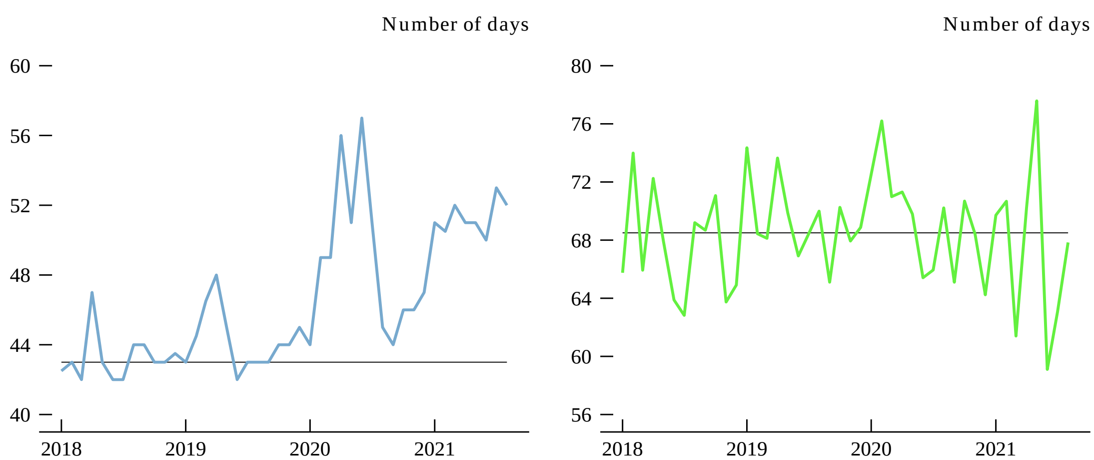
observed TEUs unloaded at a particular point of time for a vessel surpassed a threshold.[14]

Third, we must identify a specific date for a vessel unloading cargo at a U.S. port. The difficulty here lies in the fact that the "arrival date" associated with BoL records typically reflect when individual shipments clear customs. Generally speaking, a large majority of BoL import shipment records from a container vessel at a port of unlading are listed as arriving within a one or two day period. However, there are frequent exceptions in which a vessel's shipments are reported as arriving over more extended periods of time, which could lead to an incorrect inference for a vessel arrival date. These records could reflect delays in clearing customs, typos in arrival date, or differences in identifying arrival date by exporters or importers. To account for these concerns, we take our baseline dataset of daily vessel-port observations and then eliminate a daily record if that day's shipments from a particular vessel were a very low share of the vessel's (observed) maximum capacity. Finally, we consolidate a vessel's arrival date into a single day if substantial shipments occur over a period of less than five days.

Figure 11: Median number of days between vessel visits at port

(a) Port of Los Angeles/Long Beach

(b) Major East Coast Ports



Source: S&P Global Market Intelligence and authors' calculations.
Notes: This figure plots, for a given month, the median number of days since a vessel last visited the port. The black line in each figure represents the average number of days during the period 2013-2017. Major East Coast ports include the ports of Charleston, Newark/NY, Norfolk, and Savannah.

For a first look at the insights from this new dataset, we quantify the delays in vessel movements brought on by the shipping congestion experienced in 2020 and 2021. To mea-

[14]For the discussion below, we set this threshold at a relatively low value of 200, though for other purposes researchers may want to focus on vessels with larger capacity.

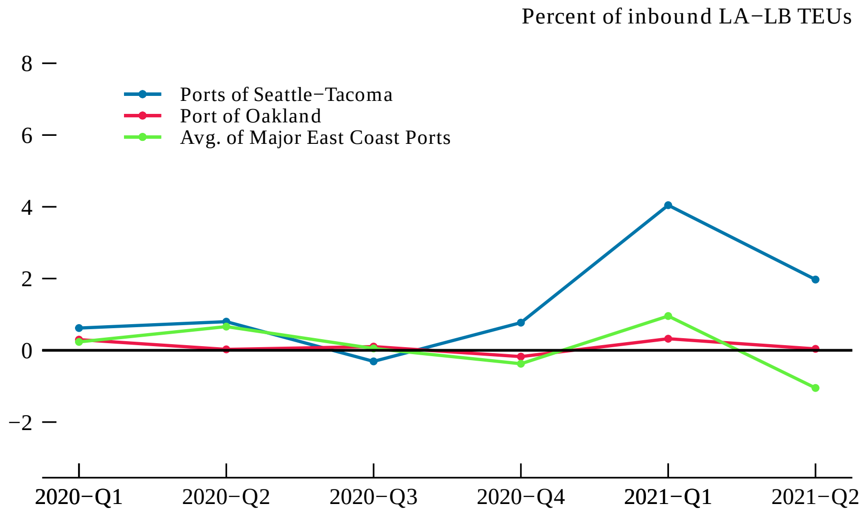
sure the typical transit times for container vessels at a given port, we calculate the number of days between return arrivals of a given vessel and calculate the monthly median value for a given port. Panel (a) of Figure 11 indicates that a typical vessel would unload cargo at the Ports of Los Angeles/Long Beach (LA/LB) about every 43 days during normal times (2013-2017). This value was relatively stable in 2018 and 2019, but spiked in early 2020 following country-level lockdowns and the more general slowdown in trade during the early period of COVID-19. Round-trip transit times normalized in the third quarter of 2020 but subsequently increased in late 2020 and early 2021 due to the congestion at the Port of LA/LB. The median number of days in between port visits of 52 during 2021-Q1 and 2021-Q2 reflects an increase of roughly 8 days from typical levels.

Panel (b) of Figure 11 shows that there has been no such systematic delays in ship processing at an average of major U.S. East Coast ports during this period. Panel (b) also shows the longer average round-trip transit time of East Coast ports, a fact which reflects the increased prevalence of multi-stage trips common for vessels servicing these ports.[15]

Given reports that the congestion at the Ports of LA/LB resulted in vessels being rerouted to unload at other ports on the U.S. West Coast, we next attempt to quantify this degree of rerouting from our BoL-based dataset of vessel-port traffic. We first identify the sample of vessels that visited the Port of LA/LB on a consistent basis in a pre-Covid period, i.e. in both Q3 and Q4 of 2019. In the subsequent six quarters (2020-Q1 to 2021-Q2) we identify the potential set of vessel re-routings as those vessels that are not observed visiting the Port of LA/LB but are observed visiting a different U.S. port. We measure the magnitude of these re-routings as the number of TEUs unloaded at alternative ports in a given quarter, which are then displayed as a fraction of total inbound TEUs at the port of LA/LB in that quarter. Finally, because what we define as re-routing may occur even in normal times, we calculate identical statistics from baseline periods in each of 2016-2018 and subtract the average of these "normal" vessel re-routings from the period of study.

[15]Median time between port visits also tends to be noisier for East Coast ports because West Coast ports have more dedicated port-to-port vessel routes, which tend to run on more predictable schedules.

Figure 12: Percent of inbound Los Angeles / Long Beach activity re-routed to other ports



Source: S&P Global Market Intelligence and authors' calculations.
Notes: This figure plots the percent of quarterly inbound LA/LB TEU imports that are identified as being re-routed to other ports. These values are net of the average observed percent re-routed to these ports during the period 2017-2019. Major East Coast ports include the ports of Charleston, Newark/NY, Norfolk, and Savannah.

The result is plotted in Figure 12 for three likely destinations of re-routings from the Ports of LA/LB: Seattle-Tacoma, Oakland, and an aggregate of four major East Coast ports. Figure 12 reveals that vessel re-routings from LA/LB to Seattle-Tacoma spiked in the first quarter of 2021 (following the onset of port congestion in late 2020) to an amount equal to roughly 8 percent of inbound TEUs at the ports of LA/LB. This re-routing declines somewhat in the second quarter of 2021 but remains elevated relative to normal levels. While some re-routings were documented in press reports to the Port of Oakland, our data indicate that these did not constitute a significant fraction of inbound TEUs from LA/LB. Similarly, the data also confirm that few, if any, vessels were re-routed on net from LA/LB to the East Coast of the United States during this period.

In summary, the unique features of BoL data, together with timely access, provide both researchers and policymakers with a useful tool to analyze disruptions to trade such as those accompanying COVID-19.

6 Conclusion

This paper provides the first detailed analysis of the utility of data from bills of lading for international trade research, specifically the information available on U.S. imports via Panjiva. These data provide a near real-time, firm-level dataset useful for addressing a variety of economic questions that cannot be addressed with other data. Furthermore, some of the limitations of U.S. import data—including a general lack of trade values, redaction of some firm names, and being restricted to vessel shipping—do not apply to the same data available for other countries.

We use the unique elements of the data to analyze international trade relationships. Over 60 percent of consignees (importers) have only one foreign shipper (exporter), but these consignees represent less than 20 percent of import volumes. Most shipper-consignee pairs ship in three or fewer months per year, though the surprisingly small number of pairs that ship every month account for over 30 percent of U.S. imports by TEU. In the COVID-induced trade collapse in 2020, the number of shippers per consignee dropped dramatically and remained volatile before recovering in mid-2021.

Finally, we explore other aspects of international trade during the COVID-19 crisis. The daily frequency shows how quickly exports from India to the United States fell following lockdowns in March 2020. Furthermore, the resulting drop in U.S. imports weeks later demonstrates clearly how international shipping lags transmit these shocks with a delay. Following the collapse, U.S. goods demand recovered briskly, and these data demonstrate the margins on which imports can rise. In the very short run, within a few months, higher imports were mostly achieved within existing shipper-consignee pairs. Over subsequent quarters, however, imports rose by consignees switching shippers or source countries, and also by the entry of new consignees.

Our work and the recent literature demonstrate that bill of lading data remains underutilized in international trade. With some caveats, these data provide a useful complementary dataset to disaggregated official public data and confidential datasets. Moreover, the ability to see most firm names of shippers and consignees opens the possibility of merging BoL data with other firm-level datasets. Researchers in international trade should consider the possibility of using BoL data in their future work.

References

Bernard, Andrew B. and Andreas Moxnes, "Networks and Trade," *Annual Review of Economics*, 2018, *10* (1), 65–85. _eprint: https://doi.org/10.1146/annurev-economics-080217-053506.

——, **J. Bradford Jensen, and Robert Z. Lawrence**, "Exporters, Jobs, and Wages in U.S. Manufacturing: 1976-1987," *Brookings Papers on Economic Activity. Microeconomics*, 1995, *1995*, 67–119. Publisher: Brookings Institution Press.

——, ——, **Stephen J. Redding, and Peter K. Schott**, "The Margins of U.S. Trade (Long Version)," Working Paper 14662, National Bureau of Economic Research January 2009.

Boehm, Christoph E., Aaron Flaaen, and Nitya Pandalai-Nayar, "Input Linkages and the Transmission of Shocks: Firm-Level Evidence from the 2011 Tōhoku Earthquake," *The Review of Economics and Statistics*, March 2019, *101* (1), 60–75.

Bonfiglioli, Alessandra, Rosario Crinò, and Gino Gancia, "Firms and Economic Performance: A View from Trade," September 2020.

——, ——, **and** ——, "Concentration in international markets: Evidence from US imports," *Journal of Monetary Economics*, July 2021, *121*, 19–39.

——, ——, **and** ——, "International Trade with Heterogeneous Firms: Theory and Evidence," *CEPR Working Paper*, June 2021, (16249).

Bruno, Valentina and Hyun Song Shin, "Dollar and Exports," April 2020.

Cerdeiro, Diego A. and Andras Komaromi, "Supply Spillovers During the Pandemic: Evidence from High-Frequency Shipping Data," *IMF Working Papers*, December 2020, (WP/20/284).

——, ——, **Yang Liu, and Mamoon Saeed**, "World Seaborne Trade in Real Time: A Proof of Concept for Building AIS-based Nowcasts from Scratch," *IMF Working Papers*, May 2020, (WP/20/57).

Dhyne, Emmanuel, Ayumu Ken Kikkawa, Magne Mogstad, and Felix Tintelnot, "Trade and Domestic Production Networks," *The Review of Economic Studies*, March 2021, *88* (2), 643–668.

Feenstra, Robert C and David E Weinstein, "Globalization, Markups, and US Welfare," *Journal of Political Economy*, August 2017, *125* (4), 35.

Ganapati, Sharat, Woan Foong Wong, and Oren Ziv, "Entrepôt: Hubs, Scale, and Trade Costs," July 2021.

Heiland, Inga, Andreas Moxnes, Karen Helene Ulltveit-Moe, and Yuan Zi, "Trade From Space: Shipping Networks and The Global Implications of Local Shocks," March 2021.

Heise, Sebastian, Georg Schaur, Justin R Pierce, and Peter K Schott, "Tariff Rate Uncertainty and the Structure of Supply Chains," January 2019.

Jain, Nitish and Di (Andrew) Wu, "Can Global Sourcing Strategy Predict Stock Returns?," SSRN Scholarly Paper ID 3606884, Social Science Research Network, Rochester, NY May 2020.

——, **Karan Girotra, and Serguei Netessine**, "Managing Global Sourcing: Inventory Performance," *Management Science*, May 2014, *60* (5), 1202–1222.

Monarch, Ryan, ""It's Not You, It's Me": Prices, Quality, and Switching in U.S.-China Trade Relationships," *The Review of Economics and Statistics*, February 2021, pp. 1–49.

Panjiva, "S&P Global Market Intelligence, Panjiva Supply Chain Intelligence Platform & Data Feed."

Wasi, Nada and Aaron Flaaen, "Record Linkage Using Stata: Preprocessing, Linking, and Reviewing Utilities," *The Stata Journal*, October 2015, *15* (3), 672–697. Publisher: SAGE Publications.

A Comparing Panjiva, U.S. Census, and Port Data

In this section of the appendix, we compare aggregated shipping volume BoL data to data directly reported by ports themselves and also to official U.S. Census Bureau data. These checks provide information to researchers considering the representativeness of the BoL data. We focus on comparing volume measures in BoL (weight and TEU), as they are more comprehensively available than imputed values, which suffer from both missing observations and extensive measurement error. Generally speaking, weight and TEU are very similar volume measures over time and could substitute for one another given most questions. In short, we find that BoL data closely track official port data in TEUs.

A.1 Comparing statistics by port

Table 6 compares measures of trade weight and number of TEUs by port, as reported in BoL data and by Census. BoL weight measures tend to exceed Census measures somewhat. Still, as the columns labeled "share" demonstrate, the proportion of imports going to each port is similar between Census and BoL, with the notable exception of Los Angeles and Long Beach: Here, the sum of the two ports is more comparable than their individual identification. This adds to the list of reasons why it is best practice to treat LA/LB as a single economic entity for most questions with these data.

The right two columns of Table 6 provide the total TEU count in 2019 by port for BoL data and data provided by the ports themselves. In most cases, these correspond remarkably closely.

Table 6: Comparison of Panjiva and Official Statistics, 2019

	Panjiva Weight*	**Census Weight***	**Panjiva Weight Share**	**Census Weight Share**	**Panjiva TEU***	**Official Port TEU***
Houston, TX	50,497	55,630	0.08	0.09	1.23	1.24
Los Angeles, CA	43,092	25,190	0.07	0.04	4.63	4.71
Long Beach, CA	35,524	54,085	0.06	0.09	3.72	3.76
Newark, NJ	51,425	56,963	0.08	0.09	3.65	3.77
Savannah, GA	20,803	20,049	0.03	0.03	2.20	2.22
Seattle/Tacoma, WA	14,726	14,398	0.02	0.02	1.49	1.37

Source: S&P Global Market Intelligence, Haver Analytics, U.S. Census, and authors' calculations.

Notes: Panjiva aggregates for Seattle/Tacoma and Houston exclude shipments where the consignee country is missing.

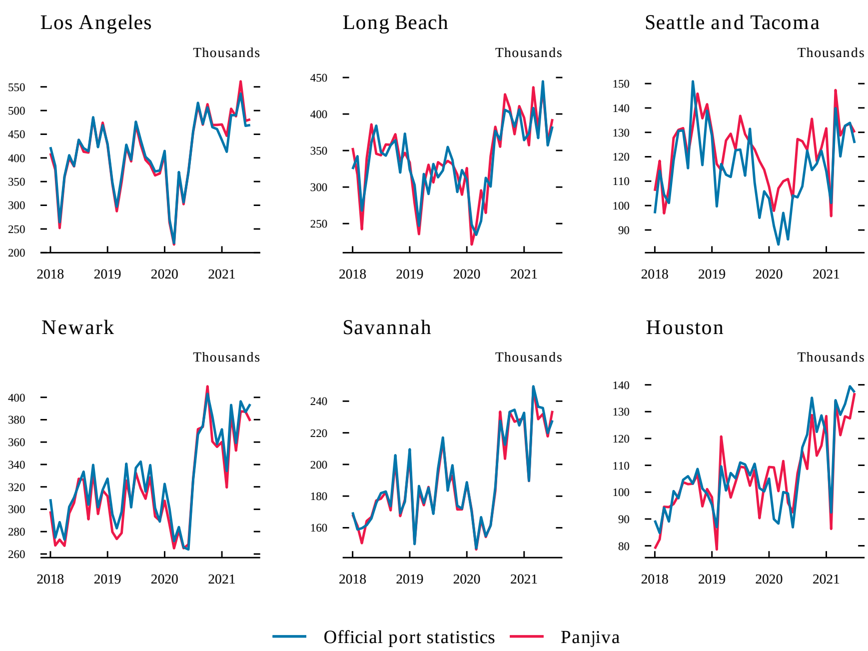
*In millions.

A.2 Comparing containers by port over time

Next, we compare the number of imported TEUs reported by Panjiva to the volumes reported by ports. In particular, Figure 13 displays monthly Panjiva and official imported TEU volumes for the top six U.S. ports. Importantly, both sources tend to give similar signals for the level and changes in trade from month to month.

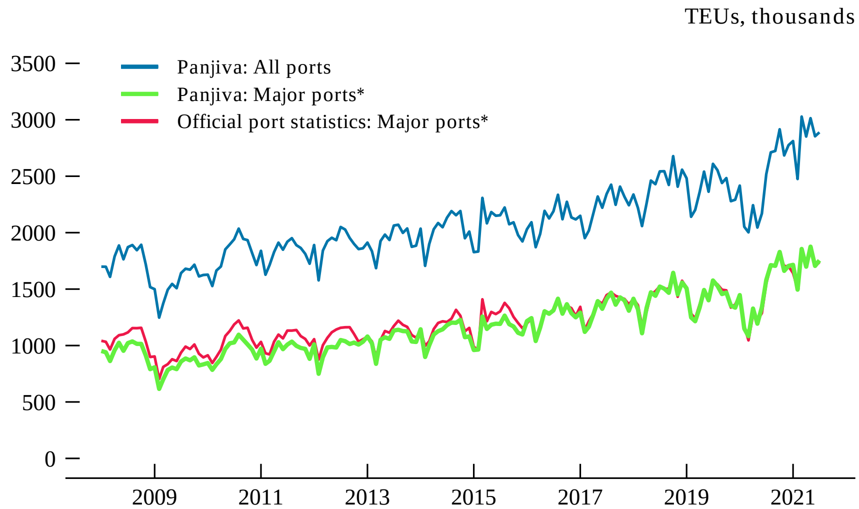
In terms of timeliness of data reporting, the official data on container volumes by port are available from Haver with a lag of about 3 weeks on average, while data are available from Panjiva with a lag of only about 7-14 days. While this improvement in timeliness from Panjiva data is relatively modest, it may nonetheless be valuable during times when shipping is being interrupted, such as during the COVID-19-related plunge and the backups at West Coast ports during the subsequent recovery.

Figure 13: Comparison of Panjiva data and official port statistics by port



Source: S&P Global Market Intelligence, individual ports via Haver Analytics, and authors' calculations.
Notes: Panjiva aggregates for Seattle/Tacoma and Houston exclude shipments where the consignee country is missing.

Figure 14: Comparison of Panjiva data and official port statistics



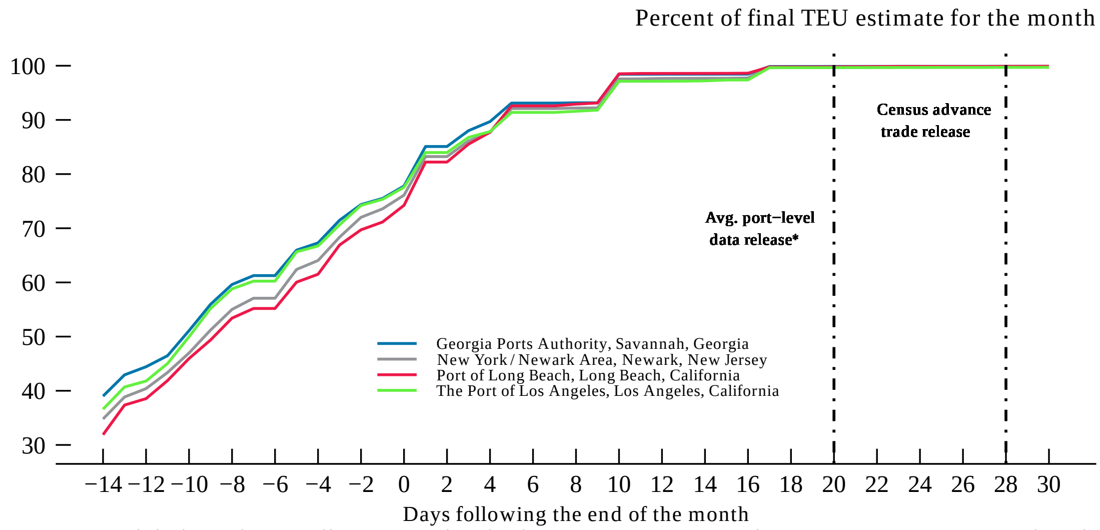
Source: S&P Global Market Intelligence, individual ports via Haver Analytics, and authors' calculations.

Notes: *Major ports include Houston, Long Beach, Los Angeles, Savannah, Seattle/Tacoma, New York, and Newark.

A.3 Comparing lags in data reporting

Figure 15 illustrates the timeliness of when data are available for a given month. As shown in the figure, data are updated continuously with roughly three-quarters of a given month's final TEU value in by the end of the month. The data then reach close to 100 percent of the final monthly value by around 7 to 14 days after the end of the month. This reporting is sooner than the port-level reporting and significantly sooner than the Advance Economic Indicators trade report released by the U.S. Census Bureau.

Figure 15: Panjiva BoL data completeness



Source: S&P Global Market Intelligence, individual ports via Haver Analytics, Census Bureau, and authors' calculations.

Notes: 100 percent reflects the "final" level of TEUs estimated for a given month.

A.4 Comparing firm-level trading information

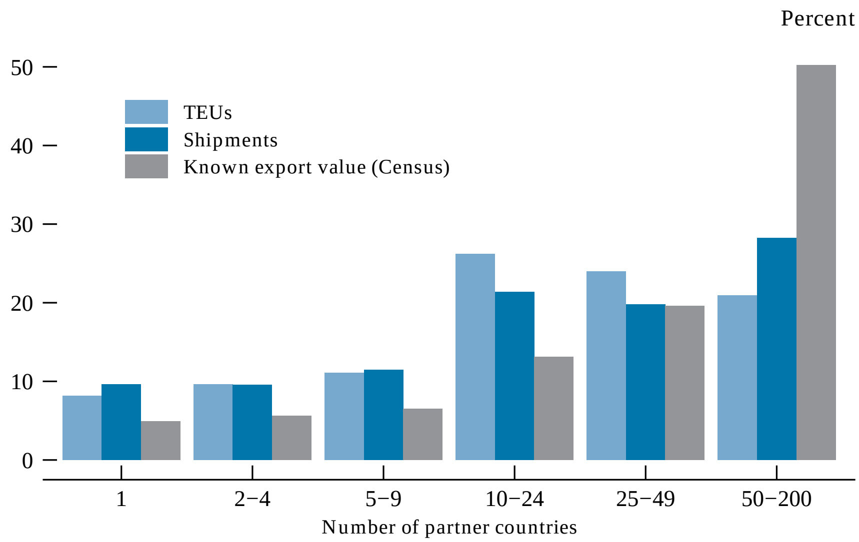
As discussed above, one of the key benefits of the BoL data, relative to public data sources, is the availability of firm identifiers for most transactions. Comparing firm-level information from BoL data to similar information in other datasets, such as the Census Bureau's Longitudinal Foreign Trade Transaction Database (LFTTD), is difficult given the confidentiality associated with official statistical datasets. Nonetheless, the Census Bureau does publish some information on characteristics of firms engaged in international trade, which can be compared to BoL sources.

One piece of information about trading firms that the Census Bureau reports is a histogram of the value of trade by the number of destination countries for each exporting firm (See top chart on page 3, Census Bureau 2020). In Figures 16 and 17, we display similar figures based on Panjiva data for both exporters and importers, respectively, though our histograms are in terms of the number of TEUs and shipments. Our figures include all firms and are therefore most comparable to the blue bars in the histogram provided by the Census Bureau.

Figure 16, for exports, shows a rightward skew of the distributions for TEUs and shipments based on BoL data, indicating the importance of firms that export to many countries in overall trade volumes. This rightward skew is consistent with, but actually somewhat less pronounced than that reported for the value of exports in Census Bureau (2020), which is reproduced in the gray bars of the Figure.

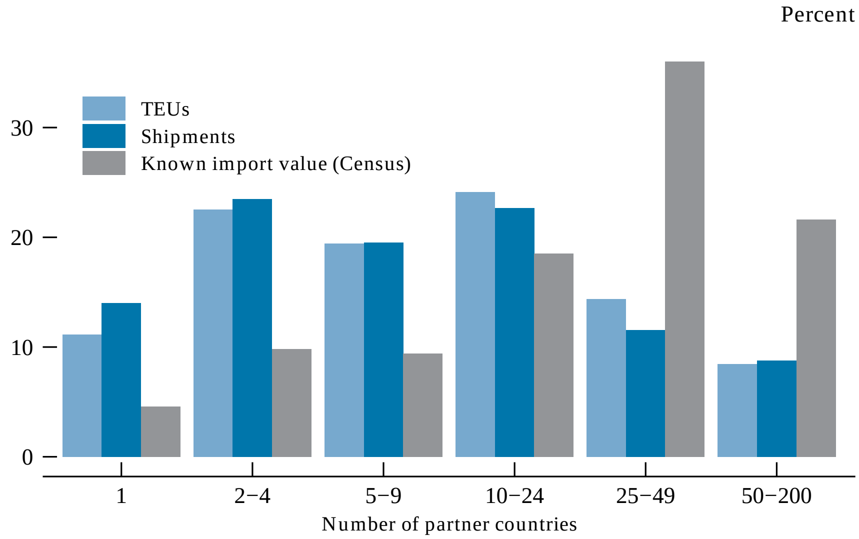
Figure 17 indicates that, in contrast to exports, firms that import from a small number of destinations account fora relatively larger share of U.S. import volumes. This difference may be indicative of smaller fixed costs associated with importing, relative to exporting.

Figure 16: Percent of TEUs, shipments, and value by number of partner countries for U.S. exports, 2018



Source: S&P Global Market Intelligence, U.S. Census, and authors' calculations.

Figure 17: Percent of TEUs, shipments, and value by number of partner countries for U.S. imports, 2018



Source: S&P Global Market Intelligence, U.S. Census, and authors' calculations.

A.5 Bill of Lading Forms for U.S. Imports and Exports

Figure 18: Customs and Border Protection Bill of Lading Form for U.S. Imports

DEPARTMENT OF HOMELAND SECURITY
U.S. Customs and Border Protection

OMB CONTROL NUMBER 1651-0001
EXPIRATION DATE: 03-03-2021

INWARD CARGO DECLARATION

19 CFR 4.7, 4.7a, 4.8, 4.33, 4.34, 4.38, 4.84, 4.85, 4.86, 4.91, 4.93, 4.99

1. Name of Vessel	2. Nationality of Ship	3. IMO No.	4. Voyage No.	Page No. of

5. Name of Master	6. Last Foreign Port Before U.S.	7. Port of Discharge	8. Date of Departure from Port of Loading	9. Time of Departure from Port of Loading (Zulu)

10. Shipper (SH) Consignee (CO) Notify address (NF)	11. Bill of Lading No.	12. Marks & Nos. (MN) Container Nos. (CN) Seal Nos. (SN)	13. No. & Kind of Packages Description of Goods Hazardous Materials (Must Provide UN Code)	Answer Col. 14 **OR** Col. 15		16. First Port/Place Where Carrier Takes Possession of Cargo	17. Foreign Port Where Cargo is Laden on Board
				14. Gross Wt. (lb. or kg.)	15. Measurement (per HTS)		

Add Table Row | Remove Table Row

Received by CBP (Signature):

Paperwork Reduction Act Statement: An agency may not conduct or sponsor an information collection and a person is not required to respond to this information unless it displays a current valid OMB control number and an expiration date. The control number for this collection is 1651-0001. The estimated average time to complete this application is 30 minutes. If you have any comments regarding the burden estimate you can write to U.S. Customs and Border Protection, Office of Regulations and Rulings, 799 9th Street, NW., Washington DC 20229.

CBP Form 1302 (3/20)

Figure 19: Customs and Border Protection Bill of Lading Form for U.S. Exports

DEPARTMENT OF HOMELAND SECURITY
U.S. Customs and Border Protection

OMB CONTROL NUMBER 1651-0001
EXPIRATION DATE: 03-03-2021

CARGO DECLARATION
Outward With Commercial Forms
19 CFR 4.62, 4.63, 4.75, 4.82, 4.87-4.89
(Oath to be taken on CBP Form 1300)

1. Name of Ship	2. Port where report is made (not required by United States)	3. Nationality of ship
4. Name of master	5a. Port of loading	5b. Port of discharge

B/L Nr.	6. Marks and Nrs. (MN) Container Nrs. (CN) Seal Nrs. (SN)	7. Number and kind of packages; Description of goods	8. Gross Weight (lb. or kg.)	9. Measurement (per HTSUS)

Add Row | Delete Row

Received by CBP (Signature):

This form may be printed by private parties provided it conforms to the official form in size, wording, arrangement style and size of type, and quality and color of paper. (19 CFR 4.99)

Paperwork Reduction Act Statement: An agency may not conduct or sponsor an information collection and a person is not required to respond to this information unless it displays a current valid OMB control number and an expiration date. The control number for this collection is 1651-0001. The estimated average time to complete this application is 30 minutes. If you have any comments regarding the burden estimate you can write to U.S. Customs and Border Protection, Office of Regulations and Rulings, 799 9th Street, NW., Washington DC 20229.

B Handling the Panjiva data feed

While the storage and analysis of data is typically a relatively minor concern in economics research, these issues require extensive consideration when handling massive datasets, such as Panjiva's BoL data. In this section, we provide detailed discussion of the computing solutions we used to effectively make use of these data. Our hope is that this information will assist other researchers as the use of BoL data becomes more widespread.

Scalable programming tools are critical to effectively ingesting and analyzing this dataset. While there are a variety of potential big data solutions to handle 100 GB of data, our research solution balanced performance and usability of the data. Harnessing primarily open source tools from Apache and the Python Software Foundation, we loaded the Panjiva data into a cluster Hadoop environment to provide scalable data storage and processing.

Panjiva provides researchers access to the underlying data through an FTP server that hosts the raw files in a zip format. Once raw files are downloaded, they are decompressed and converted out of their "phrase-separated" values file into a more useful format for querying. Panjiva's file format uses a non-standard characters to separate records and fields, which can cause performance bottlenecks. The files are large enough to warrant

Table 7: U.S. import data description for remaining variables

Variable name	Description
billofladingnumber	Bill of lading for shipment
billofladingtype	Types of bills of lading: House, Simple or Master designation
carrier	Name of the company that transports the goods
concity	City of the consignee's domestic location
concountry	Country of the consignee's domestic location
confulladdress	Full address of the consignee's location
conoriginalformat	The party to take final delivery of the merchandise (original format)
conpostalcode	Postal code of the consignee's domestic location
conroute	Street address of the consignee's domestic location
constateregion	State/region of the consignee's domestic location
containermarks	Symbols printed on boxes/crates to determine how to handle shipment
containermarksid	Symbols printed on boxes/crates to determine how to handle shipment
containernumbers	Container identification numbers
containernumbersid	Container identification numbers
containertypeofservice	Indicates the type of service provided for the customer
containertypeofserviceid	Indicates the type of service provided for the customer
containertypes	Indicates the type of container used in the shipment
containertypesid	Indicates the type of container used in the shipment
dangerousgoods	Substances or materials that pose unreasonable risk to health and safety
dangerousgoodsid	Dangerous goods ID
dividedLCL	Indicates whether shipments are combined with other shipments
dividedLCLid	Indicates whether shipments are combined with other shipments
filedate	Represents the date that the data was publicly available
FROB	Foreign cargo remaining on board
goodsshipped	Free text description of the product
goodsshippedid	Unique ID for records within goodsshipped tables
hasLCL	Denotes whether the shipment has consolidated cargo
hscodeid	Harmonized Item Description and Coding System (HS)
inbondcode	Indicates whether the shipment is In-Bond or not In-Bond
iscontainerized	Indicates whether a shipment was containerized (Panjiva derived)
manifestnumber	Identification number of manifest on which goods were listed
masterbillofladingnumber	Identification number of the master bill of lading
measurement	Additional description of measurement used on the shipment
notifyparty	Name of notify party
notifypartySCAC	Standard Carrier Alpha Code (SCAC) for notify party
numberofcontainers	Total number of containers in the shipment
placeofreceipt	Location where the goods were received for transport to the vessel
portofladingcountry	Country of port of lading
portofladingregion	Region of port of lading
portofunladingregion	Region of port of unlading
quantity	Quantity of items in the shipment
shpcity	City in which the exporter is located
shpcountry	Country of shipper
shpfulladdress	Full address of shipper
shpmtdestination	Country of shipment destination
shpmtdestinationregion	US geographic region of the final destination of the goods
shporiginalformat	Name of the shipper (original format)
shppostalcode	Entity resolved postal code of the shipper's domestic location
shproute	Street address of the shipper's domestic location (Panjiva derived)
shpstateregion	State/region of the shipper's domestic location (Panjiva derived)
transportmethod	Mode of transportation
vesselvoyageid	Voyage ID for vessel carrying shipment
volumecontainerTEU	Volume of container in TEU
volumecontainerTEUid	Volume of container in TEU
weightoriginalformat	Shipment weight as originally reported on shipment record
weightT	Shipment weight in metric tons

parallelization though their format of the files forces an initial single core processing bottleneck. Parallelization is required to effectively ingest these files since a single file can contain tens to hundreds of millions of records. Using the python packages Dask and Pandas, the data are saved into a column-oriented storage object called Apache Parquet. This file type is popular among big data experts due to its effective compression and querying performance. The structure of Panjiva updates involves large snapshot files and smaller modification files that could include updates, additions, or deletions to the primary data.

These optimized data files are partitioned based on the structure provided by Panjiva. For U.S. Imports data, records are separated into four blocks by arrival date: 2007-2009, 2010-2014, 2015-2019, 2020-2024. Partitioning the data improves query performance on the cluster by minimizing the number of files being scanned when requesting time-based subsets of the data.

The data files are structured into a partitioned Apache Hive tables with the added capabilities of Apache Impala. We then utilize either PySpark or SQL protocols to query specific data subsets from the hundreds of millions of records and produce basic summary statistics of the data.

C Data Cleaning and Vessel Standardization

For the port analysis exercise, we pull all shipment-level data from 2012 to present, resulting in approximately six million observations. As with most "big data" datasets, spelling errors and name variations are widespread. Over two-and-a-half million observations are missing a vessel International Maritime Organization identifier (IMO). We remove obvious duplicates, standardize the vessel names, and impute the missing vessel IMOs.

First, we generate a crosswalk using a subset of the data that contains a vessel IMO. We clean the vessel names by removing any non-alphanumeric characters and removing any trailing or leading spaces. Next, we collapse our data by vessel name and vessel IMO – of the 15,000 unique vessel IMOs in our known IMO dataset, over 2,000 are associated with more than one vessel name. Because of this, we create a "primary" vessel name based on the name which is most commonly associated with each individual vessel IMO. We then merge back the standardized crosswalk of known IMOs on the full dataset, resulting in almost 3.5 million observations to be standardized.

Next, we aim to add in the missing vessel IMOs of over 2.5 million observations. This process is similar to our standardization process, except we keep all observations with an unknown vessel IMO. We then clean the vessel names in the same manner (remove trailing and leading spaces and non-alphanumeric characters). Next, we merge these onto the de-duplicated known vessel IMO and standardized vessel name crosswalk such that only unique names in the unknown IMO dataset will merge to unique names in the known IMO crosswalk. Over 2 million observations merged at this point, leaving roughly half a million remaining.

The rest of the data cleaning is an iterative process. After skimming through the unmatched observations, we find that almost 8,000 observations actually have the vessel IMO in the vessel name column. We also find that of the unknown vessel IMO observations, some observations will likely remain unidentified as they have meaningless vessel names; we choose to drop names that are a random mix of numbers and letters and are also too short (less than seven characters) to accurately identify.

At this stage, we have less than half-a-million observations without a vessel IMO to identify. Using reclink2, a natural language processing package ((Wasi and Flaaen 2015)), we attempt to merge the remaining unknowns to our known dataset. We manually go through 2,000 unique name matches resulting from the fuzzy merge. The fuzzy merge provides a "score" of how accurate the unidentified string is to a known IMO string vessel name. We tag all matches over 99 percent accuracy and generate a crosswalk from the fuzzy merge which yields us an additional 50,000 observations.

Finally, we manually search for the vessel IMOs based on the vessel name and route (if there are multiple IMOs associated with a given vessel) of the largest number of unidentified vessels. Merging these imputed vessel IMOs results in almost 100,000 identified vessels. After these iterative processes, our data cleaning and standardization results in over 5.6 million of the 6 million observations.