Related-Party Trades in Vertical Integration

YongKi Hong[*]

April 12, 2023

Abstract

Despite the importance of intra-firm trades in theories of the firm, an empirical literature using proxy measures has documented surprisingly little such trade. I revisit this conclusion using economy-wide firm-level data from Korea, where related-party trades are directly observable. I show that the true prevalence and volume of the trades are much greater than previous measures indicate. Past proxy measures that rely heavily on input-output tables appear to dramatically underestimate the trades, capturing only 17.6% of related parties that trade and 32.6% of their sales volume. I propose alternative methods to infer trade within ownership networks that show substantially improved performances.

[*] hyk0127@gmail.com. I am extremely grateful for the advice and support of Jonathan Vogel, Simon Board, Pablo Fajgelbaum, and Martin Hackmann. I also thank John Asker, Oleg Itskhoki, Brian Pustilnik, Ivan Lavrov, Conor Foley, Luke Zhang, and participants at various virtual and in-person seminars for helpful comments and feedback. I thank Kiho Dong, C.P.A., for answering countless questions about rules and practices around related-party trades and their reporting. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Institute for Digital Research and Educations Research Technology Group. All errors are my own.

1 Introduction

An extensive literature on the theory of the firm focuses on why some trades are moderated within the boundary of a firm whereas others occur at arm's length. Transaction cost economics (Williamson, 1971, 1979; Klein, Crawford and Alchian, 1978) and the property rights approach (Grossman and Hart, 1986; Hart and Moore, 1990) are particularly influential perspectives on this issue. Both theories build on the premise that substantial vertical trades occur within entities.

This fundamental hypothesis has been infrequently tested; when tested, it has not fared well with data. Information on sales and purchases of inputs within-firm, or with related parties of a firm,[1] are not generally available. Hence, large-sample empirical works on the topic are scarce and are either reliant on proxy data that require strong assumptions or are based only on a specific, small portion of an economy.[2] Moreover, these empirical studies suggest that related-party trades are small and sparse, in a critical divergence from the theoretical literature.

Atalay, Hortaçsu and Syverson (2014)—henceforth AHS—were the first to empirically test this issue with a large sample, and this seminal paper has served as a benchmark for subsequent researches. In the absence of data on intra-firm trades, AHS construct a novel proxy from U.S. establishments' shipments data in the Commodity Flow Survey (CFS), using geographical information as well as industry-level proxies. Their baseline results show that almost 50% of establishments with at least one related party do not sell anything to them; and even when they do, the sales comprise only a small portion of the sellers' economic activities. This was a surprising conclusion: if there is little vertical trade, how can facilitating trade be a central driver of integration? How can vertical integration enhance efficiency?[3]

In this paper, I present the first economy-wide direct measurement of related-party trades and show that there is, in fact, substantial trade within integrated firms. I construct a novel firm-level dataset that enables this direct observation. Exploiting a South Korean accounting requirement, I web-scraped firms' annual trades with each of their related parties from the side notes of *all* publicly available financial statements in Korea—similar to 10-K reports in the U.S.—between 2013–2019. As firms are explicitly requested to report related-party transactions, this data shows the trades for all firms in the economy above a set of size thresholds,[4] without having to rely on a proxy measure.

[1] Related parties refer to entities connected through a sufficient amount of control or ownership, such as parent companies, subsidiaries, and so on. Exact definitions differ by study and dataset; see section 2 for more detail.

[2] Ramondo, Rappoport and Ruhl (2016) and Nunn and Trefler (2013) are among the strand of the literature that studies U.S. firms' trades with *foreign* subsidiaries by utilizing customs data.

[3] This discussion has been influential in economic policymaking. In September 2021, the Federal Trade Commission (FTC) issued a statement withdrawing support for the Vertical Merger Guidelines issued jointly with the Department of Justice in 2020. In the statement, the FTC cites AHS to argue that the guideline puts too much emphasis on the pro-competitive effects and efficiency gains from vertical mergers, saying: "we should be highly skeptical that EDM [Elimination of Double Marginalization] will even be realized" as "[in] many cases, vertical integration does not even prompt firms to provide the upstream input to its own downstream division." (FTC, 2021)

[4] One representative threshold is the firm's total sales surpassing roughly $8.3 million U.S. dollars; see Section 2.

I find that firms utilize related-party trades–in terms of both prevalence and value–substantially more than the previous literature has documented. Almost all manufacturing firms with a related party appear to engage in related-party trades: 87.2% of the firms report either sales to or purchases from a related party, and 77.3% report related-party sales during a fiscal year. What is more, the trades in my data are a considerably larger part of firms' activities than past estimates. These results imply that vertically integrated firms actively utilize related-party trades, consistent with a focus of the vast theoretical literature.

Subsequently, I explore why related-party trades are substantially larger in my data compared to previous studies. I show that rather than the simple difference of country, the main driver of the different results appears to be the data that is better suited to address the question. Specifically, I construct a dataset where related-party trades are directly observed, whereas the lack of such measurement has compelled previous works to infer them using proxy measures. In most cases, researchers observe only either relatedness or trades: that is, they observe trades but not whether the two sides are related, or see that two entities are related but not whether they trade.[5]

To infer internal sourcing without direct observation, researchers have long depended on a combination of (i) industry-level trade patterns represented by Input-Output Tables (IOT) and (ii) entity ownership; see, e.g., Alfaro et al. (2019), Atalay et al. (2019), Acemoglu, Johnson and Mitton (2009), and Aghion, Griffith and Howitt (2006). For example, consider two entities, A_i and A_j, with a common owner and belonging to industries i and j, respectively. The common approach regards A_i to be purchasing from A_j only if industry i uses the output from industry j more than some arbitrary threshold according to IOT. Unfortunately, this approach may not provide a close approximation for the presence of trades between A_i and A_j. First, while there may exist some IOT threshold above which A_i is likely to buy from A_j, we require data on intra-party trade to know the value of this threshold. More generally, related-party trade patterns may differ fundamentally from the general economy upon which the IOTs are built.

Indeed, I show that when compared with the directly observed data, the proxies' accuracy is heavily sensitive to the choice of cutoffs, and is generally low. A common coefficient cutoff of 1% (e.g., AHS, Aghion, Griffith and Howitt (2006)) captures only 17.6% of related-party pairs that trade in the Korean data, and less than one-third of the sales volume. In fact, using Korean data together with the 1% cutoff yields a result that remarkably resembles AHS's results. Similarly, AHS's robustness check that drops the IOT requirement and only utilizes geographical information yields results that are remarkably similar to those obtained from the true Korean data. Yet another common cutoff used in the literature of having a positive total requirements coefficient (e.g., Alfaro et al. (2019)) is, in turn, too lenient and results in assuming almost all of the related

[5] In a concurrent work, Garg, Tan and Ghosh (2021) also directly measure related-party trades by utilizing a regional dataset from a state of Karnataka in India and find larger trades compared to AHS. Relative to Garg, Tan and Ghosh (2021), the dataset utilized in this paper can represent an entire economy that is more advanced and has strong contractual enforcement, visualize international related-party trades in addition to domestic, and include the flow of services as well as physical goods. Furthermore, this paper utilizes the data to construct more accurate predictive algorithms that allow future work without access to intra-firm trade data to better predict these trades.

parties are trading.

These results suggest that the IOT-based proxies for related-party trades, when used on their own, appear to have generated incorrect conclusions and an inconsistency between theory and data.[6] Yet these same proxies have been central in addressing a range of questions. For example, the proxies have been used to discern production chains within integrated firms and subsequently answer questions such as why firms only integrate specific parts of the production chain (Alfaro et al., 2019), what factors induce more vertical integration (Acemoglu, Johnson and Mitton, 2009; Blyde and Molina, 2015; Alfaro et al., 2016), and also to separate out vertical from horizontal FDI (Fajgelbaum, Grossman and Helpman, 2015; Alfaro and Charlton, 2009). Given the importance of these questions and the sensitivity of inferring related-party trades on the assumptions used, there is a strong need for a better way of inferring trades within ownership structures.

In this vein, I propose alternative proxy measures that provide a more accurate inference of related-party trades. Here, I exploit the unique opportunity to utilize the true data on related-party trades to test and compare each method's predictive performance. Using supervised machine learning, I build prediction mechanisms that enable researchers to infer the existence of related-party trades with only widely available information on firms—sizes, countries, and industries—and the ownership links between them. Despite requiring only a small amount of additional information other than the IOT coefficients, the proposed measures greatly improve performance metrics such as accuracy, precision, and recall.

The rest of the paper proceeds as follows. Section 2 discusses the original data collection and contents. Section 3 presents the related-party trade of Korean firms and discusses the underestimation problem of the proxy measure. Section 4 presents robustness checks, and section 5 proposes alternative proxy measures for related-party trades. Section 6 concludes.

2 Data

This paper draws on several sources to construct data. I first collect and construct the key dataset of firm-to-firm related-party trades, then combine it with existing firm-level databases to match firm characteristics to both sides of the trades. Here, I describe the data collection and matching.

Related-party trade data. — The primary strengths of the data used in this paper are that related-party trades are observed directly, for a large sample of firms, and with high credibility. In most cases, researchers either observe trades without knowing whether the trade partners are related or observe related firms but not whether they trade. This is the first economy-wide dataset that incorporates both components at the same time. Moreover, the reports undergo external audits and government scrutiny, ensuring the results' accuracy.

[6] Ramondo, Rappoport and Ruhl (2016) also notes that the IOT coefficients are not correlated with the observed U.S. parent companies' trades with foreign subsidiaries in their data.

A South Korean accounting requirement enables this direct observation of related-party trades.[7] All Korean firms satisfying a set of size thresholds (henceforth *reporting firms*) are required to get an annual external audit and make the reports—analogous to 10-K reports in the U.S.—publicly accessible.[8] Crucially to this paper, in the side notes of the reports, the firms need to disclose annual trades with each of their related parties. This section offers a brief overview of the terms and definitions. Details can be found in Appendix A and are based on Korean financial reporting standards (1024, 1028, 1110 K-IFRS).

I scraped and cleaned the filings from an official website maintained by a Korean government agency (*Financial Supervisory Service*), comparable to the Electronic Data Gathering, Analysis, and Retrieval database maintained by the U.S. Securities and Exchange Commission.[9] I examine all 190,725 available financial reports from FY 2013–2019: after ruling out reports that were found to be inadequate by external auditors,[10] or had issues in scraping, 176,657 (92.6%) reports from 44,701 firms are used in the analysis. The main analysis utilizes only the firm-years where the reporting firms are confirmed to have had at least one related party. After the exclusion of singleton firms, 121,519 annual reports from 30,390 firms are used, amounting to 558,114 firm-year-related party triples.[11] While the information in this section of the financial reports has been used in previous research in small batches through hand-scraping, this is the first paper that utilizes it in its full scope and detail.

A related party is defined primarily by the voting rights that one firm possesses over the other. That is, the definition of related party in this paper is based on *control*, and not mechanically on the share of outstanding stocks owned. For example, if firm A owns a right to name the entire board of directors of firm B, A and B are deemed related parties even if A owns only a small share of B's outstanding stocks.

Related parties in this data are composed of two categories. The first category of related parties is within a network connected from the reporting firm with links of controlling voting power. The related party does not have to be a direct subsidiary or a parent company. Regardless of the number of ownership links between the reporting firm and the related party, trades between the

[7] The International Financial Reporting Standards (IFRS), which Korea has adopted, requires related-party transactions to be reported (IAS 24). However, partly due to the complexity of creating a dataset from document-based information, they have not received much academic attention from economists. Santioni, Schiantarelli and Strahan (2020), which analyzes Italian firms' related-party loans and debts, is a good example of the recent efforts to utilize this information.

[8] In the last year of the data, the fiscal year 2019, the threshold requires firms to satisfy at least two of the following: : (1) sales $\geq$ 8.3 million USD (originally 10 billion Korean Won), (2) assets $\geq$ 10m USD (12b KRW), (3) debts $\geq$ 5.8m USD (7b KRW), or (4) more than 100 employees. The thresholds vary slightly over the data period; details can be found in Appendix A.

[9] While 10-K filings of U.S. firms also include information regarding related-party transactions, the Korean filings offer much more detail and a steep advantage for digitizing as they are primarily filed in table formats, not in sentences.

[10] Specifically, reports are excluded if external auditors issued a *disclaimer of opinion* as they could not obtain sufficient audit evidence from the company.

[11] Appendix A.5 ascertains the validity of the scraped data by comparing it with existing databases that report similar information but with limited scopes or higher levels of aggregation.

two are reported as long as each link satisfies one side having a controlling voting right over the other. Second, a firm is also deemed a related party if it is *directly* linked to any entity in the first category with 20% or more voting power. In a rough summary, for two firms to be deemed a related party, up to one link between firms of 20% voting power is allowed, as long as all other links are above 50%.[12]

The financial reports disclose the related parties' names and the reporting firm's relationship with them. The relationships are reported in discrete categories, with information on (1) whether a link with less than 50% voting rights is included and (2) which direction the control runs: in other words, whether the trade partner is a subsidiary, a parent, or whether there is a lateral component in the direction of ownership—as in 'sibling' or 'uncle' firms.

Lastly, firms report transactions separately for each of their related parties. I process the transactions into four categories—sales, purchases, loans, and debts.[13]

The financial statements, including the side notes, go through strong government scrutiny on top of external audits. An intentional misreporting of the information can lead to strong punishments: in a widely reported case in 2018, public stock trading of *Samsung Biologics* was suspended as the firm had allegedly intentionally misreported the firm's relationship with a related party.[14]

Firm Information and Matching. — Aside from the transactions and relationships with related parties, all firm-level information is matched to the scraped dataset from three databases: the ORBIS database by Bureau van Dijk and the two largest firm-level databases in Korea, KISVALUE and TS2000. The linked firm-level variables include industry, location, and financial information such as total assets and sales. Each reporting firm is matched to the databases using 10-digit firm identifiers (*Business Registration Numbers*) assigned by tax authorities. The related parties of the reporting firms are matched by name, as no other identifying information is provided on the reports. Extensive checks on the matches are undertaken to ensure accuracy.[15] Some are inevitably unmatched to firm-level information, but these are not vital to the main results, as shown in the robustness check in Section 4.

Importantly, the ORBIS database is used to discern the set of reporting firms that have at least one related party. This paper excludes singleton firms from the main analysis as they inherently cannot engage in related-party trades. However, a financial report does not provide the firm's

[12] Researchers use different levels of control or shareholding to define related parties, depending on available data. Majority shareholding is common (AHS, Ramondo, Rappoport and Ruhl (2016)) but lower thresholds are also utilized, especially with the customs data (Ruhl, 2015).

[13] Firms often report more detailed categories, which I re-summarize into the four during data cleaning. See appendix A for further detail.

[14] WSJ, Jeong and Martin (2018), "South Korea Regulator Says Samsung BioLogics Violated Accounting Rules" (https://www.wsj.com/articles/south-korea-regulator-says-samsung-biologics-violated-accounting-rules-1531407128). The case is currently on trial (April 12, 2023).

[15] I use a matching and cleaning process akin to what Alfaro-Urena, Manelici and Vasquez (2022) use for firm-to-firm trades. For example, when the related party also files its own report, I examine the reports from both sides to verify the match. See Appendix for details.

full list of related parties, instead only displaying those that *transacted* with the reporting firm—be it in sales, purchases, loans, or debts—during the fiscal year. Hence, when relying solely on the reports, a firm that has related parties but did not actively trade with any will be marked as a singleton and excluded from the sample. I use ORBIS to supplement this, as it provides for each firm the list of all other firms that share the same *ultimate owner*.[16] Throughout this paper, I take a conservative approach and use the set of firms reported to have a related party *either* in ORBIS or in the financial reports.[17]

Input-Output Table. — This paper utilizes IOTs to evaluate the accuracy of the previous literature's proxy measures and show the extent of possible biases. I use the 2015 Bank of Korea Input-Output Table (*use table*) throughout the entire data period, as it is the only publicly available version offering the most finely disaggregated sectors (381 commodities and 278 industries). Using official concordance tables, I then link the IOT with 5-digit Korean Standard Industry Classification (KSIC-10) codes and also with 5-digit 2017 NAICS codes.[18]

3 Sales and Purchases with Related Parties

Two findings emerge from Korean firms' trades with related parties. First, most firms appear to engage in related-party trades, and the importance of these trades is much greater than what has been established in the literature. Second, the discrepancy from existing literature appears to stem from a considerable underestimation inherent in the proxy measure that has been widely used to infer related-party trades; specifically, measures based on IOT coefficients show poor performance in predicting the trades.

3.1 Usage of Related-Party Trades by Korean Firms

While the data spans all industries, the main analysis will primarily utilize reports from manufacturing firms. The reasons for this are twofold. First, it enables a clearer comparison with existing studies, a vast majority of which focus on the manufacturing sector. Second, it reduces concerns over the possible effects of transfer pricing. While transfer pricing is monitored closely for all industries in Korea,[19] manufacturing firms have even less room to maneuver as manufactured goods' fair market value is easier to calculate than for services.

[16] The ORBIS database's definition of related party differs in detail from my data based on Korean financial statements. For example, in the ORBIS database, the minimum ownership percentage criterion is 25% while it is 20% in the Korean data. While these minor differences may affect the sample of firms that I consider, their significance is expected to be very small: observations with smaller than 50% ownership share are already scarce in the ORBIS database (3.34%). While it is possible that there could be a bunching between 20% and 25% ownership shares, there is no reason to anticipate the bunching to exist or the magnitude of it to be consequential.

[17] See Appendix A.4 for further details.

[18] As is common, concordance tables are not one-to-one for a number of industries. Where one IOT code is matched to multiple industry codes, I split the IOT coefficients equally between industries.

[19] All related-party trades that show more than (i) a 5% difference from the market price *or* (ii) a $250,000 (300 million KRW) difference in total value from the fair market value are subject to punitive double taxation and these are among the major items inspected in tax audits by Korean authorities.

Table 1: Share of Sales to Related Parties

Related Party Share	Percentiles (%)				Fraction (%)		Weighted	
of Firm Total Sales	50th	75th	90th	95th	=0	$\geq$1	Mean (%)	*N*
Panel 1: Korean Firms								
Manufacturing Firms	2.7	17.8	57.4	90.6	22.7	2.1	33.6	54,042
All Firms	1.3	15.1	62.4	95.8	30.3	3.0	24.0	121,519
Panel 2: Literature								
AHS (2014) - Main Result	0.1	7.0	37.6	69.5	49.7	1.2	16	67,500
Panel 3: IOT Requirement								
Manufacturing Firms	0.8	8.7	37.4	69.7	26.6	1.1	16.9	26,077
All Firms	0.1	3.5	24.4	52.5	36.7	0.7	9.7	52,298
Panel 4: Requirements à la AHS								
Manufacturing Firms	0.2	4.7	30.9	69.0	34.3	1.2	9.9	18,194
All Firms	0.0	1.9	19.1	50.5	42.7	0.8	6.4	35,955
Panel 5: Intensive Margin Only								
Manufacturing Firms	7.0	31.7	82.0	99.4	13.5	3.8	38.1	26,232
All Firms	4.8	30.2	86.3	100.0	19.3	4.7	26.8	52,696

NOTE.—The table reports the share of firms' sales that is sold to related parties. *N* consists of the firm-years where firms have at least one related party. Due to missing firm-level information, the total *N* used on the table is slightly smaller than the 125,044 available firm-level reports. The weighted mean is weighted by the size of each reporting firm's total sales.

Panel 1 of Table 1 reports the share of related-party sales in total firm sales. Following the convention from AHS, I present quantiles of distributions of the shares. I first note that most firms report positive quantities of related-party trades. 77.3% of manufacturing firms report selling to a related party during the fiscal year, and in fact, 87.2% of the firms have either sales to or purchases from related parties.

Moreover, the sizes of the related-party sales appear substantially greater than what the literature has previously found. In Panel 2 of Table 1, the findings of AHS are presented as a benchmark. The results from Korean manufacturing firms illustrate the greater importance of related-party sales throughout the entire distribution. The 75th percentile firm in AHS sells only 7.0% of its sales to related parties, while the corresponding number is 17.8% for Korean firms.[20] In the 90th percentile entity, the gap is even larger, where the share of related-party sales is close to 60% in Korea but is less than 40% in AHS.

While the overall level of trades is higher, other general characteristics of the distribution are

[20] Note that the comparison is based on the manufacturing industries in Korea, while AHS also includes the mining, wholesale, and select retail industries included in the CFS. Section 4 shows that using the same set of industries as AHS simply generates almost identical results as Table 1.

similar to the literature (AHS, Ramondo, Rappoport and Ruhl (2016)). The internal trade share distribution is highly skewed, although to a lesser degree. Also, the share of related-party sales is positively correlated with the selling firm's size. The weighted mean of the shares of related-party sales, weighted by the total sales of firms, is significantly larger than the unweighted mean or the median. However, this paper shows much higher shares of intra-party sales overall and, importantly, a much longer left tail: a vast majority of manufacturing firms are selling to at least one of their related parties, even for only small amounts.

3.2 Underestimation Problem in Using Input-Output Tables to Infer Vertical Trading between Related Parties

If the present Korean data show such a consistent and significant difference from the existing literature, what causes this divergence? Since each paper on the topic uses a different dataset, the apples-to-oranges problem renders a rigorous decomposition of the source of the differences difficult. In this section, I highlight a single main driver of the disparity: the low accuracy of the widely-used proxy for related-party trades. Specifically, using IOTs to construct proxies appears to be responsible for much of the inaccuracy.

Confronted with a lack of data on related-party trades, IOTs are widely used to proxy for the related parties that trade with each other. This process is rooted in Fan and Lang (2000) and utilizes firms' ownership and industry information which is often accessible to researchers. Assume firm A is in industry J and has related parties a_1, a_2, $\cdots$, a_n, each in industries j_1, j_2, $\cdots$, j_n, respectively. The commonly used method assumes that A sells to a_i only if the industries J and j_i trade *significantly* according to the IOT. The criteria for *significant* trade varies by studies. In AHS, the criteria is more than 1% of industry J's output being used as an intermediate input in industry j_i, but studies have utilized different cutoffs.

These criteria based on IOT cutoffs are then used to study vertical integration. As an example, HY, a Korean firm in the *Manufacture of dairy products and edible ice cakes* industry, has a related party in the same industry (IOT coef. 0.042), and another in the *Manufacture of truck and motor vehicles for transportation of goods and special purpose* industry (IOT coef. 0.001).[21] Based on the industry-level trading patterns, with the 1% cutoff, one would assume that HY is selling only to the first related party but not to the second.

While a paucity of true data has necessitated the use of such measures, using IOT coefficients in this context inherently entails limitations. Firstly, industry codes may not be able to reflect the firm's entire line of businesses, which is problematic especially for multiproduct firms. Secondly, while IOTs represent industries' average input requirements, how firms source from related parties may fundamentally differ from arm's length sourcing. For example, data shows that HY purchases a large amount from the related party that produces special-purpose trucks, despite

[21] The IOT coefficients refer to the proportions of HY's industry's intermediate sales that are directed to each related party's industry.

such trades being a very small share of an average dairy producer's behavior. This is because HY stands out from its peers by operating a large fleet of small refrigerating motor vehicles which work as roaming retail stores for the firm's products. As these were highly specialized vehicles that no other company was using, their production was integrated into the firm to remove holdup costs. At the same time, the specialized nature of the product dictated the IOT coefficients to be small.

Even when taking IOT coefficients as accurate reflections of related-party trades, without actual data to maximize accuracy, the traditional method is inherently simple and reliant on arbitrary components. It utilizes a single cutoff of IOT coefficients to proxy complex decisions, and more importantly, the choice of the cutoff varies widely among studies. Studies such as Aghion, Griffith and Howitt (2006) and Monarch, Park and Sivadasan (2017) echo AHS's criteria, while Alfaro and Charlton (2009), Blyde and Molina (2015), and Alfaro et al. (2019), use criteria based on total requirement coefficients, thus taking into account indirect input use by industries as well. Moreover, the specific cutoffs utilized in previous papers range widely from as small as zero or 0.0001 up to 0.05 in their baseline specifications.[22] Additionally, papers including Acemoglu, Johnson and Mitton (2009), Fort (2017), and Altomonte et al. (2021) do not utilize cutoffs, but instead use the coefficients to construct vertical integration indices for each firm.

The benchmark example, AHS, ingeniously utilizes shipments and geographical data along with the proxy method. The paper is unique in two dimensions. First, by utilizing shipment data, it can speak to the *sizes* of related-party trades, while the traditional proxies mostly concentrate on approximating the *existence* of trades between two related parties. Second, it exploits the geographical locations of entities to achieve a more accurate inference.

AHS's proxy for U.S. establishments' related-party sales is constructed with shipment data in the Commodity Flow Survey. For each establishment in the survey, CFS samples shipments and records their counts, values, and destination zip codes. However, whether the shipment is internal to the same firm is not observable. The paper, therefore, classifies a shipment as related-party sales only if it is sent to zip codes where (i) a related party of the sending establishment is located *and* (ii) that related party is in an industry that uses more than 1% of the sender's industry's output, according to IOT.[23] In a way, this method treats a group of entities as possible participants of related-party trades by first using the IOT cutoff. Then, it utilizes geographical information to distinguish the shipments that are likely sent to the group.

To demonstrate the extent of limitations that stem from utilizing the IOT cutoffs, I re-create the main results after imposing the same data-generation process used in the previous literature.

[22] Papers often provide robustness checks where they use different cutoffs and show that the qualitative results do not change. However, in a more general context, different cutoffs produce widely different predictions of which related parties are trading. Table 3 shows how two of the most popular cutoffs perform in the present Korean related-party trade data.

[23] AHS recognize that retail and wholesale industries are not represented well in the IOT, and utilizes other supplemental datasets. Section 4 shows that this limitation of the IOT does not affect this paper's results.

Panel 3 of Table 1 reports the results from the Korean data when counting only the sales to the related parties that satisfy the IOT cutoffs (1%).[24] It is immediately apparent that the limitations of the proxy measures take a substantial toll: the results are now much more similar to AHS, where the use of related-party sales is almost identical to the existing literature for the 90th and 95th percentile firms. In fact, only 17.6% of reporting firm–related party pairs identified in the full Korean dataset satisfy the IOT cutoff, representing 32.6% of the total value of trade in the data.

Panel 4 constructs a more complete comparison with AHS by imposing two additional limitations that pertain to the paper's data. First, as the CFS only provides detailed destination information for domestic sales, AHS only utilizes sales to domestic related parties. Also, only the related parties connected through majority shareholdings are included, while the present study's Korean data also includes a small group of related parties linked with minority voting power, as described in section 2. The two additional limitations produce smaller, but still significant changes to the result. The results of Panel 4 now look surprisingly similar to AHS. In this way, the restrictions in the customary data generation process can influence how the true state of vertical trades is perceived.

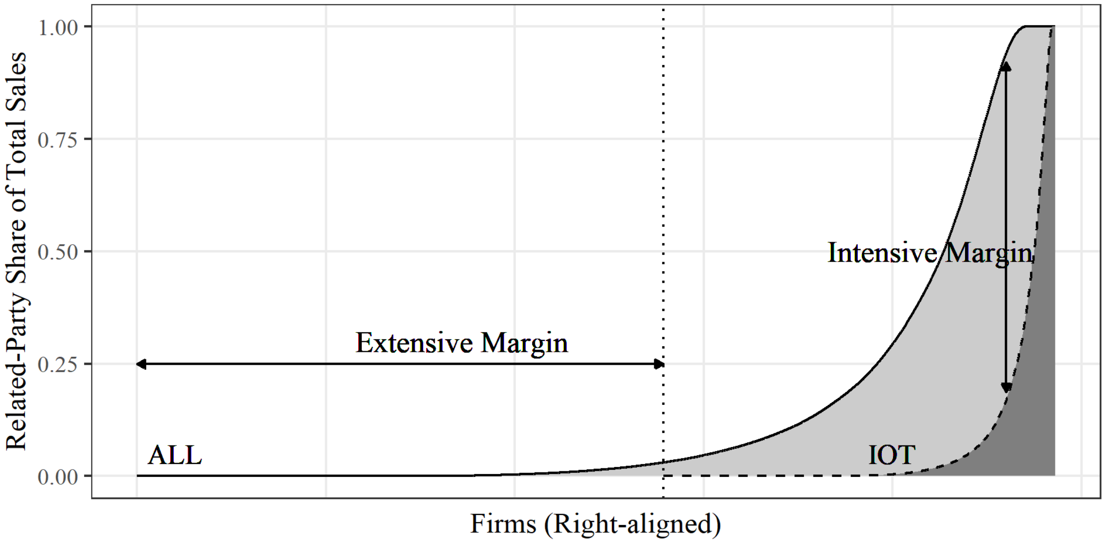
The effect of the customary data-generation process can be broken down into extensive and intensive margins. On the one hand, utilizing the 1% IOT cutoff with a firm's related parties will result in underestimating the firm's related-party trades, as only the trades with its related parties above the cutoff will be counted (*intensive margin*). At the same time, researchers commonly limit the target of their analyses to firms that have at least one related party. This means that the 1% cutoff also reduces the total sample of firms considered by reducing the scope of *firms with at least one related party* (*extensive margin*). The firms that *only* have related parties in industries below the 1% cutoff would have been excluded from the sample altogether, deemed as having 'no (tradeable) related party.'

The extensive margin stands out when comparing the sample size between Panel 1 of Table 1 with Panels 3 and 4. While 54,042 manufacturing firm-year pairs have a related party during the year, only 26,077 (48.2%) are confirmed to have at least one related party in an industry that satisfies the IOT coefficient requirement. Moreover, related-party trades of only 18,194 (33.7%) firm-years are confirmed to remain after imposing all requirements in AHS. That is, the extensive margin limits researchers' focus to a much smaller set of firms and their activities.

Lastly, the breakdown implies that the intensive margin is much larger than what is visible from a simple comparison. Panel 5 of Table 1 demonstrates the limitations of IOT cutoffs when only

[24] In Panels 3 and 4 of Table 1, if a firm is selling to related parties that are missing industry information, I assume that a fraction of the firm's sales—the 4-digit reporting firm industry's average share of sales to related parties—are to related parties that satisfy IOT coefficient cutoffs. This is a conservative assumption in the sense that it would make the effect of additionally required restrictions (IOT requirement, and the extra restrictions in AHS) appear smaller. Section 4 shows that alternative treatments of related parties without industry information make only marginal differences in the results.

Figure 1: Comparison of sample limitation on intensive and extensive margins



Note.—"ALL" refers to the distribution of firms' true share of total sales that is directed to related parties using all firms in the Korean data. Each horizontal location is a reporting firm, and the vertical location shows the related-party share of the firm's sales. "IOT" refers to a similar distribution, starting from the right, drawn with only the firms that have at least one related party satisfying the 1% IOT cutoff required in the literature (equivalent to Panel 3 of Table 1). As the extensive margin decreases the number of reporting firms in the sample, the "IOT" graph is horizontally narrower. Note that a simple overlay of the distributions is presented, so the same horizontal location does not imply the same firm.

taking the intensive margin into account. It shows what would appear when one uses only the sample of firms that have at least one related party satisfying the IOT cutoff, but measures their true magnitude of related-party trades. Now changes appear strikingly large—more than a quarter of the sampled firms sell more than 30% of their sales to related parties. This is because the now excluded set of firms, on average, have fewer related parties and therefore have a smaller share of related-party sales. That is, when analyzing a fixed number of firms (i.e., by removing the extensive margin), the limitations of the standard proxy measure will be much more pronounced than for the baseline result.

One caveat is on the treatment of related parties that could not be matched with firm-level information. The size of the extensive margin depends on the treatment of the related parties that were not successfully matched with the industry information. The main analysis reports the most conservative specification that would make the sum of the intensive and extensive margin (i.e., the difference between Panels 1 and 4 of Table 1 appear the smallest). This, however, likely overestimates the extensive margin. In Section 4, I report alternative specifications and confirm that while the relative size of the extensive margin may change, the margin remains consistently large and the qualitative results remain unchanged.

4 Robustness Checks

In this section, I present two sets of robustness checks of the results presented in Section 3. The first set of checks investigates avenues where imperfections in the present dataset could have affected the results. Several possibilities are explored: using a different data source to define the set of 'firms with at least one related party'; alternative treatment of the trades with related parties that were unsuccessfully matched to industry information; inclusion of more reporting firms in industries that are covered in CFS; lastly, an alternative treatment of retail and wholesale industries, which are represented with insufficient detail in input-output tables. The results consistently support the main conclusion, with only minor discrepancies from Table 1.

In the second set of robustness checks, I also show that analyzing specific segments of the data generates results that are in line with intuition: the results are robust to using related-party purchases as opposed to sales, and the results are also much higher for firms in industries with a prior belief of high usage of related-party trades.

Note that the results of robustness checks are presented primarily for the reporting firms in manufacturing industries for the sake of conciseness unless declared otherwise. The results from the full sample of firms in all industries are consistent with the manufacturing sample and are listed in the Appendix Table 9.

4.1 Alternative Treatments of the Data

First, I test an alternative sample selection, by using a different method to find the set of firms that have at *least one related party*. As Section 2 describes, the sample in Table 1 is a union of two groups of firms: (i) those that have related-party transactions in the data year, or (ii) the firms that have related parties in the ORBIS database. However, to the extent that only the firms with some related-party transactions are represented in group (i), the inclusion of the firms only in the group (i) may still exert upward pressure. In Panel 2 of Table 2, I only use group (ii) as the sample. The results confirm that the concern is unfounded. In fact, related-party sales appear even larger in this alternative specification compared to the benchmark results presented in Panel 1 of Table 2.[25]

The next two robustness checks examine treatments of the related parties that were not able to be matched to industry information. As detailed in Section 2, related-party trade data from Korean financial statements provide only limited information about the related parties. Information on the related parties, such as industry, total sales, and cost of goods sold are matched from multiple existing firm-level databases by the firms' names. However, some related parties are inevitably unmatched to firm-level information, which may affect the results. Specifically, the results of the exercises that impose additional constraints on the data may be impacted. For example, If a

[25] Panel 1 of Table 2 simply provides the manufacturing firms' results from Table 1 for a concise comparison with the results from robustness checks.

related party's industry is not known, then whether it satisfies an IOT cutoff is also unavailable.[26]

The main analysis in Table 1 utilizes the reporting firm's industry-level averages in place of missing values. That is, using all related parties *with* industry information, I first calculate the share s_i of related-party sales of firms within each 4-digit KSIC industry i that are directed to the related parties that satisfy the IOT cutoffs. Then if a reporting firm in i sells to a related party that is missing industry information, the share s_i of such sales is assumed to satisfy the IOT cutoffs. Assuming that the share of related party sales that satisfy the cutoffs are not systematically different between the matched and the unmatched groups, this assumption will not bias the results of Table 1.

However, this is a strong assumption. Appendix Table 10 shows that the matched and unmatched groups exhibit differences in both the sizes and types of trades with reporting firms. While there is no evidence that the differences extend to the firms' industry, a validation of this assumption is in order.

Panels 3 and 4 of Table 2 test two assumptions on opposing ends of the spectrum and confirm that the main results remain consistent in either case. Panel 2 is calculated assuming that all unmatched related parties do not satisfy the IOT cutoff. In this specification, imposing proxy-generating processes eliminate the largest share of observed related-party transactions, and therefore the common proxies' limitations appear the strongest.

In Panel 4, the opposite case is tested where all unmatched related parties are assumed to satisfy the IOT cutoff. The results do not show a large difference from the main results. This suggests that even when the IOT cutoffs are applied just to the set of related parties with known industry information, the limitations are already strong. Here, note that both the intensive and extensive margins of the restrictions are affected. The intensive margin effect is smaller than in Table 1, as the IOT cutoff now generates false negatives. On the other hand, the extensive margin is smaller as well: now, with IOT and AHS requirements, a larger number of reporting firms have 'at least one tradeable related party'. Hence, the number of observations shows a less dramatic difference between the true observation and the simulated results.

However, regardless of the specification, the key takeaways remain unchanged. First, descriptive statistics for firms' related-party trades remain unaffected from Table 1 and are much larger than the existing literature's estimates. Second, common restrictions on the data consistently assert a substantial downward pressure on estimating the related-party sales, in both intensive and extensive margins.

The fourth robustness check tests how the treatment of the retail and wholesale industries affects the results. IOTs in most cases, including the one utilized in this study, do not define these indus-

[26] Note that the limitation discussed here does not affect the calculation of Korean firms' related-party sales' share in Panel 1 of Table 1. Any possible effects are limited to the results in Panels 3 and 4 of the same table.

Table 2: Share of Sales to Related Parties, Manufacturing Firms' Reports Only

Related Party Share	Percentiles (%)				Fraction (%)		Weighted	
of Firm Total Sales	50th	75th	90th	95th	=0	≥1	Mean (%)	*N*
Panel 1: Main Result - Share at Reporting Firm Industry-Level								
All trades	2.7	17.8	57.4	90.6	22.7	2.1	33.6	54,042
IOT Requirement	0.8	8.7	37.4	69.7	26.6	1.1	16.9	26,077
AHS Requirement	0.2	4.7	30.9	69.0	34.3	1.2	9.9	18,194
Intensive Margin Only	7.0	31.7	82.0	99.4	13.5	3.8	38.1	26,232
Panel 2: All Firms with a Related Party in the ORBIS Database								
All trades	4.8	25.6	74.0	98.3	20.4	3.3	37.4	25,696
IOT Requirement	0.7	8.5	37.2	66.6	27.6	0.9	17.8	15,717
AHS Requirement	0.0	2.0	22.8	55.9	44.5	0.9	11.1	8,018
Intensive Margin Only	8.5	35.2	86.4	99.9	13.4	4.3	40.1	15,826
Panel 3: Assume No Related Party without Industry Information Vertically Related								
All trades	2.7	17.8	57.4	90.6	22.7	2.1	33.6	54,042
IOT Requirement	0.0	2.2	22.2	54.2	54.2	0.8	8.3	26,236
AHS Requirement	0.0	2.0	23.2	61.4	52.2	1.0	5.0	18,276
Intensive Margin Only	7.0	31.7	82.0	99.4	13.5	3.8	38.1	26,232
Panel 4: Assume All Related-Parties without Industry Information Vertically Related								
All trades	2.7	17.8	57.4	90.6	22.7	2.1	33.6	54,042
IOT Requirement	0.8	9.2	36.4	67.0	31.5	1.0	24.9	49,234
AHS Requirement	0.2	4.2	22.2	50.3	36.1	0.7	12.5	35,888
Intensive Margin Only	3.2	19.2	60.0	92.2	21.1	2.3	34.2	49,237
Panel 5: All Sales to RW industries as Vertically Related								
All trades	2.7	17.8	57.4	90.6	22.7	2.1	33.6	54,042
IOT Requirement	1.2	10.5	41.2	73.2	24.2	1.0	21.8	30,772
AHS Requirement	0.4	6.3	34.5	73.4	29.4	1.2	11.4	21,804
Intensive Margin Only	6.3	29.1	78.2	98.8	14.1	3.5	36.9	30,950
Panel 6: Reporting Firms in All CFS industries								
All trades	2.5	16.8	56.7	90.6	22.3	2.1	31.3	66,126
IOT Requirement	0.5	6.9	32.8	63.4	27.2	0.9	15.0	29,455
AHS Requirement	0.1	4.1	28.3	64.9	35.1	1.1	9.1	19,558
Intensive Margin Only	8.5	36.1	87.7	99.9	10.7	4.4	38.9	19,669
Panel 7: Related-party Purchases								
All trades	3.8	17.5	41.8	62.0	23.8	0.8	28.8	53,175
IOT Requirement	0.9	8.4	24.9	40.9	26.3	0.2	14.9	23,368
AHS Requirement	0.2	4.6	18.9	35.3	37.2	0.1	7.3	18,701
Intensive Margin Only	7.7	25.5	52.8	73.4	14.1	1.4	32.2	23,483
Panel 8: Reporting Firms in Industries with Prior of High RPT Use								
All trades	5.0	27.4	79.4	99.4	20.8	3.8	36.9	7,745
IOT Requirement	4.6	25.6	72.4	94.4	21.3	2.5	23.2	5,347
AHS Requirement	1.6	15.9	63.2	94.4	27.4	2.4	13.5	4,198
Intensive Margin Only	10.5	42.1	93.6	100.0	13.2	5.3	38.0	5,347

NOTE.—This table reports the share of firms' sales that is sold to related parties. Sample *N* consists of firm-years that have at least one related party, regardless of whether the reporting firm has any related-party transactions: due to missing firm-level information, total *N* used in the table is slightly smaller than the 125,044 available firm-level reports. The weighted mean is weighted by the size of each reporting firm's total sales.

tries finely enough.[27] For example, Korean IOT treats the entire retail and wholesale industries as one segment, while dividing the manufacturing industries into 234 sectors. Moreover, IOTs show the value of intermediates used to *produce* the retail and wholesale services, which does not accurately represent the flow of good and services that go through these industries.

Panel 5 of Table 2 treats all sales to related parties in the retail and wholesale industries as satisfying the IOT criteria. Again, this is the most conservative measure that would make the impact of imposing IOT requirements the smallest. However, the results remain qualitatively unchanged. While the estimates with additional restrictions appear slightly larger, the differences are small.

The last robustness check in this section fully mirrors AHS in terms of which industries' reporting firms are used. AHS's sample is from the Commodity Flow Survey (CFS), which contains establishments in mining, manufacturing, wholesale, and select retail industries. As the main results in Table 1 limit to manufacturing industries only, or use all industries, in Panel 6 of Table 2 I report the results using the same set of industries as in CFS. Again, this produces only very minor differences in results.

4.2 Analysis of Distinct Sub-Segments of the Data

Panel 7 examines firms' purchases from their related parties, instead of the sales. In particular, this robustness check explores whether firms' reliance on related parties is different for input purchases compared to sales of output. For instance, while the median firm sells 2.7% of its total sales to related parties, it may be purchasing a much higher share of its needed inputs from them. I report Panel 7 using purchases from related parties as shares of the firms' cost of goods sold (COGS). The differences from the main result are minor: more than 75% of firms purchase from their related parties, and the limitations of proxy measures substantially underestimate the outcomes.

The final robustness check computes the share of sales for firms in industries where existing literature has underscored the importance of related-party transactions. Specifically, Panel 8 utilizes fifteen 4-digit industries that are reviewed in Lafontaine and Slade (2007).[28] Reassuringly, the results confirm prior beliefs. The share of trades is larger at every reported percentile based on the alternative construction: the 75th and 90th percentile values are 27.4% and 79.4%, approximately 10 and 20 percentage points greater than the corresponding values in Table 1, respectively.

[27] Many previous studies that utilize IOTs recognize this issue as well (Fan and Lang, 2000; Acemoglu, Johnson and Mitton, 2009). AHS utilizes information from the Annual Wholesale Trade Survey and the Annual Retail Trade Survey to address this issue: similar data is unavailable for South Korea.

[28] The procedure is almost identical to a robustness check performed in AHS, with small differences in the specific industries considered. The included industries are coal mining, petroleum refining, footwear manufacturing, soft drink bottling, organic chemicals manufacturing, cement manufacturing, auto parts manufacturing, aircraft parts manufacturing, iron ore mining, pulp manufacturing, and shipbuilding. The last three industries are not included in AHS due to CFS's scope or confidentiality. Removing them from consideration creates no qualitative change in the results, and in fact, it makes the related-party trades appear slightly larger than is reported in Panel 8 of Table 2.

5 Alternative Proxies from Supervised Machine Learning

5.1 Data and Methodology

Despite the common proxies' limited performances, they have been vital for numerous researches in vertical and horizontal integration. Such works commonly require sorting out firms within business groups that are linked through supply chains, and the proxy has been widely applied to this end. In this section, I propose an alternative, novel method to proxy for such firms by utilizing supervised machine learning. Through this, I show that by using actual intra-party trade data, researchers can obtain dramatically improved performances even with relatively off-the-shelf methods such as random forests.

The strength of this proxy is threefold. First, this is the first measure that uses actual trades to optimize prediction performance, and therefore it does not need to rely on arbitrary cutoffs as in the previous measures. Second, it utilizes a substantially wider range of information than the industry-level IOT coefficients. Lastly and simultaneously, this measure preserves the previous proxies' broad applicability to diverse data environments, as it only requires the most commonly available variables to produce predictions.

It is important to note that this method is only the first step in the right direction and is not a perfected panacea. As the data used to train the algorithm can only come from Korea, a rigorous evaluation of the external validity of its performances in other countries is inherently impossible. However, this method is valuable as it can easily be improved by expanding the scope with the addition of other countries' data. Such expansion is probable considering that the accounting regime that this paper exploits, the IFRS, has been adopted in a broad range of countries, including most of Europe. Moreover, even in its current form, its advantages over traditional proxies are clear as this method is firmly rooted in actual data.

I train a random forests model (Breiman, 2001) using a set of predictors that have been discussed in the literature as related to firms' participation in related-party trades: IOT coefficient, firm size measured in assets and sales, group size measured in the number of firms, industry contractibility from Rauch (1999), an indicator of whether the reporting firm has direct majority control over the related party, firm location, and industry of both of the related firms.

In this section, I utilize a smaller subset of firms compared to Section 3 to ensure that I have the complete list of related parties for each reporting firm. As a result, I utilize a total of 2,607 firms over the FY 2013–2019 that have 420,428 firm-year-related party triples. Among the triples, 105,367 (25.1%) report transactions in terms of sales, purchases, loans, or debts, while the rest do not trade with the reporting firms. The main body of this section will primarily report the algorithms trained only with manufacturing firms. This leaves 1,600 firms over the same data period, with 197,021 firm-year-related party triples. Among the triples, 56,158 (28.5%) report intra-party transactions. Further details of the method, construction of the data, and predictors

Table 3: Confusion Matrix: Predicting Related-Party Trades Based on IOT

(a) IOT Cutoff ≥ 0.01

		Actual	
		Trade	No Trade
Pred.	Trade	6,058	7,115
	No Trade	16,796	52,277

(b) Total Requirements > 0

		Actual	
		Trade	No Trade
Pred.	Trade	22,854	59,392
	No Trade	0	0

(c) ML Algorithm from 2019 Data

		Actual	
		Trade	No Trade
Pred.	Trade	440	178
	No Trade	249	1,270

Note.—Panels (a) and (b) represent confusion matrices of utilizing only a single cutoff of the chosen IOT coefficients to discern the pairs of related parties that trade, from those that do not. Panel (a) shows the results of predictions based on the cutoff of 1% of the seller industry's intermediate sale going to the buyer industry. Panel (b) shows the performances of using the total requirements coefficient having a strictly positive value. Panel (c) shows the confusion matrix from using a random forests algorithm trained with data from 2019. *Pred.*, or *Prediction*, denotes whether the method in question predicts the pairs to trade, and *Reference* denotes whether the pairs trade in the true data. The data spans all publicly traded manufacturing firms in Korea, over 2013-2019, and their related parties.

are outlined in Appendix B.

5.2 Prediction Results

Several key measures are utilized to assess prediction performances. As an illustrative example, I first show the classification performances of two widely used methods from the literature that each relies on a single IOT coefficient cutoff. The confusion matrices in Tables 3a and 3b divide pairs of related parties in the data according to how well each method *predicts* whether the pair trades.

Table 3a presents the confusion matrix for predictions of related-party trades using 1% of the IOT coefficient as the cutoff. At first glance the method appears to perform relatively well; it classifies 70.9% of the related party pairs correctly, with 6,058 true positives and 52,277 true negatives.

However, this simple *accuracy* measure[29] of 70.9% masks underlying problems. The correctly classified observations are composed mostly of the pairs that do not trade, while not many of the trading pairs are picked up. Out of the related-party pairs that are actually trading, this method detects only 6,058 (26.5%); in other words, the *recall* rate is low.[30] Moreover, the group of pairs predicted to trade is instead composed of more pairs that actually do not. Only 6,058 (46.0%) of the pairs predicted to be trading are true positives; therefore, the *precision* rate of this method is low as well.[31] To summarize, attempts to study vertically trading entities through this proxy measure would capture only a small portion of those that are actually trading, and the constituted sample would in fact consist of more non-trading entity pairs than trading pairs.

Table 3b, using another prominent proxy from the literature, displays a different problem. When we deem the pairs with a positive total requirements coefficient to be vertically integrated, the

[29] Accuracy = (True positives + True negatives) / All observations

[30] Recall = True positives / (True positives + False negatives)

[31] Precision = True positives / (True positives + False positives)

Table 4: Prediction Performance Metrics: 2013-2019 Algorithms

Year (1)	Accuracy (2)	Precision (3)	Recall (4)	Specificity (5)	AUC (6)	PR-AUC (7)
2019	0.800	0.712	0.639	0.877	0.861	0.708
2018	0.816	0.686	0.658	0.879	0.888	0.719
2017	0.770	0.678	0.547	0.876	0.848	0.701
2016	0.757	0.658	0.586	0.844	0.830	0.700
2015	0.703	0.657	0.593	0.781	0.778	0.712
2014	0.820	0.672	0.660	0.880	0.881	0.737
2013	0.757	0.648	0.599	0.837	0.826	0.730

NOTE.—This table reports the prediction performance metrics of algorithms created from each year's training data, tested on the same year's out-of-sample testing dataset. Information from all public manufacturing firms in Korea and their related parties is utilized.

cutoff proves to be too lenient, not categorizing any firm-related party pairs as non-trading.[32] As such, while it detects all (27.8%) observations that are actually trading and records 100% recall, both its accuracy and precision are very low at 27.8%.

In contrast, the random forests algorithm provides much more accurate and consistent predictions. Table 3c reports the performance of an algorithm trained with the most recent year's data, 2019, when applied to the same year's out-of-sample testing set. The algorithm detects 62.7% of the actual trading pairs, a substantial jump from 26.5% in Table 3a. Moreover, among the pairs predicted to be trading, 71.1% are, in fact, doing so, again showing a sizable increase from 46.0% and 27.8% based on the more traditional measures. This result is all the more notable, considering that the model requires only a small amount of additional information over the traditional measures.

The improved performance is not limited to 2019 but is consistent throughout all years. Table 4 shows the out-of-sample performance of algorithms built with each year's data. Here, a separate algorithm is trained each year with how 80% of firms in the year's data trade with their related parties, then is tested on the remaining 20%. In accuracy, precision, and recall, the performance gains over the traditional cutoffs are strong and consistent. Specificity, which measures how well the mechanism detects the non-trading pairs, is consistently strong as well.[33]

What is more, the algorithms fare well in other key metrics that are widely used in the prediction literature. Appendix Figure 3a plots the Receiver Operating Characteristic (ROC) curve from 2019. This curve plots the tradeoffs between true positive rates (*recall*) and false positive rates

[32] This threshold intends to account for indirect supply chains, and therefore is more inclusive by design. As such, it has proven to exclude only a minimal fraction of observations in other datasets as well. Alfaro et al. (2019) reports that in the WorldBase dataset that they utilize, 98.0% of the related parties of parent firms satisfy the total requirements criteria.

[33] Specificity = True Negatives / (True Negatives + False Positives)

($1 - specificity$) as the threshold becomes more relaxed for declaring a pair to be trading. As the method becomes more lenient in declaring a pair to be trading, any classification approach detects more true positives and produces higher recall. At the same time, it is more likely to misconstrue non-trading pairs as trading, and have higher false positive rates. Encapsulating this curve, the AUC score calculates the *Area Under the (ROC) Curve* which provides a measure of the estimated probability that a positive case, in this case a trading pair, will be ranked higher by the algorithm than a negative case (Hosmer Jr, Lemeshow and Sturdivant, 2013). As column (6) of Table 4 reports, the AUC scores consistently report a strong result.

When the positive cases are scarce such that there is a large class imbalance, AUC scores could be overly optimistic (Davis and Goadrich, 2006). In this case, the PR-AUC scores are often used to evaluate the model performances. This score calculates the area under Precision-Recall curves, plotted in Appendix Figure 3b, which shows the tradeoff between *precision* and *recall* as the predictive threshold changes. While the class skew is not strong in this paper, with 27.8% trading, PR-AUC scores are also consistently strong as shown in column (7) of Table 4, sufficiently addressing any possible concerns.

6 Conclusion

A major theoretical focus in vertical integration has been on its ability to facilitate the trade of goods and services along production chains. While the lack of related-party trade data has made direct observation or measurement of the trades difficult, the empirical literature has utilized proxy measures to infer them and has found trades within related parties to be surprisingly small.

I construct novel firm-level data on Korean firms' related-party trades from their financial reports, and demonstrate for the first time the true size, direction, and prevalence of these trades. In contrast with the existing empirical literature, most firms appear to be engaged in related-party trades and the trades are shown to assume a substantially larger share of the firms' total sales and purchases.

A commonly used proxy for related-party trades, based on IOTs, appears to have caused much of this disparity between the theoretical and the empirical literatures. The traditional approach to creating a proxy only captures the tip of the iceberg, missing a much larger share of trades between related parties in seemingly unrelated industries. Out of all related entities that purchase from the reporting firms, only 14.4% are in industries that satisfy the common IOT requirements, and the sales to them represent only 31.7% of total related-party sales in the economy.

This strongly signals that (i) vertical integration involves active trading, and that (ii) related-party trades are utilized in contexts that are tailored to specific circumstances or needs of the firms, and do not simply follow the economy-wide trade patterns represented in input-output tables.

Using IOTs to infer related-party trades from the network of related firms, it seems, is a perilous approach that should be utilized with caution.

The usage of the proxy, however, was inevitable in many papers due to the lack of available data. Given this, what should researchers do when faced with this lack of data about the true state of the world? How can we better infer the existence and the magnitude of related-party trades from more commonly available data sources?

To this end, I utilize the random forest method to construct predictive algorithms that researchers can apply to their own datasets. While the method requires only a small amount of additional publicly available data, it shows a marked improvement in performance. Moreover, I show that variables such as group and firm sizes, as well as the firm's control over the related party, are important predictors of whether two related entities are trading.

It should be emphasized that the proposed method is only a first step in a better direction. It is a demonstration of what can be achieved by leveraging actual data even with relatively simple methods, and may benefit from integrating new data and methods. For example, while its strong out-of-sample prediction performance is documented in various ways, the lack of data currently prohibits verification of its performance across a diverse array of countries and contexts. By expanding the related-party trade data that this algorithm relies on, it can be updated to account for possible country-specific nuances. This a highly achievable goal as numerous countries share the accounting requirements that this paper exploits.

This paper indicates that widely used proxy measures appear to have caused notable biases in our perception of the size and prevalence of vertical trades, and presents possible alternatives to the proxies. Then, a natural question follows: what other aspects of vertical integration could be better understood by this improvement in the long-standing measurement problem? I intend to pursue this in separate papers.

References

Acemoglu, Daron, Simon Johnson, and Todd Mitton. 2009. “Determinants of vertical integration: financial development and contracting costs.” *The Journal of Finance*, 64(3): 1251–1290.

Aghion, Philippe, Rachel Griffith, and Peter Howitt. 2006. “Vertical integration and competition.” *American Economic Review*, 96(2): 97–102.

Alfaro, Laura, and Andrew Charlton. 2009. “Intra-industry foreign direct investment.” *American Economic Review*, 99(5): 2096–2119.

Alfaro, Laura, Davin Chor, Pol Antras, and Paola Conconi. 2019. “Internalizing global value chains: A firm-level analysis.” *Journal of Political Economy*, 127(2): 508–559.

Alfaro, Laura, Paola Conconi, Harald Fadinger, and Andrew F Newman. 2016. “Do prices determine vertical integration?” *The Review of Economic Studies*, 83(3): 855–888.

Alfaro-Urena, Alonso, Isabela Manelici, and Jose P Vasquez. 2022. “The effects of joining multinational supply chains: New evidence from firm-to-firm linkages.” *The Quarterly Journal of Economics*, 137(3): 1495–1552.

Altomonte, Carlo, Gianmarco Ottaviano, Armando Rungi, and Tommaso Sonno. 2021. “Business Groups as Knowledge-Based Hierarchies of Firms.” *CEPR Discussion Paper*, No. 16677.

Antras, Pol, and C Fritz Foley. 2015. “Poultry in motion: a study of international trade finance practices.” *Journal of Political Economy*, 123(4): 853–901.

Antràs, Pol, and Davin Chor. 2013. “Organizing the global value chain.” *Econometrica*, 81(6): 2127–2204.

Atalay, Enghin, Ali Hortaçsu, and Chad Syverson. 2014. “Vertical integration and input flows.” *American Economic Review*, 104(4): 1120–48.

Atalay, Enghin, Ali Hortaçsu, Mary Jialin Li, and Chad Syverson. 2019. “How wide is the firm border?” *The Quarterly Journal of Economics*, 134(4): 1845–1882.

Athey, Susan, and Guido W Imbens. 2019. “Machine learning methods that economists should know about.” *Annual Review of Economics*, 11: 685–725.

Blyde, Juan, and Danielken Molina. 2015. “Logistic infrastructure and the international location of fragmented production.” *Journal of International Economics*, 95(2): 319–332.

Breiman, Leo. 2001. “Random forests.” *Machine learning*, 45(1): 5–32.

Davis, Jesse, and Mark Goadrich. 2006. “The Relationship between Precision-Recall and ROC Curves.” *ICML '06*, 233240. New York, NY, USA:Association for Computing Machinery.

Efron, Bradley. 2020. “Prediction, estimation, and attribution.” *International Statistical Review*, 88: S28–S59.

Fajgelbaum, Pablo, Gene M Grossman, and Elhanan Helpman. 2015. "A Linder hypothesis for foreign direct investment." *The Review of Economic Studies*, 82(1): 83–121.

Fan, Joseph PH, and Larry HP Lang. 2000. "The measurement of relatedness: An application to corporate diversification." *The Journal of Business*, 73(4): 629–660.

Fort, Teresa C. 2017. "Technology and production fragmentation: Domestic versus foreign sourcing." *The Review of Economic Studies*, 84(2): 650–687.

FTC. 2021. "Statement of Chair Lina M. Khan, Commissioner Rohit Chopra, and Commissioner Rebecca Kelly Slaughter on the Withdrawal of the Vertical Merger Guidelines." *Commission File No. P810034*.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. 2022. "Predictably unequal? The effects of machine learning on credit markets." *The Journal of Finance*, 77(1): 5–47.

Garg, Shresth, Brandon Tan, and Pulak Ghosh. 2021. "Within Firm Supply Chains: Evidence from India." *Working Paper*.

Grossman, Sanford J, and Oliver D Hart. 1986. "The costs and benefits of ownership: A theory of vertical and lateral integration." *Journal of Political Economy*, 94(4): 691–719.

Hart, Oliver, and John Moore. 1990. "Property Rights and the Nature of the Firm." *Journal of Political Economy*, 98(6): 1119–1158.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression.* Vol. 398, John Wiley & Sons.

Jeong, Eun-Young, and Timothy W Martin. 2018. "South Korea Regulator Says Samsung Bio-Logics Violated Accounting Rules." *The Wall Street Journal*.

Klein, Benjamin, Robert G Crawford, and Armen A Alchian. 1978. "Vertical integration, appropriable rents, and the competitive contracting process." *The Journal of Law and Economics*, 21(2): 297–326.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *The Quarterly Journal of Economics*, 133(1): 237–293.

Lafontaine, Francine, and Margaret Slade. 2007. "Vertical integration and firm boundaries: The evidence." *Journal of Economic Literature*, 45(3): 629–685.

Li, Kai, Feng Mai, Rui Shen, and Xinyan Yan. 2021. "Measuring corporate culture using machine learning." *The Review of Financial Studies*, 34(7): 3265–3315.

Monarch, Ryan, Jooyoun Park, and Jagadeesh Sivadasan. 2017. "Domestic gains from offshoring? Evidence from TAA-linked US microdata." *Journal of International Economics*, 105: 150–173.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine learning: an applied econometric approach." *Journal of Economic Perspectives*, 31(2): 87–106.

Nunn, Nathan. 2007. "Relationship-specificity, incomplete contracts, and the pattern of trade." *The Quarterly Journal of Economics*, 122(2): 569–600.

Nunn, Nathan, and Daniel Trefler. 2013. "Incomplete contracts and the boundaries of the multinational firm." *Journal of Economic Behavior & Organization*, 94: 330–344.

Ramondo, Natalia, Veronica Rappoport, and Kim J Ruhl. 2016. "Intrafirm trade and vertical fragmentation in US multinational corporations." *Journal of International Economics*, 98: 51–59.

Rauch, James E. 1999. "Networks versus markets in international trade." *Journal of International Economics*, 48(1): 7–35.

Ruhl, Kim J. 2015. "How well is US intrafirm trade measured?" *American Economic Review*, 105(5): 524–29.

Santioni, Raffaele, Fabio Schiantarelli, and Philip E Strahan. 2020. "Internal capital markets in times of crisis: The benefit of group affiliation." *Review of Finance*, 24(4): 773–811.

Williamson, Oliver E. 1971. "The vertical integration of production: market failure considerations." *The American Economic Review*, 61(2): 112–123.

Williamson, Oliver E. 1979. "Transaction-cost economics: the governance of contractual relations." *The Journal of Law and Economics*, 22(2): 233–261.

A Appendix: Data

This section of the appendix provides details of the data construction process. The first half describes details of the related-party data. Subsection A.1 outline the scope of Korean firms that report the information, and A.2 provides detailed definitions of a *related party* and the types of trades that are reported. The information is based on Korean accounting regulations, which I translated and summarize here. For the original text, see Korean International Financial Reporting Standards (K-IFRS), especially the section on Related-Party Disclosures (1024).

Then, the section details how the rest of the data were constructed. Subsection A.3 details how information other than the related-party trades (e.g. firm size and industry) were matched to each firm. Subsection A.4 explains how the main analysis finds the *reporting firms that have at least one related party*.

Lastly, subsection A.5 checks the validity of the data construction process by cross-checking the numbers with established firm-level databases. While the scope of the data in these databases is much more limited compared to my own, the data used in this paper appear highly similar when aggregated up to the same level available in the databases.

A.1 Related-Party Trade Data

The related-party trades are web-scraped from side notes of firms' annual financial reports. This section describes the sample of firms included in this data and the scraping process.

Firms[34] in South Korea that satisfy a set of size thresholds are legally required to receive an

Table 5: Reporting Firm Criteria

FY start date	Types of Firms	Criteria*
Before 2018.11.1	Corporations	Total Assets ≥ $10m Publicly traded or will be in the next FY Total Assets ≥ $5.8m *and* Total Debts ≥ $5.8m Total Assets ≥ $5.8m *and* Employees ≥ 300
On 2018.11.1 –	Corporations**	Total Assets ≥ $41.6m Total Sales ≥ $41.6m Satisfies 2 of $\begin{cases}\text{Total Assets} & \geq \$10\text{m}\\ \text{Total Debts} & \geq \$5.8\text{m}\\ \text{Total Sales} & \geq \$8.3\text{m}\end{cases}$

** All monetary units are converted to USD from the original Korean Won using a rough exchange rate of 1200 Won = 1 USD.*
*** Also includes limited companies from FY starting on 2019.11.1 and after.*

[34] More specifically, a type of firm, corporations, are subject to this reporting requirement during this paper's data period. As corporations are consistently more than 94% of all firms in Korea during the same period, I use the terms interchangeably.

annual external audit and publicly disclose the reports. The thresholds have undergone a small update during the data period. While the details of which are outlined Table 5, data from the only period affected by this change, the fiscal year 2019, does not demonstrate a material difference from other years.

A.2 Definition of Terms

A.2.1 Related Parties

On the 'Related-Party Trade' section of the side notes to the financial reports, firms disclose the names and relationships of the related parties that are involved in trades with the reporting firm. Specifically, the reporting firms are required to disclose relationships with the related parties in the following seven categories.

(1) Parent firm (holding a majority voting power)
(2) Parent firm (holding a minority voting power, or a joint parent firm)
(3) Subsidiary (holding a majority voting power)
(4) Affiliate (holding a majority voting power)
(5) Jointly owned subsidiary
(6) Board members of reporting firm or its parent firm
(7) All other related parties ('sibling'/'cousin' firms, etc.)

In practice, firms often report the relationships using much more finely defined categories. In that case, I re-categorize them into the seven categories.

A.2.2 Variables Reported ("Trades")

The firms are required to disclose the following categories of transactions with their related parties. From the data, all transactions corresponding to sales, purchases, debts, and loans are taken and aggregated into the correct category. Firms often report transactions using more granular categories, which I re-summarize into the four categories. Note that the provision of debt guarantees or collateral is excluded from the data used in the main analysis.

(1) Sales and purchase of goods (final and intermediate)
(2) Sales and purchase of real estate or other assets
(3) Sales and purchase of services
(4) Lease
(5) R & D
(6) License
(7) Loans, debts and other investment
(8) Provision of loan guarantee / collateral
(9) Uncompleted contract
(10) Provision of debt payment on behalf of the other

A.3 Matching Firm-level Information to Related-Party Trade Data

The related-party trade data from Korea provides a unique firm-level 10-digit identifier (Business Registration Number) for the reporting firm. Firm-level information is easily matched to the reporting firms, using the identifier, from three existing firm-level databases: KISVALUE, TS2000, and ORBIS. KISVALUE and TS2000 are the two largest firm-level databases in Korea. The two databases, when combined, contain all reporting firms in my related-party trade sample. The ORBIS database is compiled by Bureau Van Dijk, and includes a large subset from the related-party trade dataset as well as a larger number of firms outside of South Korea.

Values from the three databases are consistent but often display minor differences. In case of a difference in numerical accounting data (e.g., total sales, cost of goods sold, etc), I assign priority to the information from TS2000, KISVALUE, then ORBIS, following the number of reporting firms that can be matched with each database. When the databases report different industry codes for the firms, (i) the most detailed industry information is used, and (ii) if the industry codes display the same level of details or digits, the databases are given the same priority as the accounting data. Note that information from separate financial statements (at the individual firm level) is used instead of the consolidated financial statements, which report combined information of parent companies with their subsidiaries.

On the other hand, for the related parties that are reported, the only available information is their firm names. The related parties are then matched to data in existing databases by the names. The firm names are cleaned, then matched with 10-digit identifiers by (i) historical firm name data from the DART website as well as from existing domestic databases, then in case the firm names cannot be matched to the domestic 10-digit identifiers, it is then matched by (ii) firm names in the ORBIS database.

The name-matching process goes through extensive checks, and conservative criteria are used to check the validity of matching in order to ensure accuracy. Here I summarize the key steps taken. The key is to use a dual-reporting feature: if firm A reports trades with firm B that is large enough, firm B would also be reporting trades with firm A in the same year. Then, the trades reported by firm A should be equal to the trade reported by firm B. In the following, firm A denotes the original reporting firm and firm B denotes one of A's related parties.

1. If a unique firm identifier is matched with firm B,
    (a) If there is an annual financial report from the matched firm identifier in the given year, the match is regarded correct only if B's financial report also lists firm A as its related party.
    (b) If there is no annual financial report from the matched firm identifier in the given year, then the match is assumed as a correct match.
2. If multiple firm identifiers are matched with firm B's name,

(a) If there is a unique firm identifier that has an annual financial report in the same year that lists firm A as its related party, then that firm identifier is regarded as a correct match.

(b) If there are multiple identifiers that have annual financial reports in the same year that lists firm A as its related party, or if there is no such 10-K report, then no firm identifier is deemed a correct match.

3. After (1) and (2), if firm B is not matched with an existing firm identifier from Korea, firm B's name is checked against the ORBIS database. Information from the ORBIS database is deemed correct only if it is a unique match with a firm identifier.

After the matching process, 52.1% of the observations are matched with information on the related parties.

The related-party trade numbers go through a similar verification process. The process is largely in line with the data verification process outlined in the Data Appendix of Alfaro-Urena, Manelici and Vasquez (2022).

A.4 Finding Firms with at Least One Related Party

The analysis of the paper excludes singleton firms, or firms without any related party, from the sample as they inherently cannot engage in related-party trades. However, the financial reports do not provide a firm's full list of related parties, instead only displaying those that have *transacted* with the reporting firm—be it in sales, purchases, loans, or debts—during the fiscal year. Hence, when relying solely on the reports, a firm that has related parties but is not actively trading with any will be marked as a singleton and excluded from the sample.

I use the ORBIS database to supplement the 10-K reports, as it provides for each firm the list of all other firms in the database that share the same *ultimate owner*. Throughout this paper, I take a conservative approach and use the set of firms reported to have a related party *either* in ORBIS or in the financial reports.

Two caveats exist in the process. First, the ORBIS database's definition of a related party differs in detail from my data based on Korean financial statements. For example, in the ORBIS database, the minimum ownership percentage criterion is 25% while it is 20% in the Korean data. While these minor differences may affect the sample of firms that I consider, their significance is expected to be very small: observations with smaller than 50% ownership share are already scarce in the ORBIS database (3.34%). While it is possible that there could be a bunching between 20% and 25% ownership shares, there is no reason to anticipate the bunching to exist or the magnitude of it to be consequential.

Second, As the ORBIS database only provides the most recent ownership information for each firm, the set of firms that have related parties in ORBIS in 2019 are utilized throughout the entire data period. I contend that any possible biases from this limitation are not significant to the main results in Table 1. First, any bias will act to include singleton firms in the sample that I consider, more so for the earlier years of the data, and therefore result in underestimating firms' related-party trades. As the main finding of the paper is that these trades are larger and more prevalent than previously thought, this paper's results will only be stronger without the possible biases. Second, the possible biases are likely small, as the attrition rate is small: using a sub-sample of all public firms in Korea, for which the complete list of firms' related parties are observable, I confirm that 91.3% of firms with a related party still do after 3 years.

A.5 Comparison of data with existing databases

This section provides a test of the validity of the scraped related-party trade data's accuracy by comparing the results with two existing databases. In general related-party information has rarely been utilized as it is listed only in the side notes of financial statements. The rare exceptions are KISVALUE and TS2000, the two largest firm-level databases in Korea, which each contain a part of the data. Here, I show that a comparison of the scraped data with existing databases yields almost identical results.

KISVALUE provides the related-party data aggregated at the reporting firm's level only. Even though trades with individual related parties are not visible, the database includes a large subset of firms in my related-party trade data: it includes 19,464 firms, compared to the 30,390 firms in my own. In Table 6 I report the comparison of related-party trades for the sample of firms that are included in KISVALUE and the scraped data at the same time. The results confirm the accuracy of the scraping process. The difference between my data and KISVALUE is minuscule. Moreover, the related-party trades appear slightly larger in the existing database, suggesting that the results in Table 1 are not overestimating, if underestimating slightly.

It is worth noting that KISVALUE reports NA values for the firms that do not engage in any related-party trades, as well as for some firms that post zero trades. Therefore, by utilizing only the firms with non-NA related-party trade values in KISVALUE, as in Panel 2, the numbers may overrepresent the firms with some level of related-party transactions, be it sales, purchases, loans, or debts. In Panel 1, I mitigate the effect by including more firms. First, using my scraped data I identify the firms that have no related-party trades (but have all NA values in KISVALUE), then include them in KISVALUE samples as having zero related-party trades.

In contrast, TS2000 provides the data defined at the most similar levels of detail to my own, but for a much narrower scope of firms and with different variable definitions. The database reports the trades for 2,284 public firms in Korea, compared to the 30,390 firms used in the main paper. Moreover, the reporting firms in this sample are significantly larger: the median total sales of firms in this sample is roughly $61 million USD, while for the full sample it is $16 million USD.[35] Lastly, TS2000 divides the trades differently: the sales and purchases are divided into those related to the firm's core business operations, and those that are not (*operating* vs. *non-operating* income and cost), and *others*, a category assigned when the database could not determine whether a reported transaction is about the core operation or not. The problem is, this *others* category was defined too liberally: the numbers are too large, and it does not distinguish sales from purchases, simply listing the sum of reported numbers.

Panel 3 of Table 6 compares the related-party trades reported by TS2000 with the scraped data. To account for the *others* category, here the numbers reported are the sum of sales and purchases, as a share of each firm's total sales.

Table 6: Share of Sales to Related Parties, Comparison with KISVALUE

Related Party Share	Percentiles (%)				Fraction (%)		Weighted	
of Firm Total Sales	50th	75th	90th	95th	=0	≥1	Mean (%)	*N*
Panel 1: All firms with RPT in KISVALUE: Sales								
Scraped - All Industry	1.4	13.8	54.5	89.7	29.2	2.4	27.0	85,776
KISVALUE - All Industry	1.6	14.3	55.5	90.0	26.9	2.0	26.2	85,850
Scraped - Manufacturing	2.6	16.6	52.5	85.4	22.0	1.8	33.3	41,053
KISVALUE - Manufacturing	2.9	17.0	52.7	84.8	19.7	1.4	32.9	41,077
Panel 2: All firms with RPT in KISVALUE: Sales								
Scraped - All Industry	2.8	18.4	63.3	94.4	18.4	2.8	27.8	74,456
KISVALUE - All Industry	3.1	19.0	64.2	94.2	15.7	2.3	27.0	74,530
Scraped - Manufacturing	4.0	19.9	58.0	89.8	13.0	2.1	34.2	36,793
KISVALUE - Manufacturing	4.3	20.2	57.7	89.4	10.4	1.6	33.5	36,817
Panel 3: All firms with RPT in TS: Sales + Purchases								
Scraped - All Industry	13.5	40.8	81.6	101.0	4.8	5.6	48.9	11,768
TS2000 - All Industry	12.2	39.0	81.2	101.6	11.9	5.8	42.0	12,440
Scraped - Manufacturing	14.4	39.5	73.8	96.4	4.7	4.0	58.0	7,717
TS2000 - Manufacturing	13.1	37.8	70.9	94.6	12.5	3.8	56.1	7,957

NOTE.—The table compares the related-party trades reported in KISVALUE with the scraped and cleaned data used in this paper.

[35] The median total sales of reporting firms in KISVALUE's related-party trade sample is $27 million USD.

B Constructing New Method to Proxy for RPT with Machine Learning

In this section, I provide more details on the construction of the new proxy in Section 5. In section 5, I propose an alternative proxy by utilizing supervised machine learning. This is the first such measure that is optimized based on how it performs to predict actual intra-party trades. Hence it is able to (i) move away from the arbitrary cutoffs used in the previous measures, and (ii) utilize a wider range of information. Here, the goal is to come up with an algorithm that only requires *commonly available data* to produce predictions on which pairs of related parties are trading and which are not, so that it can be used in more standard data environments.

B.1 Methodology

Supervised machine learning offers a number of advantages for this task. First, it is best suited to out-of-sample predictions, compared to the more traditional econometric toolbox that focuses on the in-sample fit of a model. As the desired end-product is an algorithm that researchers can apply to their own dataset, I opt for machine learning to optimize performance and avoid overfitting to my own data. Second, the models are flexible as they allow for a large number of predictors as well as non-monotonic relationships between the outcome and the predictor (Athey and Imbens, 2019). Lastly, when the set of potential predictors is large and the key predictors are not clearly identified, researchers can establish the relative importance of each predictor (Breiman, 2001).[36]

Specifically, I apply the random forests approach to generate predictive algorithms classifying pairs of related parties as either trading or not trading. As the approach is widely in use, I only provide a brief description. The random forests approach estimates an individual decision tree by sequentially splitting the data based on optimized cutoffs of the most informative predictors. To illustrate a simple example, if the IOT coefficient is the only predictor, a tree would find a cutoff (or cutoffs) of the coefficient that best divides the data into groups of trading pairs and non-trading pairs. In practice, each tree is drawn from many predictors and cutoffs. A random forest aggregates a multitude of decision trees that are created by bootstrapping different subsets of both the data and the predictors. This aggregation aims to address the limitations of relying on a single tree, such as the volatility of the results and an overdependence on the variables used in early splits.

In order to form a prediction on which pairs of related parties are trading and which are not, I use a variety of possible predictors—referred to as 'features' in the machine learning literature—derived from related literature. The predictors include a range of basic industry, firm, and group-level characteristics; for a complete list, see Table 8.

[36] Machine learning is increasingly being used by economists in the academic domain. See, for example, Kleinberg et al. (2018), Fuster et al. (2022), and Li et al. (2021). Athey and Imbens (2019) and Mullainathan and Spiess (2017) provide excellent reviews of machine learning applications for economists. Efron (2020) presents a succinct comparison between predictive algorithms and standard regression techniques.

Table 7: List of Predictors and Related Literature

Category	Predictors
1. IOT coefficients	Direct and total requirements, share of sales to RP industry (Atalay et al., 2019; Alfaro et al., 2019)
2. Firms' basic information	Size (assets, sales) of firms (Ramondo, Rappoport and Ruhl, 2016)
3. Group's basic information	Size of group (number of firms) (Ramondo, Rappoport and Ruhl, 2016)
4. Industry contractibility	Index derived from Rauch (1999) (Rauch, 1999; Nunn, 2007)
5. Control over RP	Dummy for whether RP is a subsidiary (Antràs and Chor, 2013)
6. Location of RP	Dummy for whether RP is a domestic firm (Antras and Foley, 2015)
7. Industries of firms	2-digit KSIC (3-digit NAICS) codes (Lafontaine and Slade (2007), AHS)

Note.—This table provides the list of predictors used in Section 5, and examples of the papers that discuss relevance between the predictor and related-party trades.

B.2 Choice of Features

The first set of features involves both parties' industry-specific characteristics, including (i) IOT coefficients, (ii) a measure of industry contractibility, and (iii) industry dummies. IOT coefficients reflect technologically determined intermediate input needs for each industry, and have been most widely used to infer related-party trades. I include three different types of coefficients that have each been utilized in the literature: the direct requirements coefficients, total requirements coefficients, and the share of intermediate sales directed to a specific industry.[37] Also, I include a measure of how contractible an industry's output is—or more specifically, how relationship-specific it is—to account for potential holdup problems.[38] Past studies have also pointed to specific industries as more likely participants in intra-party trades. To account for this, in some specifications I also include dummy variables indicating the industry code.[39] of both firms

[37] Direct requirements show the share of industry i's intermediate inputs that come directly from another industry j, while total requirements represent both direct and indirect inputs from j. The last coefficient, the share of industry i' intermediate sales directed to j, is utilized in AHS.

[38] I use a measure of relationship-specificity developed and used in Rauch (1999). This method classifies commodities by whether they are sold on organized exchanges, have a reference price in a trade publication, or neither. By reflecting the depth of the potential market for the commodity, the degree of potential holdup problems is inferred. Following Nunn (2007), I create a dummy variable indicating whether a good falls into the first two categories. Rauch's classification that was revised in 2007 groups goods into 1,189 SITC Rev.2 industries. I use official concordance tables to match them with appropriate KSIC or NAICS codes of each firm.

[39] I use 3-digit NAICS and 2-digit KSIC codes in order to maintain a similar level of disaggregation.

The second set of predictors measures the size of firms and groups. Both AHS and Ramondo, Rappoport and Ruhl (2016) report that a firm is more likely to engage in related-party trades when that firm is larger and when it belongs to a larger group. I use firms' total assets and total sales, and the number of firms within a group as possible predictors. The relatively parsimonious nature of the variables in this set have an added benefit for future use, as total assets and sales are often the most widely available information in firm-level databases.[40]

The last set of predictors is loosely defined as a firm's 'control' over the related-party in trade. I include an indicator of whether the reporting firm controls a *majority* of the related party's voting power; in other words, whether the related party is a subsidiary of the firm. In contrast, control over the firm's affiliates with *minority* vote-holding, or other related parties such as parent companies, parent companies' other subsidiaries, etc. are deemed weaker and are therefore distinguished with this dummy variable. In a similar vein, I also include an indicator showing whether the related party is domestic or foreign. This partly reflects the firm's possible control and supervision over the affiliate's activities (Antras and Foley, 2015); at the same time, it is expected to pick up differences in domestic vertical ownership and FDI.

Compared to Section 3, a smaller subset of firms is utilized here to ensure that for each firm there is a full list of its related parties. For all publicly traded firms in Korea, TS2000 reports the list of their related parties. I merge the primary related-party trade dataset with the list from TS2000, keeping only those firms that appear in both datasets. The consolidated data contains the complete list of all public firms' related parties, as well as how much each firm trades with them. As a result, I utilize a total of 2,607 firms over 2013–2019 that have 420,428 firm-year-related party triples. Among the triples, 105,367 (25.1%) report transactions in terms of either sales, purchases, loans, or debts, while the rest do not trade with the reporting firms. The main body of this section will primarily report the algorithms trained with the firms in the manufacturing sector only. This leaves 1,600 firms over the same data period, with 197,021 firm-year-related party triples. Among the triples, 56,158 (28.5%) report intra-party transactions.

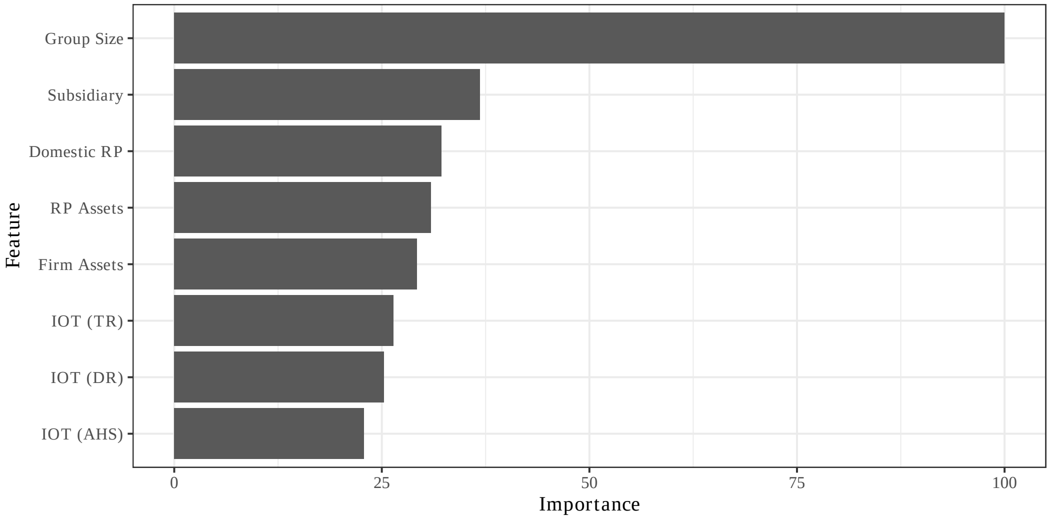
The data for each year is then randomly divided into training and testing sets according to an 80:20 split at the firm level. The principle is to have no overlapping information between the two sets: thus, no reporting firm will appear in the same year's training set *and* testing set. I train each year's prediction algorithm separately using the `R` package `caret`.

B.3 Alternative Measures of Prediction Performances

In section 5, I report the algorithm's performances mainly using three measures: *accuracy*, *precision*, and *recall*. In this section, I show that the algorithms fare well in other key metrics that

[40] Total assets and sales are matched to firms from existing databases, following the process described in section 2. However, not all related parties in the data are successfully matched with assets and sales information. These cases are primarily driven by firms that are too small and thus not included in the databases at hand. Therefore, missing values are assigned 0, and a dummy variable indicating imputation is added.

Figure 2: Importance of Top Variables, 2019 Algorithm



are widely used in the prediction literature. Appendix Figure 3a plots the Receiver Operating Characteristic (ROC) curve from 2019. This curve plots the tradeoffs between true positive rates (*recall*) and false positive rates ($1 - specificity$) as the threshold becomes more relaxed for declaring a pair to be trading. As the method becomes more lenient in declaring a pair to be trading, any classification approach detects more true positives and produces higher recall. At the same time, it is more likely to misconstrue non-trading pairs as trading, and have higher false positive rates. Encapsulating this curve, the AUC score calculates the *Area Under the (ROC) Curve* which provides a measure of the estimated probability that a positive case, in this case a trading pair, will be ranked higher by the algorithm than a negative case (Hosmer Jr, Lemeshow and Sturdivant, 2013). As column (6) of Table 3 reports, the AUC scores consistently report a strong result.

When the positive cases are scarce such that there is a large class imbalance, AUC scores could be overly optimistic (Davis and Goadrich, 2006). In this case, the PR-AUC scores are often used to evaluate the model performances. This score calculates the area under Precision-Recall curves, plotted in Appendix Figure 3b, which shows the tradeoff between *precision* and *recall* as the predictive threshold changes. While the class skew is not strong in this paper, with 27.8% trading, that PR-AUC scores are consistently strong sufficiently addresses any possible concerns.

B.4 Relative Variable Importances

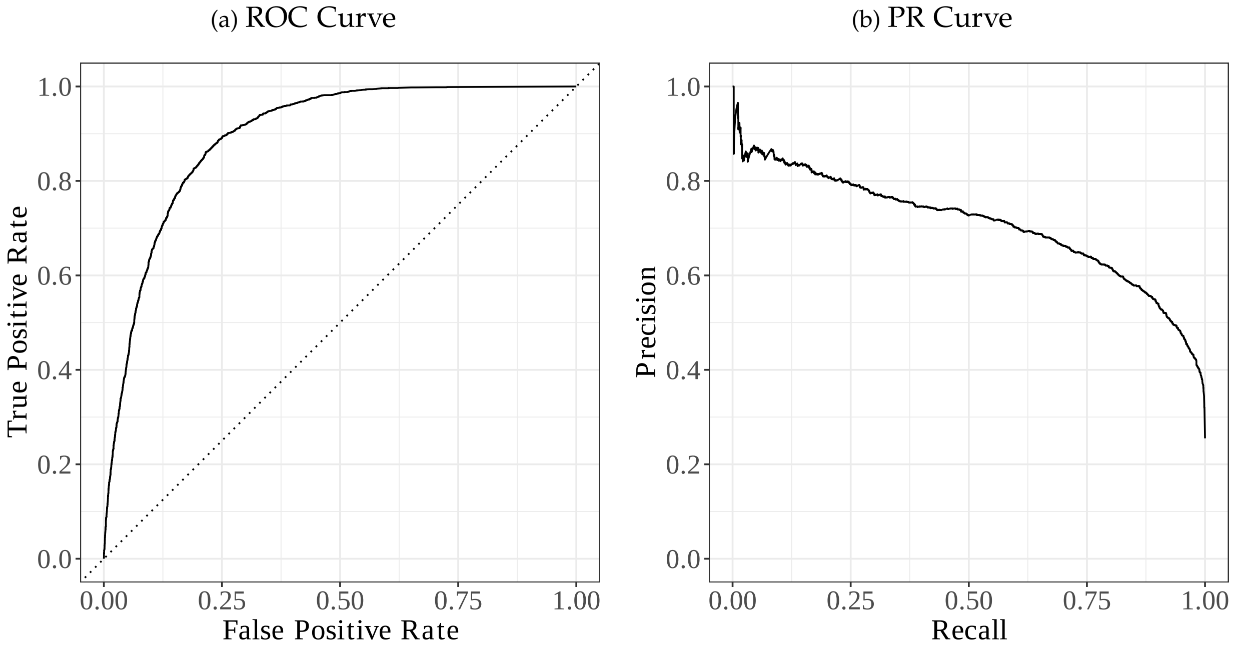
In addition to reporting algorithm performance, I report the top 8 predictors in terms of variable importance in Figure 2. The importance is calculated based on how information from each variable decreases mean node impurity: this is closely analogous to the residual sum of squares in regression. The most important variable is given the index value of 100, while other variables are given values in relative terms to it. The most notable trait from the variable importance plot is the large importance of group size. While Ramondo, Rappoport and Ruhl (2016) has shown that group size could be a significant predictor of related-party trades, that its relative importance far surpasses the other features is surprising.

Moreover, the IOT coefficients, although important, nevertheless appear far from the most decisive predictors. Other variables that represent reporting firms' control over the related party are shown to be more important, such as whether a majority of the related party's voting rights are owned by the firm (*Subsdidiary*) or whether the related party is located domestically. Then follow the sizes of both firms in terms of total assets, and only lastly the input-output coefficients.

C Supplementary Tables and Figures

C.1 Figures

Figure 3: Performance of Algorithm based on 2019 Data



(a) ROC Curve

(b) PR Curve

C.2 Tables

Table 8: List of Industries with Prior of High RPT Use from Lafontaine and Slade (2007)

KSIC Codes	Industry Names
0510	Mining of coal and lignite
1120	Manufacture of ice and non-alcoholic beverages; production of mineral waters
1521	Manufacture of footwear
1921	Petroleum refineries
2011	Manufacture of basic organic chemicals
2331	Manufacture of cement, lime and plaster
3031	Manufacture of parts and accessories for motor engines (new products)
3032	Manufacture of parts and accessories for motor vehicle body (new products)
3033	Manufacture of power transmission devices and electrical and electronic equipment for motor vehicles (new products)
3039	Manufacture of other parts and accessories for motor vehicles (new products)
3040	Manufacture of parts and accessories for motor vehicles (remanufacturing products)
3132	Manufacture of engines and parts for aircraft
0610	Mining of iron ores
1711	Manufacture of pulp
3111	Building of ships and floating structures

NOTE.—The original list of industries are from Lafontaine and Slade (2007), and this table outlines the corresponding 4-digit KSIC industries and their codes used in this paper.

Table 9: Share of Sales to Related Parties, Firms in All Industries

Related Party Share	Percentiles (%)				Fraction (%)		Weighted	
of Firm Total Sales	50th	75th	90th	95th	=0	≥1	Mean (%)	*N*
Panel 1: Main Result - Share at Reporting Firm Industry-Level								
All trades	1.3	15.1	62.4	95.8	30.3	3.0	24.0	121,519
IOT Requirement	0.1	3.5	24.4	52.5	36.7	0.7	9.7	52,298
AHS Requirement	0.0	1.9	19.1	50.5	42.7	0.8	6.4	35,955
Intensive Margin Only	4.8	30.2	86.3	100.0	19.3	4.7	6.8	52,696
Panel 2: All Firms with a Related Party in the ORBIS Database								
All trades	2.2	20.9	74.3	99.1	28.2	3.9	25.6	62,147
IOT Requirement	0.0	3.2	22.7	51.2	37.0	0.6	10.1	33,354
AHS Requirement	0.0	0.7	12.6	42.4	50.1	0.6	6.5	18,087
Intensive Margin Only	5.3	31.6	86.8	100.0	19.3	4.9	27.7	33,649
Panel 3: No Related Party without Industry Information Vertically Related								
All trades	1.3	15.1	62.4	95.8	30.3	3.0	24.0	121,519
IOT Requirement	0.0	0.3	10.9	36.1	66.3	0.5	5.3	52,761
AHS Requirement	0.0	0.4	11.9	40.8	63.8	0.6	3.7	36,197
Intensive Margin Only	4.8	30.2	86.3	100.0	19.3	4.7	6.8	52,696
Panel 4: All Related-Parties without Industry Information Vertically Related								
All trades	1.3	15.1	62.4	95.8	30.3	3.0	24.0	121,519
IOT Requirement	0.1	5.3	31.5	65.9	40.5	1.3	15.0	107,028
AHS Requirement	0.0	2.5	19.3	49.4	43.4	0.8	8.1	78,090
Intensive Margin Only	1.7	17.1	66.4	97.1	28.1	3.2	24.5	107,017
Panel 5: All Sales to RW industries as Vertically Related								
All trades	1.3	15.1	62.4	95.8	30.3	3.0	24.0	121,519
IOT Requirement	0.2	5.2	28.3	57.9	32.2	0.8	12.9	65,461
AHS Requirement	0.1	3.4	23.4	56.3	35.3	0.9	7.5	46,205
Intensive Margin Only	4.1	26.7	81.6	99.8	19.4	4.2	25.9	65,909
Panel 6: Reporting Firms in All CFS industries								
All trades	2.5	16.8	56.7	90.6	22.3	2.1	31.3	66,126
IOT Requirement	0.5	6.9	32.8	63.4	27.2	0.9	15.0	29,455
AHS Requirement	0.1	4.1	28.3	64.9	35.1	1.1	9.1	19,558
Intensive Margin Only	6.6	30.2	80.6	99.2	13.2	3.6	36.2	29,661
Panel 7: Related-party Purchases								
All trades	3.1	20.0	62.3	95.5	26.5	4.2	27.0	105,563
IOT Requirement	0.2	5.5	21.5	39.7	33.2	0.5	11.2	37,944
AHS Requirement	0.0	3.0	16.0	33.1	41.6	0.5	5.9	29,928
Intensive Margin Only	7.7	29.9	71.1	99.7	14.6	4.9	30.6	38,110
Panel 8: Reporting Firms in Industries with Prior of High RPT Use								
All trades	5.0	27.4	79.2	99.4	20.8	3.7	36.9	7,767
IOT Requirement	4.6	25.6	72.4	94.4	21.3	2.5	23.2	5,347
AHS Requirement	1.6	15.9	63.2	94.4	27.3	2.4	13.5	4,203
Intensive Margin Only	10.5	42.1	93.6	100.0	13.2	5.3	38.0	5,347

NOTE.—The table reports the share of firms' sales that is sold to related parties. Sample *N* consists of firm-years that have at least one related party, regardless of whether the reporting firm has any related-party transactions: due to missing firm-level information, total *N* used in the table is slightly smaller than the 125,044 available firm-level reports. The weighted mean is weighted by the size of each reporting firm's total sales.

Table 10: Comparison of Related-Parties with vs. without Industry Information

RPT	Info	$\text{Share}_{>0}$	q_{25}	q_{50}	q_{75}	q_{100}	Mean
Sale	Yes	0.642	0	20.0	548.0	247,063.0	3,211.6
	No	0.538	0	1.8	281.3	150,477.3	1,850.5
Purchase	Yes	0.614	0	15.1	540.5	190,664.2	2,788.5
	No	0.495	0	0	255.0	126,707.5	1,608.5

NOTE.—This table compares trades with related parties that are (i) successfully matched to industries and (ii) not, using pooled observation throughout 2013-2019. For each of the two groups, $\text{Share}_{>0}$ reports the share of firm-year-related party triples that have a positive amount of each type of trade. q_n reports the size of the trades for the n-th percentile. As q_{25} is already at the smallest possible value of zero, q_0 is omitted from the table. Columns 4-8 use Millions of Korean Wons as units. The table reports the numbers after truncating 0.5% of observations from each end.

Table 11: Performance: Predict 2019 Intra-party Trades

Year (1)	Accuracy (2)	Precision (3)	Recall (4)	Specificity (5)
2019	0.797	0.711	0.627	0.878
2018	0.906	0.828	0.807	0.940
2017	0.883	0.806	0.722	0.939
2016	0.859	0.758	0.687	0.921
2015	0.846	0.740	0.633	0.922
2014	0.825	0.709	0.602	0.908
2013	0.825	0.709	0.552	0.921

NOTE.—Metrics for 2013–2018 represent the performances of each year's algorithms in predicting intra-party trades of 2019. In doing so, I utilize all firms in 2019 that have not been used in the training sets of each year: in this way, there are no overlaps between training and testing datasets. The metrics for 2019 represent the out-of-sample performances of the 2019 algorithm.

Table 12: Performance: Predict 2013 Intra-party Trades

Year (1)	Accuracy (2)	Precision (3)	Recall (4)	Specificity (5)
2019	0.825	0.692	0.708	0.873
2018	0.825	0.696	0.733	0.864
2017	0.845	0.712	0.756	0.880
2016	0.842	0.723	0.763	0.876
2015	0.871	0.776	0.774	0.910
2014	0.887	0.801	0.830	0.911
2013	0.768	0.668	0.604	0.850

NOTE.—Metrics for 2013–2018 represent the performances of each year's algorithms in predicting intra-party trades of 2019. In doing so, I utilize all firms in 2019 that have not been used in the training sets of each year: in this way, there are no overlaps between training and testing datasets. The metrics for 2019 represent the out-of-sample performances of the 2019 algorithm.

Table 13: Out-of-Sample Confusion Matrix, ML Algorithm from 2019 Data

		Reference	
		Trade	No Trade
Prediction	Trade	432	176
	No Trade	257	1,272

Table 14: Prediction Performance Metrics: 2013-2019 Algorithms (target: AUC)

Year	Accuracy	Precision	Recall	Specificity	AUC	PR-AUC
(1)	(2)	(3)	(4)	(5)	(6)	(7)
2019	0.797	0.711	0.627	0.878	0.861	0.708
2018	0.816	0.694	0.639	0.887	0.889	0.719
2017	0.770	0.678	0.547	0.876	0.848	0.699
2016	0.755	0.663	0.564	0.853	0.829	0.695
2015	0.705	0.660	0.595	0.783	0.779	0.712
2014	0.820	0.672	0.660	0.880	0.881	0.734
2013	0.768	0.668	0.604	0.850	0.829	0.730

NOTE.—This table reports prediction performance metrics of algorithms created from each year's training data, tested on the same year's out-of-sample testing dataset. To create this table, algorithms that optimize the ROC AUC scores were utilized, while Table 3 uses algorithms that optimize PR-AUC scores. Information from all manufacturing public firms in Korea and their related parties are utilized.