Insights from firm-to-firm transaction data: Reassessing buyer-seller networks*

Alejandra Martinez

July 2025

PRELIMINARY AND INCOMPLETE: PLEASE DO NOT CITE

Abstract

The ability to access administrative data has enabled researchers to study and understand the nature of trade relationships. As more studies utilise firm-to-firm transaction data, they rely on the few countries that collect and allow access to it. In cases where firm-to-firm data are available on both sides of the transaction, foreign firms often lack unique identifiers, resulting in researchers relying on text algorithms to generate them, often leading to over-counting the number of foreign firms (Krizan et al., 2020). In this paper, I document systematic reporting patterns on foreign firm names that can facilitate researchers' efforts to develop unique identifiers from firm-to-firm administrative data. I use Colombian customs data, which provides the names of foreign firms and is easily accessible. I present long-term patterns of network statistics, which illustrate how networks changed over time. Finally, I discuss the limitations of using random matching models of link formation to predict relevant distribution statistics.

JEL Classification: C81, F14, L14 , D85
Keywords: Data Management, Firms, Networks, Trade

* Martinez: School of Economics, University of Nottingham, Nottingham, UK and CITP; e-mail: alejandra.martinez@nottingham.ac.uk.

1 Introduction

Having access to firm-to-firm transactions allows for a deeper understanding of firms engaged in international trade. By accessing firm-to-firm administrative data, researchers have shed light on the understanding of production networks, primarily through the use of empirical observations to develop theoretical models of buyer-seller link formation.[1] Access to firm-to-firm administrative data remains limited because only a few countries systematically record flows of foreign firms, or if they do, make it publicly accessible. As a result, empirical research on production networks is limited to a few contexts, and is contingent on government agencies' ability to record transactions systematically, or the red tape required to access the data.

The use of customs administrative data for empirical analysis remains challenging, even when firm-to-firm data are available. Significant discrepancies in total trade figures often emerge when researchers conduct mirror reviews (comparing data from both sides of the trade flow). These reviews have identified several key factors contributing to discrepancies in disaggregated and aggregate data records: poor statistical systems, misinvoicing, measurement errors, shipment lags, unintentional underreporting, and evasion through smuggling (Benita and Urzúa, 2016; Liu, Wheiler, Ganguly, and Hu, 2020; Vincent, 2004).[2] But, beyond discrepancies in reported values (intensive margin), analysis of individual firm-to-firm data has revealed less documented sources of discrepancies in accounting for the number of firms trading (extensive margin). Foreign firms typically lack systematic identifiers within customs datasets, forcing researchers to create identifiers from available information, potentially introducing errors. Limited access to firm-to-firm customs records hinders researchers' ability to cross-check and document these discrepancies through mirror exercises, obscuring potential impacts on aggregate statistics.[3]

Accurate trade statistics between countries are crucial beyond mere accountability. Discrepancies in trade flows can disproportionately affect small, trade-oriented open economies (Sen, 2000). And even if developed countries (e.g., OECD members) have improved their accounting accuracy more rapidly than developing nations (Makhoul and Otterstrom, 1998), accurate trade accounting remains relevant for most countries, as trade imbalances and bilateral deficits have tangible consequences for trade policies (Feenstra, Hai, Woo, and Yao, 1999).[4] As firm-to-firm data becomes increasingly available, researchers must consider the limitations of these datasets and account for potential issues in disaggregated and aggregate statistics. Moreover, policymakers need to make more informed decisions about future data collection systems to properly assess the impacts of trade policies.

For researchers to address discrepancies in the number of foreign firms, access to firm-to-firm transactions from both the countries of origin and destination is necessary (i.e.,

[1] See Bernard and Moxnes (2018) for a literature review on production networks.

[2] See Hamanaka, 2012 for a review of mirror comparison exercises.

[3] Using aggregate data, Helpman, Melitz, and Rubinstein (2007) demonstrate that traditional estimates of intensive and extensive trade margins are biased, primarily due to the omission of the extensive margin rather than selection bias.

[4] In 2018, the US initiated a trade war with China, largely based on the US-China trade balance. See Fajgelbaum and Khandelwal (2021) for a review of research following this policy.

mirror review). Krizan, Tybout, Wang, and Zhao (2020) is one of the few studies that documents discrepancies when merging individual buyer-seller transactions from both sides of customs reports. They utilise administrative firm-to-firm data for Colombian exporters and US importers. The authors construct a buyer-seller network by cleaning and harmonising Colombian export records to identify importers based on information about the reported foreign buyer, their final addresses, and the transaction value. They find that the number of US importers is twice as high compared to official reports (when using US data that contains the actual number of domestic firms). Reversing the analysis and using US customs records along with Colombian firm identifiers constructed by US authorities, they arrive at a network that also yields twice as many Colombian exporters. The key takeaway is that, in the absence of harmonised firm identifiers between different customs agencies, reported names or IDs of foreign counterparts often result in an overcount of foreign firms compared to what is observed using harmonised firm identifiers from domestic authorities.

While Krizan et al. (2020) are the only ones to document discrepancies found in customs records—on the number of foreign firms—they do not provide guidance on how to use these datasets to alleviate such discrepancies. In this paper, I document systematic patterns in the recording of foreign names within Colombian customs data. To the best of my knowledge, this is the first study to document sources of discrepancies in the implicit content of foreign buyers' names reported in customs data, with the aim of obtaining a more accurate count of foreign buyers. Colombian customs records are relatively easy to access compared to other administrative datasets, so understanding how to utilise these data and address their limitations can be extremely valuable. I use data from 2007 to 2019, covering almost all available years where information on the foreign buyer is included. After correcting patterns that may lead to over-counting the number of foreign firms, I construct a production network for a specific sector and compare the observed network patterns with those documented in other contexts. A limitation of this paper is the lack of access to US data, which prevents verification of the accuracy of the buyer count.

Using the newly constructed buyer-seller network for the flower sector between US and Colombian firms, I illustrate changes in this network over time. First, I replicate the main stylised facts of production networks for the entire panel of my data, examining key features such as the density, firm size distribution, and degree assortativity. I find that these patterns are qualitatively similar to those reported in other empirical studies. However, by extending the analysis over a longer time span, I document significant changes in the network's sparsity, shifts in the degree distribution of buyers and sellers, and relative changes in firms' connectivity over time. These dynamics are often overlooked when focusing solely on cross-sectional statistics, underscoring the importance of longitudinal analysis in understanding the evolution of buyer-seller networks.

In section 2, I describe the customs data and its variables, providing an overview of the sources of discrepancies that can lead to systematic errors in counting the number of foreign firms. I find that one possible source of these discrepancies is the failure to clean and separate the information in the reported buyer name. I classify transactions based on whether the additional information reported is useful for identifying the firm.

One percent of the transactions contain additional information that is not useful, while six percent include information that can be used to identify a particular firm. The second issue is the absence of an actual firm name in the relevant variable. This could involve other types of information being reported instead of the firm's name, or the name may not be reported at all. Some sectors experience these reporting issues, while others do not. This indicates that, for certain sectors, creating a buyer-seller network may not be advisable since the majority of reported buyer names do not correspond to a firm, which prevents the creation of a unique identifier. Coffee (HS 09), mineral fuels (HS 27), and miscellaneous edible preparations (HS 21) account for 97% of the transactions with no reported buyer name.

In Section 3, I focus on the flower sector (HS 06), which accounts for 81% of transactions and 31% of trade value, to construct a buyer-seller network from 2007 to 2019. I clean the reported buyer name variable with a focus on the content of the text. Despite employing standard text cleaning procedures, I find numerous transactions with unusually long reported buyer names. Upon closer examination, I find that this is due to the reporting of multiple buyer names within a single transaction. This practice is systematic across sectors and years, with approximately one-third of all reported buyer names containing multiple firm names. I then construct the buyer-seller network using only transactions with a single reported buyer.

In the flower sector, approximately one-third of buyers engage in transactions with other buyers. Such multiple-buyer transactions are particularly common among larger and more interconnected firms, which account for 70% of total trade. If a foreign firm identifier is constructed without distinguishing between multiple buyers within a transactionoften relying on machine learning algorithms to standardise firm namesresearchers may overestimate the number of buyers. This misidentification can lead to several biases in network statistics, including: (1) underestimation of trade per firm, as importers appear to have fewer links and, consequently, fewer purchases; (2) undercounting of connections per buyer while overcounting those per seller, as more buyers will decrease the average number of connections per firm while increasing those of sellers. These inaccurate first-order network statistics can also affect higher-order statistics such as the importers' degree assortativity, which may appear steeper due to the inflated buyer count. Therefore, it is crucial for future researchers interested in constructing buyer-seller networks when using similar data to consider the additional information provided by the reported buyer names.

With the buyer-seller network, I reproduce some of the commonly documented facts in production networks. I do this for every year between 2007 and 2019, relying on the five stylised facts described in Bernard and Zi (2022): sparsity, heterogeneity, market access, degree assortativity, and hierarchy. Because of the sample I use, it is not possible to contrast these stylised facts quantitatively with other production networks; however, I compare them to the qualitative characteristics observed in other cross-sections. The findings for this buyer-seller network align with observations in other contexts: well-connected firms tend to engage in larger transactions, form matches with less well-connected firms, and access a broad range of markets. However, I also study changes over time in the network distribution. Specifically, the increase in connections observed between 2007 and 2011

was partially driven by a rise in links among the most well-connected sellers, compared to those with median-connected sellers. In contrast, the increase in connections between 2011 and 2015 was driven by a growth in connections among median-connected sellers relative to the well-connected ones. The analysis shows that while theoretical models of link formation based on cross-sectional network statistics may be essential for modelling general network patterns, they may be inadequate for capturing shifts in the distribution of connections over time.

In Section 4, I test how well random matching models, designed to fit stylised facts, align with the network. I estimate a generalised balls-and-bins model based on the framework by Armenter and Koren (2014) and the firm-level adaptation in Bernard and Zi (2022). The predictive power of the generalised balls-and-bins model is evaluated by comparing it with standard trade models (such as gravity) and flexible specifications based on both bins and firm fixed effects. The generalised balls-and-bins model performs worse than gravity models and other more flexible specifications. I re-scale the estimations of the generalised balls-and-bins model to align the predicted number of links with the data and assess the fit of other network statistics. I examine differences in the in-degree and out-degree distributions of firms between model draws and the observed data. The generalised and re-scaled balls-and-bins model predicts the distributions qualitatively; it overestimates the number of connections in the top percentiles but performs better in the middle of the distribution.

This paper relates to the literature on trade misreporting. There is a larger documentation on examining agents' motivations for under-reporting trade. Various theoretical frameworks have been developed to predict tax rates and evasion, building on the work of Allingham and Sandmo (1972), but the predictions of these models are sensitive to their underlying assumptions (Slemrod and Yitzhaki, 2002). Since evasion cannot be directly observed, empirically quantifying it has proven challenging. To better quantify the tax-evasion relationship, researchers require highly disaggregated data and reports from both the origin and destination of transactions. The 'evasion gap,' as proposed by Fisman and Wei (2004), varies between 0.1% and 3% across different empirical studies. Differences in quantification effects, such as higher evasion rates, can be explained by using data flows with inherently different characteristics, such as differentiated products in the sample.[5] Additionally, cross-country evidence suggests that factors such as corruption, auditing standards, and trade agreements also influence trade misreporting (Javorcik and Narciso,

[5]Fisman and Wei (2004) measure evasion in China's imports from Hong Kong using product data and find that the 'evasion gap' is correlated with Chinese tax rates. Specifically, they find that a one-percentage point increase in the combined tariff and VAT rate is associated with a 2-3% increase in evasion. Mishra, Subramanian, and Topalova (2008) rely on 6-digit product data and exploit a tariff reform in India in 1990, finding a 0.1% increase in evasion. This lower magnitude is reconciled with a product sample that is less biased towards differentiated goods. Stoyanov (2012) use US-Canada data and exploit the FTA of 1989, finding a strong relationship between the evasion gap and tariff rates. Specifically, an additional percentage point in the tariff rate reduces the value of reported imports by 3-5% in the US and by at least 1% in Canada. Ferrantino, Liu, and Wang (2012) examine Chinese and US trade flows and find strong statistical evidence of under-reporting of exports at the Chinese border to avoid VAT. They also find indirect evidence of transfer pricing, i.e., over-reporting at the US border to avoid higher US corporate income tax for US-based multinationals, and avoidance of Chinese capital controls, i.e., money laundering.

2017; Kellenberg and Levinson, 2019). Although this research focuses on utilising disaggregated firm-to-firm data to highlight channels behind underreported trade, my study also employs firm-to-firm data to assess the nature of misreporting. However, I do not focus on agents' motivations for such misreporting.

Discrepancies in trade statistics are often attributed to methodological differences between statistical agencies, as each agency may edit trade data according to its own procedures (Stoyanov, 2012). By focusing specifically on the discrepancies observed in the data, this paper complements the work of Krizan et al. (2020), who highlight the magnitude of the discrepancies, but do not identify their causes or propose solutions. This paper describes specific sources of mismatches in the identification of foreign buyers that can lead to over-counting when relying solely on exporters' customs records.

Customs data from Colombian importers and exporters are readily accessible and offer the advantage of reporting foreign buyer and seller names. Consequently, much empirical work using firm-to-firm data relies on Colombian customs records, with importers' customs records being the most commonly utilised source of trade flows. For instance, Blum, Claro, and Horstmann (2012) construct a buyer-seller network linking Chilean exporters with Colombian importers, while Benguria (2022) builds a network between Colombian importers and French exporters. Similarly, Bernard, Bøler, and Dhingra (2019) use the complete dataset of Colombian importers to develop a buyer-seller network spanning from 1995 to 2014, and Bernard and Dhingra (2019) analyse Colombian importers in trade with the US from 2009 to 2014. In contrast, Eaton, Eslava, Jinkins, Krizan, and Tybout (2021) focus on Colombian customs records for exporters, highlighting significant discrepancies between the LFTTD data (US flows) and the DIAN (Colombian data) regarding the number of reported foreign firms. Building on this research, I document methodologies for handling Colombian customs records to construct a buyer-seller network that mitigates some of the discrepancies in the reported number of foreign firms.

With more firm-to-firm data becoming available, there is a growing literature on domestic and international production networks. Several studies have focused on empirical regularities in buyer-seller relationships and the evolution of production networks, including Blum, Claro, and Horstmann (2012) for Chilean and Colombian firms, Atalay, Hortaçsu, Roberts, and Syverson (2011) for the US, Kramarz, Martin, and Mejean (2020) for France, Carballo, Ottaviano, and Volpe Martincus (2018) for Costa Rica, Uruguay, and Ecuador, and Lim (2018) for the US. Much of this work has established stylised facts that help to inform models aimed at understanding the nature of buyer-seller connections and testing predictions. For example, Chaney (2014) documents future and current export destinations for French firms, while Bernard and Moxnes (2018) examine the roles of buyers and sellers and their adjustments to shocks. Sugita, Teshima, and Seira (2023) explore the impact of removing the Multi-Fibre Arrangement on Mexican textile exports to the US. Additionally, Bernard, Bøler, and Dhingra (2019), using Colombian importers, document stylised facts including trade margins in firm-to-firm relationships, buyer-seller churning over time, and changes in trade shares as relationships age, but do not address changes within the network itself. This paper examines long-term patterns and changes within the network over time.

Finally, this paper contributes to the literature on models of buyer-seller link formation that aim to align with empirical regularities. The balls-and-bins model proposed by Armenter and Koren (2014), and its adaptation to firm-level transactions by Bernard, Moxnes, and Ulltveit-Moe (2018), are examples of random allocation models. These models face challenges in accurately reflecting network statistics, as they often encounter allocation problems and impose additional assumptions that may be inconsistent with empirical contexts. In more recent work, Bernard and Zi (2022) introduce a model designed to characterise the organisational principles of production networks, demonstrating its utility as a benchmark for selecting informative statistics. Similarly, Herkenhoff, Krautheim, and Sauré (2021) offer a re-interpretation of the classical Krugman (1980) model, incorporating randomly bundled varieties to account for many empirical regularities. Sheveleva (2019) presents a statistical model with random product sizes, focusing on explaining the intensive profitability of multi-product exportersspecifically, the differences in sales of best and least-selling products among large and small exporters. This paper further contributes to the discussion by evaluating the extent to which the predictions of random allocation models align with the stylised facts observed in the buyer-seller network. Additionally, it demonstrates how these models can predict distributional network patterns over time.

The paper is organised as follows. Section 2 describes the Colombian customs data and documents the details on reporting of foreign buyers that can lead to over-counting the number of foreign firms in a network. Section 3 describes how I build the buyer-seller network for the Colombia-US flower sector, and replicates the stylised facts for all the years in the data. Section 4 estimates a random matching model of buyer-seller link formation, and tests the model predictions. Section 5 concludes.

2 Data

A description of the data is provided in this section, as well as a description of the challenges associated with grouping buyers into unique identifications based on reported buyer names from administrative firm-to-firm data. I use the universe of export transactions and indicate which sectors are not suitable for building buyer-seller networks.

2.1 Colombian customs data

To build the buyer-seller network, I use customs data from the Colombian National Directorate of Customs and Taxes (in Spanish, DIAN), focusing solely on export data. Since importer names are only available from 2007 onwards in customs records of exporters, I rely on the period from 2007 to 2019. The dataset includes 57 variables, including detailing the sellers (i.e., exporters) and buyers (i.e., importers), such as their addresses and names. For Colombian sellers, the data includes the unique identifier (NIT3) which corresponds to a tax number and refers to either a firm, a person, or a foreign entity. 99% of transactions are from domestic firms, which report their unique tax identification that can be matched to other external data sources. There are three pieces of information reported on the buyer's

side: the buyer's name (RAZN_IMP), the buyer's destination address (DIR_PDES), and the buyer's country of destination (COD_PAIS3).

2.2 Standardising reported buyer names from Colombian customs data

In the customs export records from DIAN there are about 5.9 million observations between 2007 and 2019.[6] The number of unique reported buyer names using the variable (RAZN IMP) is around 124K, which need to be standardised and corrected to identify the buyers across all years.

To construct and standardise the buyer name, several filtering and correction procedures are applied that are standard when dealing with this type of data. Utilising Colombian import data, Bernard, Bøler, and Dhingra (2019) and Bernard and Dhingra (2019) clean the reported seller names by eliminating non-alphanumeric characters, removing common prefixes, and employing machine learning algorithms to group likely spelling variants or misspellings. Similarly for the same data—at least in the Colombian side—Krizan et al. (2020) standardise names and typos, e.g., "st" to "street", and "corp" to "corporation", and in addition they also rely on the reported addresses and the transaction value to match with the US transaction data (LFTTD).

In the same spirit, I also remove some non-alphanumeric characters and standardise the reported buyer names for typos, as well as grouping similar words. While this step is necessary to begin the process of running text matching algorithms, it is not sufficient, as it still leaves the reported buyer names with other problems to solve before using them to build the network. The main problem is that about 6% of the observations (355K) seem to include additional information in the reported buyer names. The additional information is not random and cannot be purged easily, as it sometimes conveys useful information to identify and group different buyers.

After revising the reported buyers' names, I carefully proceed to clean those that contain additional information. First, I remove information related to errors in the reported buyer names, i.e., reporting addresses or tax identifications instead of a firm name. This first step helps to reduce the length of some reported names, which is problematic for the observations with additional information. For reported buyers' names that still contain additional information after the first cleaning run, there is a second step to be carried out. I split the remaining reported buyer names into multiple names using the different separation *keys* contained in the text. The separation *keys* can be non-alphanumeric characters e.g.,'/', a combination of non-alphanumeric characters and letters e.g.,'Y/O', key words, e.g., 'care of', 'dba', or in few cases, additional information that conveys the recipients of the purchase or shipping vessels.

Using the separation *keys* to divide the reported buyers' names into multiple names, I classify the observations with multiple names in two groups. The observations *Type 1* are those in which none of the names within the reported buyer names, provides additional information about the buyer. These texts most likely report the transportation mode, e.g.,

[6]The 5.9 million are a result of aggregating trade values per month from all the 12 million transactions into monthly observations with unique values for the rest of the variables in the data.

Master Ship, (M/S), or *Master Vessel*, or they can mention a particular person or recipient of the product e.g., *attention to*, *notify to*. If, on the contrary there is additional information that can be used to identify the buyer, I classify the observation as being of *Type 2*. These include multiple names that refer to the same buyer, i.e., *doing business as*, dba, a 'division of', or 'a branch of', or report another partner firm, that is involved in that particular transaction. The names within a reported buyer name from observations classified as *Type 2* are almost always separated by a non-alphanumeric character '/' or a combination of character and letter (C/O).

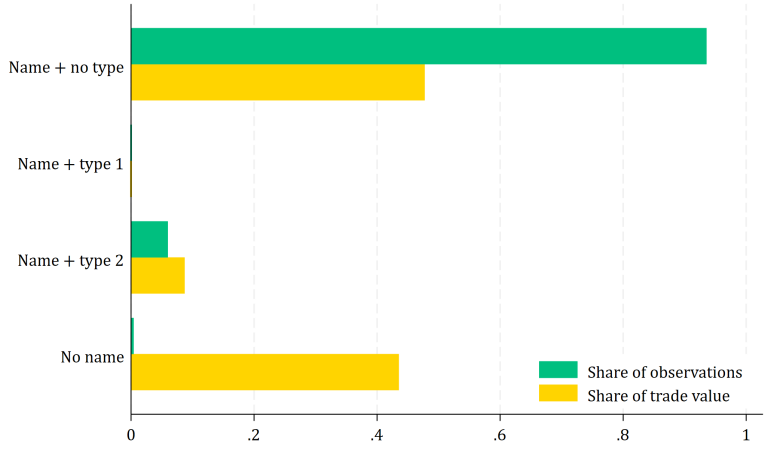
It is possible to have some cases reporting multiple names from either classification, i.e., a reported buyer name including both *Type 1* and *Type 2* names. In such cases, I classify the observation as *Type 2*, since I can obtain useful information from the multiple names. Within transactions classified as *Type 2* there are inconsistencies reporting the same exact firm names but with different separation *keys*. For example, a reported buyer name is 'Mega Labs (doing business as) Farmazona S.A', but in another observation, the same combination of names is written as 'Mega Labs Y/O Farmazona S.A'. In the first case one might think that the buyer is the same firm that can be referred to by either name, which is not unusual as firms also have pseudonyms. In the second case, this can be interpreted as being different buyers that are involved in the same transaction. When two names are used interchangeably, despite the separation *key*, I consider the names to refer to the same buyer; otherwise, I consider them to be different buyers.

There are some observations where the buyer name is not available. This can be a missing value in the reported buyer name variable, implicitly a text that reads 'not reported', or the reported name was a number or address without a firm's name. Figure 1 shows the share of observations in the data from 2007 to 2019 for different classifications based on the contents of the reported buyer name. In green, I count the participation using the 5.9 million observations, and in yellow, the trade shares. Regarding the share of observations, 93% have a name and no type, around 6% of the observations are *Type 2*, and only 0.02% of the transactions are *Type 1*. The remaining 0.4% of the observations do not have any reported name. Underneath in yellow, around 47% of trade value belongs to observations with a name and no type, about 9% to those with a name and are *Type 2*, 43% for those observations that do not report a name, and the remaining 1% are *Type 1*. Overall, most observations have one buyer name that can be identified, but that only accounts for half of the total trades.

Nearly half of the total trade is associated with observations that lack a reported name, raising concerns about how this might affect the representation of the network. To determine whether the absence of reported buyer names poses a significant issue for certain sectors more than others, I examine this aspect in greater detail. Figure 2 illustrates, in green, the proportion of observations within each HS section that include a name, out of the total 5.9 million observations within that section. The yellow bars beneath indicate the corresponding share of trade value.

Among the 21 sections, three stand out as having the highest proportion of missing buyer names: vegetable products (HS codes 06 to 14), prepared foodstuffs (HS codes 16 to 24), and mineral products (HS codes 25 to 27). In these sections, although nearly 80% of

Figure 1: Distribution of observations and trade value by buyer name classification



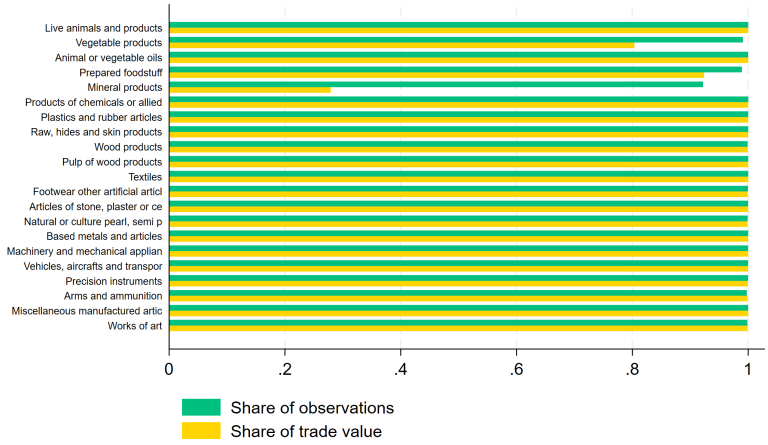
Notes: The figure plots the share of observations (blue) and share of trade (red) for each category. "Name + no type" refers to all observations where the reported buyer name is available and does not have additional information. *Type 1* refers to reported buyer names with additional information not useful for identification. *Type 2* refers to reported buyer names with additional information useful for identifying the buyer. "No name" refers to all transactions where reported buyer name is missing. Values of trade are in constant US dollars.

the observations include a name, a substantial portion of the trade value remains unidentified, with only about 30% of the trade value in the mineral products section associated with identifiable buyers. Notably, 97% of the trade value linked to missing names is concentrated in just three sectors: coffee, tea, and spices (HS code 09); miscellaneous edible preparations (HS code 21); and mineral fuels, mineral oils and products of their distillation, bituminous substances, and mineral waxes (HS code 27). When constructing an accurate network of Colombian exports, it is crucial to account for the limitations inherent in these three sectors.

I present the share of observations and trade for each section based on the ability to identify the buyer. Figure 3 depicts the share of observations with an identifiable buyer name in green and their corresponding share of trade value in yellow. The section 'vegetable products' has the highest proportion of observations (27%), followed by 'textiles' (17%). In terms of trade value, the most significant sections are 'minerals' (30%) and 'vegetable products' (14%). Within the two-digit HS classifications of 'vegetable products,' flowersclassified under live trees and other plants; bulbs, roots, and the like; cut flowers and ornamental foliageaccount for 81% of the observations and 35% of the trade value, while 'coffee, tea, and spices' account for 6% of the observations but 35% of the trade value. For the remainder of this paper, I focus exclusively on flowers (HS 06), one of Colombia's primary exports and the product with the most observations after excluding those without buyer names.[7]

[7]Eaton et al. (2021) exclude coffee and oil when constructing their firm-to-firm data, arguing that these flows are dominated by a few sellers. Indeed, the median number of sellers per year using the tax ID for the 2-digit sectors is around 131, while the coffee and oil sectors have approximately 209 and 201 sellers, respectively. The prevalence of missing names in these sectors presents a significant concern when including them in a buyer-seller network analysis.

Figure 2: Share of observations and total trade with reported buyer names within an HS section



Notes: The figure plots the share of observations (green) and share of trade (yellow) of observations with reported buyer names within an HS section.

3 The revised buyer-seller network

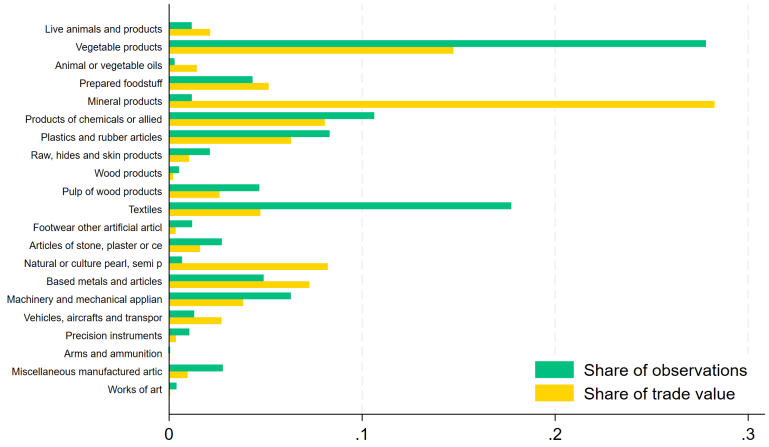
This section documents the construction of a buyer-seller network between Colombian flower exporters and US importers. I use the network to document five stylised facts about production networks (based on Bernard and Zi, 2022) over time.

3.1 Building a buyer-seller network from Colombian firm-to-firm data

The flower sector has the highest number of observations in the raw data (22%), and also has the most different reported buyer names from the variable RAZN_IMP than in other sectors. I harmonise individually each of the names obtained from the classification of observations into types, which, for some observations, generated multiple names and, for others, removed additional text. For each observation, I create a final harmonised buyer name or several final buyer names, ensuring consistency regardless of whether they are the only name or part of multiple names. Failing to account for multiple names in some transactions and potentially generating a single ID for such cases can lead to an overestimation of the number of buyers.

In the flower sector, approximately one-third of buyers are involved in transactions with other buyers, who account for about 70% of total trade value, and, on average, 38% of these buyers' connections also involve other buyers. Moreover, multiple buyers are more commonly associated with larger buyers, which disproportionately affects the overall number of connections and traded value in the network. By counting more foreign firms, it is likely to overestimate the number of connections per seller, underestimate the number of connections per buyer, and to overestimate the sparsity of the network (less connections overall). Figure A2.1 illustrates the differences between the 'non-corrected' and 'corrected' buyer-seller networks in terms of total number of buyers, seller connections,

Figure 3: Share of observations and trade value by HS section



Notes: The figure plots the share of observations (green) and share of trade value (yellow) for the HS sections, restricting to all observations with reported buyer names.

buyer connections, and total purchases. The deviations from the 45-degree line indicate there are discrepancies that arise when the correction for buyer names is not applied. In particular, the number of buyer connections and total purchases are the most problematic statistics when overcounting importers.

After harmonising the reported buyers' names I am left with 2,694 different names. I apply several filters to these observations and obtain the final count of both buyers and sellers. Table 1 shows the reduction in data when each filter is applied individually. The first row represents the baseline data used to create the network, which consists of 759k observations (from the original 5.9 million). The second row describes the first filter where I eliminate all observations without a reported buyer name. The second filter eliminates non-US final recipient addresses. This information is extracted from the variable DIR_PDES, from which it is possible to obtain a transaction's ZIP code. The third filter eliminates observations where the obtained final buyer name was counted only once throughout the entire period. These are around 466 buyer names for which there was no similar name to theirs that can be grouped with.

Since it is not possible to determine which buyer is the final buyer in 77,000 transactions involving multiple buyers, I consider only observations in which only one buyer is involved. The fourth filter eliminates all transactions involving multiple buyers. This filter has the greatest impact of all the filters, dropping about 200 buyers and almost 20% of the total trade value.[8] The fifth filter standardises inconsistent exporters' tax identifiers (64 exporters).[9] Customs data from DIAN do not exclude unpaid transactions, such as free

[8] An alternative network could be constructed by combining the consortium as a unique buyer. Due to the fact that these are separate firms, doing so might lead to an overcounting of the number of buyers. Another option would be to separate the buyers and create new connections. This splitting of the network would only work for extensive margins, as accounting for the trade value allocated to each buyer in that transaction would not be feasible.

[9] Some exporters report an additional digit on their exporter ID, so the last digit needs to be removed

samples. The final filter eliminates all transactions recorded with a trade value below $100 US dollars (or equivalently transactions with total weight below 1 kg). After applying all the filters, the final network contains most of the total trade value relative to the baseline (around 80%).

Table 1: Filters applied to firm-to-firm data to construct final buyer-seller network

#	Filter	# buyer	# sellers	# observations
	Baseline	2,694	1,347	759,030
1	Transactions with reported name	2,664	1,346	758,897
2	US Zip codes only	2,623	1,345	756,911
3	Buyer name repeats	2,157	1,324	756,443
4	Single buyer in transaction	1,957	1,307	679,406
5	Fix typos in exporter's tax ID	1,957	1,243	679,406
6	Transaction value <$100	1,833	1,260	662,900

Notes: The table reports the number of buyers and sellers for the different cleaning and filtering processes applied to the baseline data. The first filter eliminates those observations were buyers' names are missing or not reported. The second filter eliminates observations where buyers addresses are not to the US. The third filter eliminates observations were the buyers' names only appears once. The fourth filter eliminates observations were there is more than one buyer in a transaction. The fifth filter revises the tax ID of the sellers, and the sixth filter eliminates transaction with a value less than $100 US dollars (or less than 1 kg).

3.2 Key statistics on firm-to-firm trade

In this section, I present stylised facts about the buyer-seller network between Colombian flower exporters and US importers. The analysis focuses not only on the characteristics of buyers and sellers within the revised buyer-seller network but also on their evolution over time, encompassing all cross-sections from 2007 to 2019. Additionally, I compare the network statistics with cross-sectional analyses of the Belgian importers' network (Bernard et al., 2022) and the Colombian importers' network (Bernard, Bøler, and Dhingra, 2019) for the year 2014.

Fact 1. *Production networks are sparse.*

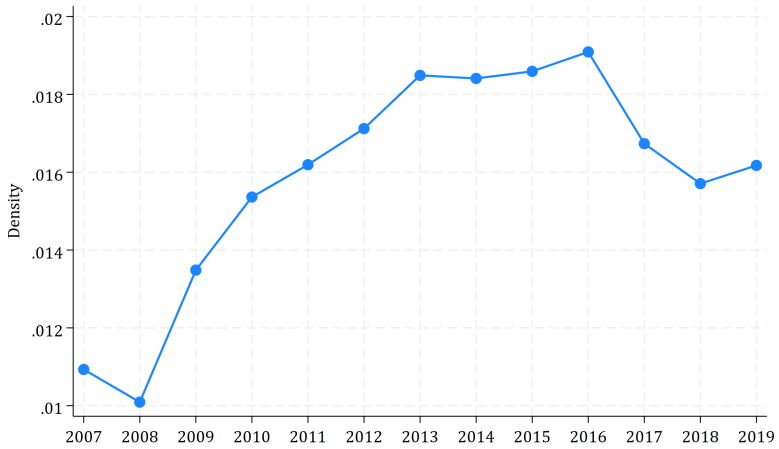
"In every production network examined to date, most buyers and sellers are not connected" (Bernard and Zi, 2022, p.7). In Japan's domestic network, Bernard, Moxnes, and Saito (2019) document that fewer than 1 in 30,000 potential buyer-sellers are active, while in Belgium, Magerman et al. (2015) document fewer than 1 in 23,000. For Colombia, Bernard, Bøler, and Dhingra (2019) find that around 1 in 15,000 buyer-seller connections exist between Colombian importers and foreign firms exporting to Colombia. Using the same cross-sectional sample as Bernard, Bøler, and Dhingra (2019), I find that approximately 1 in 80 potential buyer-seller pairs are active. This figure differs from those observed in other networks for several reasons. Firstly, I exclude links between firms from different

for those IDs. Second, typos were found in some of the IDs, which were corrected on an individual basis using the exporter name also found in the data and by checking the ID against the Chamber of Registration database.

sectors, focusing solely on connections between flower producers and consumers. Secondly, the network I constructed is a sub-sample that includes only Colombian exporters and excludes other countries that source from Colombia or sell to the US. Not including other countries to which Colombian exporters sell flowers is not a significant issue, as in 2014 the US accounted for 75% of the Colombian flower market, with domestic consumption constituting less than 5% of total flower production. This suggests that, from all sellers' perspective, the majority of trade value from buyer-seller connections is captured in the network.[10] For non-Colombian exporters to the US, the absence of US import records presents a challenge, given that in 2014, Colombian exporters supplied approximately 45% of flowers to the US.

In light of these limitations, it is more informative to compare the density within the network over time rather than against other datasets. Figure 4 illustrates the density for each cross-section of the data from 2007 to 2019, considering only buyers and sellers transacting each year and calculating the ratio of active connections to possible connections. From 2007 to 2019, the network has become relatively less sparse. In 2008, the network exhibited a density of approximately 1 in 95 connections. By 2015, this density had nearly doubled, with 1 in 50 connections active. However, in the latter part of the period, there was a decrease in connections, with a density of 1 in 58 active connections by 2019.

Figure 4: Evolution of the network density



Notes: The figure plots the density of the network in each year. Density is defined as the number of active connections divided by the number of possible connections. Each year, I use only Colombian exporters and US importers that are active (i.e., those with a positive transaction).

Fact 2. *Firms in a production network are heterogeneous in the number of links and value per link.*

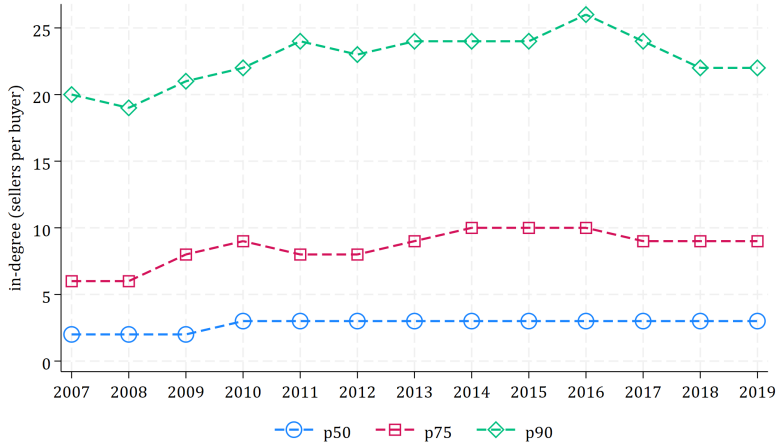
Several studies document heterogeneity in production networks, both in the out-degree distributions (number of buyers per seller) and in the in-degree distributions (number of

[10] *Sources*: Export shares from The Observatory of Economic Complexity. Domestic consumption data from Asocolflores (2016).

sellers per buyer). Carballo et al. (2018) reports that firms belonging to the 90th percentile of exporter connections in Costa Rica, Ecuador, and Uruguay can have between 6.5 to 11 times more foreign customers than the median exporter. Bernard, Bøler, and Dhingra (2019) also find a similar relationship for Colombian importers.

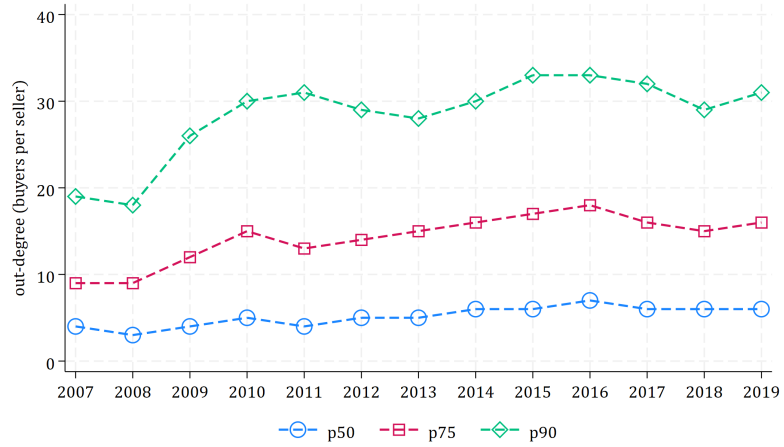
I plot the 90th, 75th, and the 50th percentiles for the in-degree, out-degree, and relationship value distributions over time. Figure 5 shows the in-degree, Figure 6 shows the out-degree, and Figure 7 the relationships values. The ratio between well-connected firms and the median is similar to that of other networks ranging from 6 to 11. Compared to other contexts, relationship values in the 90th percentile ratio to the median relationship is 17.

Figure 5: Buyers' in-degree for selected percentiles



Notes: The figure plots the average in-degree across all years (number of buyers per seller) for the 50th , 75th and 90th percentiles.

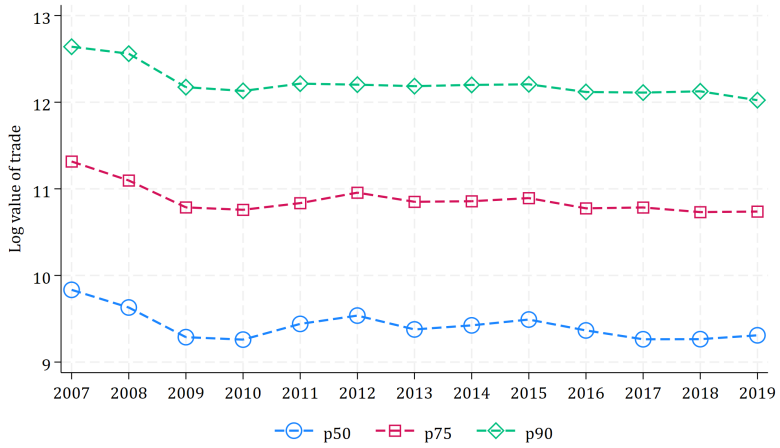
Figure 6: Sellers' out-degree for selected percentiles



Notes: The figure plots the average out-degree across all years (number of sellers per buyer) for the 50th , 75th and 90th percentiles.

Several trends stand out in the evolution of these distributions: (1) the number of connections for buyers (in-degree) and sellers (out-degree) increases over time, with sellers having the highest increase (50% more connections) and buyers having a lower increase; (2) over the entire sample well-connected sellers experience a disproportionate increase in connections relative to less well-connected sellers, while well-connected buyers see a decrease in connections relative to less well-connected buyers; (3) the average trade value of all relationships does not increase with the number of connections, in fact there is a sharp decrease during 2007-2010, followed by a period where the value remains constant.

Figure 7: Relationship log trade value for selected percentiles



Notes: The figure plots the log of trade value in US dollars (deflated) between buyer-seller relationships for the 50^{th} , 75^{th} and 90^{th} percentiles.

Fact 3. *The largest firms in terms of sales have the most buyers and suppliers and reach the largest number of markets.*

Evidence from Belgium (Magerman et al., 2015), Japan (Bernard, Moxnes, and Saito, 2019) and in Norway, (Bernard et al., 2018) show that largest firms in terms of sales are also the firms that reach and participate in the most markets.[11]

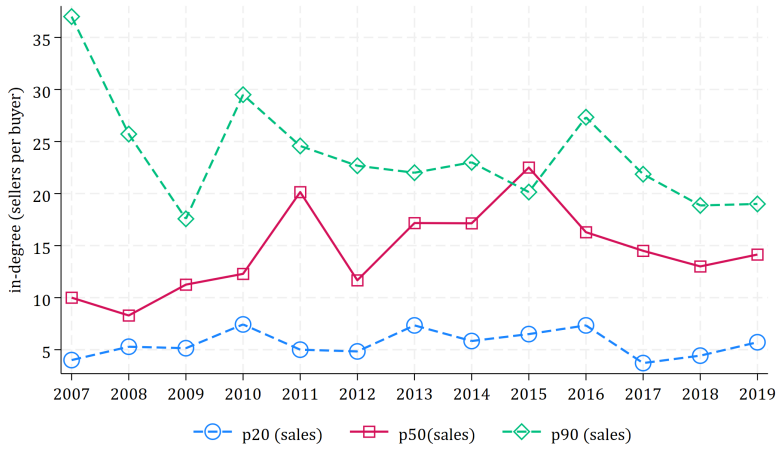
Figure 8 shows the buyers' in-degree for three percentile groups of firm-level sales (90^{th}, 50^{th}, and 20^{th} percentile). As documented in other contexts, largest buyers have the most connections. Excluding 2015, the number of connections for buyers in the 90^{th} percentile of sales are higher than those of the median firm, but the ratio p^{90}/p^{50} decreases over time, meaning large buyers become less well-connected relative to median-size buy-

[11]Fact 3 has implications for theoretical models. One-sided heterogeneity canonical models (such as in Arkolakis, 2010, Bernard et al., 2018, Eaton et al., 2019 and Lim, 2018) are not able to predict this relationship. With two-sided firm heterogeneity, Bernard et al. (2022) develop a theoretical framework of buyer-supplier production network and provide a micro-foundation for Fact 3. In their model, sellers with higher productivity have lower marginal costs and prices, which results in more buyers matching with them and greater total sales. Due to the more productive upstream suppliers, the buyers end up with lower marginal costs and more matches.

ers. Figure 9 presents the out-degree for each of the selected percentile groups of the sellers' sales distribution (90^{th}, 50^{th}, and 20^{th} percentile). The largest sellers (90^{th} percentile) start with as many connections as firms in the 50^{th} percentile, and in 2019 all sellers, regardless of size, have on average the same number of connections. Focusing on cross-sectional statistics alone would not reveal the dynamics observed between the first and last year of the sample, in which small and median sellers have in average more connections than large sellers.

I conduct additional analysis to reconcile the patterns from the seller-side, with what has been documented in other production networks. First, the low number of buyers can be due to not accounting for non-US connections. When counting the average number of non-US destinations for all three percentile groups, on average large sellers reach three more countries than smaller firms (see Figure A2.3). But when looking at their market participation, large sellers (90^{th} percentile of sales) export on average only 16% of their product to non-US countries, whereas those in the 20^{th} percentile export around 30% (see Figure A2.2). Even though large sellers reach more markets, most of their production is still sold in the US. Therefore, it is unlikely that not accounting for connections in other markets explains the generally lower number of connections. In this network, Fact 3 holds true for buyers, but does not always hold for sellers.

Figure 8: Buyers' in-degree (number of sellers per buyer) by firm size percentile

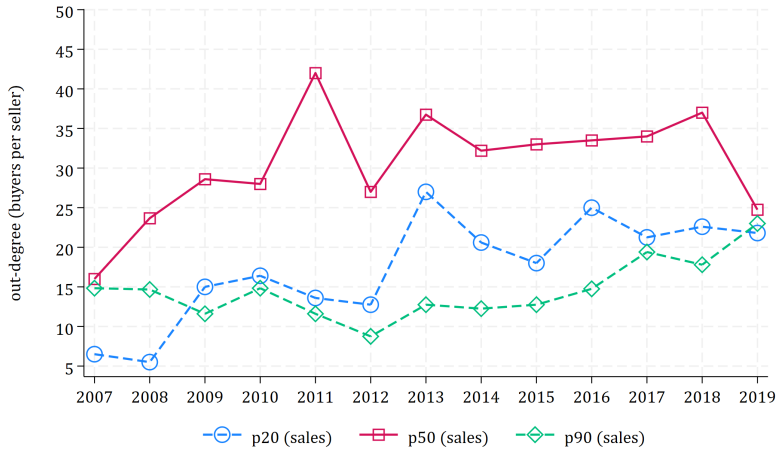


Notes: The figure plots the in-degree (sellers per buyer) for the 20^{th} , the 50^{th} and the 90^{th} percentiles of buyers' purchases.

Fact 4. *Negative degree assortativity: Firms with many connections, on average, trade with firms that are less well-connected.*

There is a tendency in social networks for highly connected individuals to be linked with other well-connected individuals. In contrast, production networks often exhibit significant negative degree assortativity between buyers and sellers. Bernard, Bøler, and Dhingra (2019) using Colombian importers in 2014 find a negative and significant relationship in buyer-seller assortativity. Colombian firms that have large number of suppliers are

Figure 9: Sellers' out-degree (number of buyers per seller) by firm size percentile



Notes: The figure plots the out-degree (buyers per seller) for the 20th the 50th and the 90th percentiles of sellers' sales.

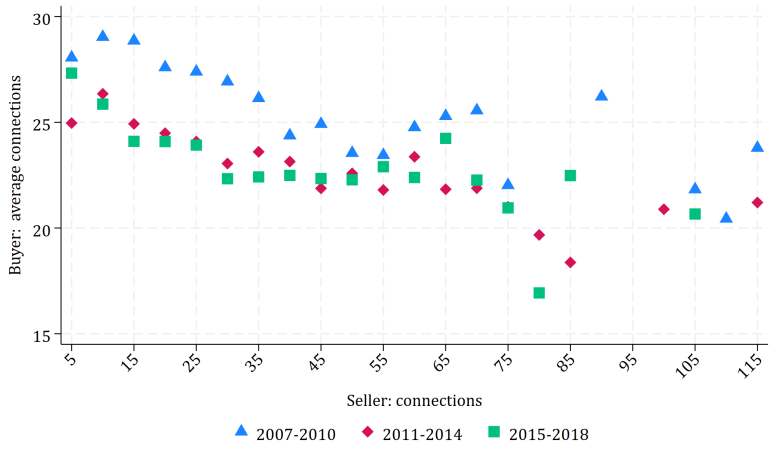
on average connected with suppliers that have fewer Colombian partners.[12]

Figure 10 illustrates exporters' degree assortativity by plotting sellers' connections across the horizontal axis and average buyers' connections across the vertical axis. To examine changes across all cross-sections, I divide the sample into three equal-sized periods, 2007 to 2010, 2011 to 2014, and 2015 to 2018, and plot the average. It is clear that there is a negative relationship for the three periods. Similarly, Figure 11 shows the same figure but illustrates importers' degree assortativity. For sellers, it is possible to see a negative relationship as well, i.e., buyers with more connections match with sellers with fewer connections, and this negative relationship does not differ greatly between the three periods.

Bernard et al. (2022) conduct a log-log regression analysis of sellers' connections and buyers' average connections using Belgium data from 2014. Their findings indicate that doubling the number of customers is associated with a 5% decline in the average number of suppliers per customer. I replicate this analysis for both buyers and sellers for each year within the network. Figure 12 presents the coefficients and the 95% confidence intervals for these log-log regressions, with exporters on the left and importers on the right. The results reveal that the fitted regression is not consistently significant at the 95% confidence level. For both exporters and importers, the negative relationship is not statistically significant during the periods 2009 to 2011 and in 2014. The coefficients exhibit a qualitatively similar pattern across both groups, increasing and decreasing in the same years, with exporters displaying larger absolute values. Moreover, the magnitudes of the coefficients are comparable to those found by Bernard et al. (2022). Given that importers are more connected and exporters less connected after correcting for reported names, it is probable that the average connection between sellers will decrease, resulting in flatter relationships.

[12]Other production networks have also shown evidence of negative degree assortativity, including studies by Bernard et al. (2018), Lim (2018), and Bernard, Moxnes, and Saito (2019).

Figure 10: Exporters' degree assortativity



Notes: The figure plots the exporters' degree assortativity. In the horizontal axis, I plot exporters' connections sorted from low to high (following left to right), using identical bins of 5 connections each. In the vertical axis, I plot the respective importer's average connections. The interval periods 2007-2010, 2011-2014 and 2015-2018 refer to the average of each bin across these years.

Fact 5. *Production networks follow hierarchy: Well-connected firms trade with a range of partners from the best connected to the least. Firms with few connections match with well-connected partners.*

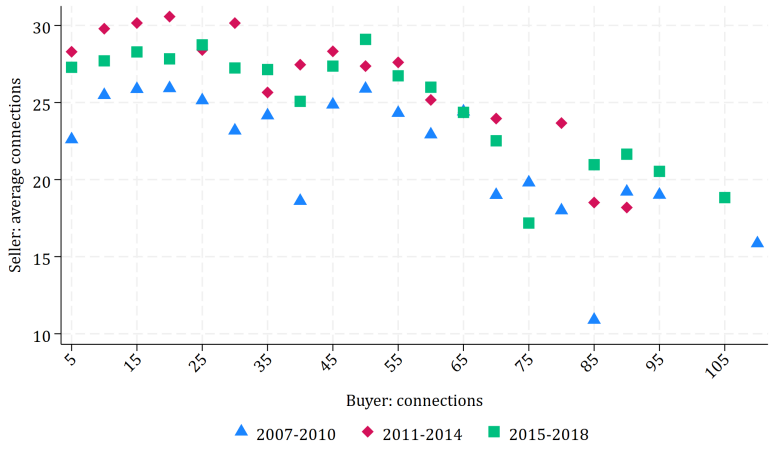
This fact underscores the hierarchical nature of production networks. Using domestic data for Japan, Bernard, Moxnes, and Saito (2019) document the hierarchical patterns where access to distant locations is more common among larger firms compared to smaller ones. Bernard et al., 2018, and Eaton et al. (2011), document the pervasiveness of buyer hierarchies, in which firms follow a hierarchical pecking order in their choice of connections.

I follow a similar procedure akin to that of Bernard et al., 2018. First, I rank buyers (sellers) from most to least connected, where $r = 1, ..., R$, the top to bottom rankings. The probability of linking to a buyer (seller) ρ_r is the number of matches to buyers (sellers) in r relative to the number of sellers (buyers). Under independence the probability to connect to well-connected buyers (sellers) is $p_1 = \rho_1 \prod_{i=2}^{R} (1 - \rho_i)$. The probability of matching to the second ranked buyer (seller) is $p_1 = \rho_1 \rho_2 \prod_{i=3}^{R} (1 - \rho_i)$ and so on. The likelihood of hierarchy under independence is $\sum_{i=1}^{R} p_i$.

I compare this likelihood relative to what is found in the data for each year in the panel, for both buyers and sellers. To compute the observed shares in the data, I use the same rankings used under the independence exercise and count matches to the top sellers (buyers), then to the top and second, and so on. If there were no pecking order of buyers (sellers) when matching to sellers (buyers), then one would expect that the likelihood under independence is equal to the shares obtained from the data.

Figure 13 and Figure 14 show the relationship between the shares of matches obtained in the data following a hierarchy, with the likelihood of following a hierarchy under independence ($\sum_{i=1}^{R} p_i$). It is clear that for 2009 onwards most linkages follow hierarchical pattern, otherwise we would have observed a close relationship between the independent

Figure 11: Importers' degree assortativity



Notes: The figure plots the importers' degree assortativity. In the horizontal axis, I plot importers' connections sorted from low to high (following left to right), using identical bins of 5 connections each. In the vertical axis, I plot the respective exporters' average connections. The interval periods 2007-2010, 2011-2014 and 2015-2018 refer to the average of each bin across these years.

likelihood and the observed shares following the 45-degree line.

In general, the network of Colombian flower exporters and US buyers exhibits the stylised patterns commonly observed in production networks. While Facts 1 to 3 illustrate the evolving patterns of the network over time, Facts 4 and 5 present higher-order complexities that require more nuanced interpretation regarding changes in the network distribution. By analysing an extended time series, I demonstrate how networks evolve, an aspect that has not been extensively documented in other production networks. Notably, some discrepancies observed in the network may be attributed to necessary corrections in the number of buyers, which could contribute to variations from what has been documented elsewhere.

4 Predicting network patterns

In this section, I simulate a random matching model of buyer-seller link formation using a balls-and-bins model (from 2014), and the implementation in firm-to-firm data from 2018. Random assignment models are often used as statistical benchmarks that are "equally capable of generating sparse production networks and the additional empirical facts" (Bernard et al., 2022, p.1). Based on their predictions, I document how well they match the Colombian-US network for all the years of the network data. Toward the end of the section, I use the estimated balls-and-bins model to showcase how quantitatively and qualitatively the predictions replicate specific distribution statistics.

Figure 12: Degree assortativity slope

Notes: The figure displays the estimated slopes for each cross-section. In the left panel, representing exporters, the slope reflects the relationship between the log of the exporters' connections and the log of their importers' average connections. In the right panel, representing importers, the slope reflects the relationship between the log of the importers' connections and the log of their exporters' average connections. The vertical lines represent the 95% confidence intervals.

4.1 Simulating a balls-and-bins model

In the balls-and-bins model modified by Bernard et al. (2018) for firm-to-firm data, J are the buyers, I the sellers, and n the balls. The number of bins is IJ—the total number of possible buyer-seller combinations. The probability that a ball lands in a bin of size ij is s_k, where $s_k = s_j s_i$, and s_j is the probability buyer j matches with a seller, and s_i is the probability that a seller matches with a buyer; they are assumed to be independent. The outcome of the balls-and-bins model is a random variable itself, which is the expected number of non-empty bins:

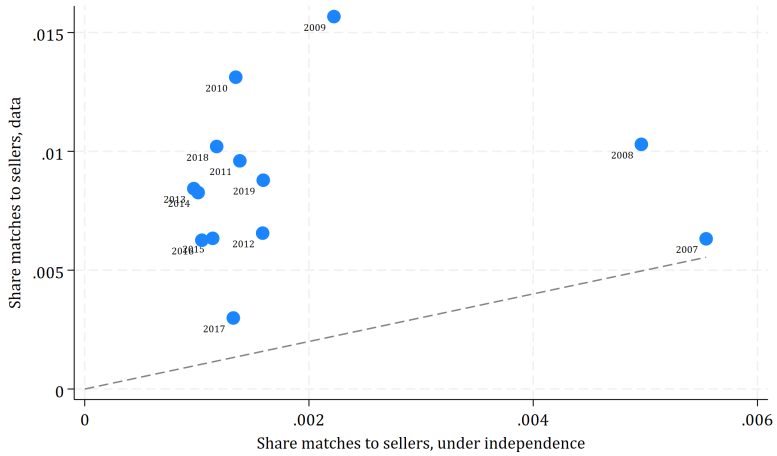
$$E(\kappa|n) = \sum_{k=1}^{K} [1 - (1 - s_k)^n],$$

where $\kappa \in \{1, 2, ..., K\}$ are the non-empty bins. This simple version of balls-and-bins model has some major drawbacks. First, the model has strong assumptions on the number of balls which are given to the model, i.e., n, is assumed to be given, and second, the model is not able to reproduce all five stylised facts (Bernard and Zi, 2022).[13]

A more general balls-and-bins model is proposed by Bernard and Zi (2022). Their "elementary model" of production networks does not have as strong assumptions as the benchmark model, and is able to reproduce Facts 1 to 5, at least qualitatively. The "elementary model" incorporates: (1) variation across the buyers' level of purchases and the sellers' ability to attract buyers; and (2) buyer-seller matches that are non-deterministic. In this setup, the probability that a buyer j matches to a seller i equals the probability that i

[13] In the simulation using Norwegian data Bernard et al. (2018) assume the number of balls, n, is the total number of transactions (5,000 on average across the panel).

Figure 13: Pecking order hierarchy across buyers



Notes: The figure plots the actual shares of firms following the hierarchy on the vertical axis and the simulated shares under the assumption of independence on the horizontal axis. The shares on the vertical axis represent the number of buyers matching to the top seller, the top and second top seller, and so on. The horizontal axis represents the simulated shares under the assumption that connection probabilities are independent ($\sum_{i=1}^{R} p_i$).

receives a purchase from j:

$$p_{ij} = 1 - (1 - s_i)^{b_j},$$

where s_i varies for every buyer and is the probability that any "purchase ball" lands in a seller i's bin, with $s_i \in [0, 1)$ and $\sum_i s_i = 1$. A main caveat of this specification is that all buyers "throw balls" of equal sizes l, so the total number of balls $b_j \geq 0$ thrown by buyer j become b_j/l, where l is not estimated but assumed.

Based on the "elementary model" from Bernard and Zi (2022), I consider a generalised balls-and-bins model where buyers j throw a number of balls b_j to sellers i. Sellers receive balls based on the size of their bins, s_i, relative to the size of all other sellers, where $\sum_i s_i = 1$. The probability that a buyer j matches to a seller i equals the probability that a seller receives one purchase from j:

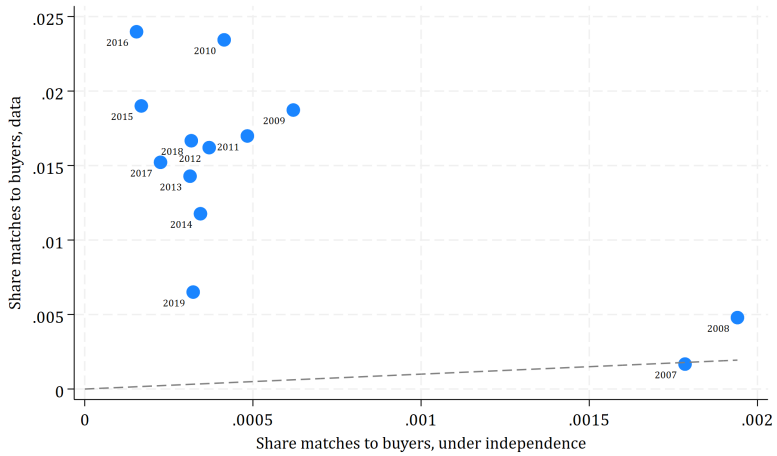
$$p_{ij} = 1 - (1 - s_i)^{\gamma b_j / d_{ij}^{\eta}},$$

where link formation attempts by j—the number of balls thrown b_j—increases linearly with j's size, x_j, and decays with i's distance from j, d_{ij}, at constant elasticity $\eta > 0$. Following the benchmark balls-and-bins models, it is possible to interpret γ as a parameter where $\gamma b_j = x_j/l$, and γ is the inverse of ball size that is not assumed.

4.2 Estimation and predictive power

I predict $\hat{p}_{ij}$ independently for every cross-section of the data. Note that all variables have time subscripts, but to ease the reading, I omit the subscript. For each cross-section, I construct the bin sizes s_i as the total sales of seller i to the US, where $x_i = \sum^{J} x_{ij}$ is the total

Figure 14: Pecking order hierarchy across sellers



Notes: The figure plots the actual shares of firms following the hierarchy on the vertical axis and the simulated shares under the assumption of independence on the horizontal axis. The shares on the vertical axis represent the number of sellers matching to the top buyer, the top and second top buyer, and so on. The horizontal axis represents the simulated shares under the assumption that connection probabilities are independent ($\sum_{i=1}^{R} p_i$).

sales from seller i across all buyers. The balls, b_j, are equivalent to the total purchases from buyer j across all sellers, where $x_j = \sum^I x_{ij}$. For simplicity, I assume a distance elasticity $\eta = 0$. I estimate γ by non-linear least squares estimations, and use the estimated parameter $\hat{\gamma}$ to predict $\hat{p}_{ij}$.

I compare the generalised balls-and-bins model to other generalisations of trade models to see how well each model predicts the observed data. In addition to the generalised balls-and-bins model, I use a non-parametric approximation that obeys the general structure $\Pr\left(y_{ij} = 1 \mid x_i, x_j, d_{ij}\right)$. Since any structure of this type will involve estimating non-linear models with high-dimensional fixed effects that can induce a bias on the coefficients due to incidental parameters, I estimate a set of non-parametric models based on linear approximations.[14]

For comparing the fitness of the generalised balls-and-bins model, I estimate four other specifications. First, a non-parametric linear model with only gravity, i.e., including buyers' and sellers' fixed effects. Second, a non-parametric balls-and-bins model with $\mathcal{S}(\cdot)$ partitioning the relevant range of (x_i, x_j) into $B^{\mathcal{S}}$ bins. Third, a model that includes both gravity and the non-parametric balls-and-bins. A final model involves a more flexible non-parametric specification that allows for interaction of observables with origin and destination effects.[15]

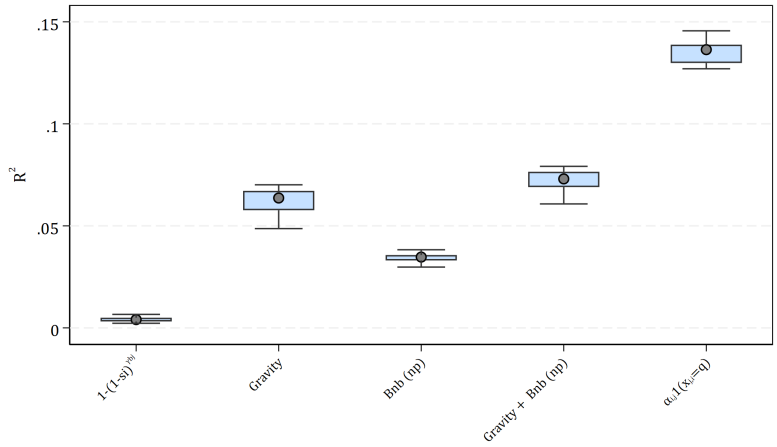
Figure 15 shows the distribution of the R^2 statistic for all five models. The predic-

[14] For estimation of non-linear models, jackknife bias-corrected estimators (see Hughes, 2022) are necessary.

[15] For the non-parametric balls-and-bins, I split sellers and buyers into different quantiles q from their observed sizes, and construct 10 bins for sellers and 10 for buyers. I construct the 100 bins, $B^{\mathcal{S}}$, from all the combinations of the buyer-seller bins. The flexible specification uses the interaction of buyers (sellers) fixed effects and the corresponding seller (buyer) bin. For q quantiles this can be represented as the following bin and fixed effects interactions: $\sum_q \left(\alpha_i^O \mathbb{1}\{\mathcal{Z}(x_j) = q\} + \alpha_j^D \mathbb{1}\{\mathcal{Z}(x_i) = q\}\right)$.

tion of the generalised balls-and-bins have the lowest predictive power. It seems that gravity on its own does a better job than the generalised balls-and-bins model and the non-parametric balls-and-bins model. Combining gravity with the non-parametric balls-and-bins is marginally better than just using gravity. Despite its lack of predictive power, the balls-and-bins specification can explain some of the variation in the data. Finally, the more flexible approach of using the combination of firm bins with firm fixed effects yields the highest predicting power.

Figure 15: Distribution of R^2 statistics for estimated models



Notes: The figure shows the distribution of the R^2 statistic for all cross-section estimations in each model. The circle displays the median, while the lines from top to bottom represent the maximum value, the 75$^{\text{th}}$ percentile, the 25$^{\text{th}}$ percentile, and the minimum value, respectively. The left-hand dependent variable use the observed connections in the data. The first model is the estimated generalised balls-and-bins model; the second model is gravity using exporter and importer fixed effects; the third model is a non-parametric balls-and-bins model using 100 bin combinations; the fourth model combines gravity with the non-parametric balls-and-bins model; and the fifth model uses interactions from firm size quantiles (with $q = 10$) combined with fixed effects.

4.3 Simulations and network statistics

The generalised balls-and-bins model can qualitatively reproduce stylised Facts 2, 4, and 5. However, the predictions cannot reproduce Fact 1 or Fact 3. This is not surprising given the model's low predictive power. Regarding Fact 3, the patterns observed in the network differ from those documented in other contexts. In this section, I focus on Fact 2, the in-degree and out-degree distributions, and analyse how well the generalised balls-and-bins model can predict different statistics from these distributions both qualitatively and quantitatively.

To compare the data with the estimated networks, the random draws must replicate the number of connections observed in the data. To ensure this, I scale the estimated model using a scaling factor $\delta \in (0, 1]$. This scaling affects buyers, as their ball sizes are adjusted to $l = 1/(\hat{\gamma}\delta)$. A higher value of δ results in smaller ball sizes. With reduced ball sizes, buyers have more balls to throw, which increases the probability that a ball will land in

any given bin.[16] Figure A2.4 displays the predicted density before and after re-scaling, while Figure A2.5 illustrates the ball sizes before and after re-scaling. The densities from the estimated model are 20 times smaller than those obtained after re-scaling by δ, and the ball sizes without re-scaling are 44 times larger.

To obtain results independent of specific draws, I present averages across 50 simulations for each cross-section. Each draw is derived from a binomial distribution with the re-scaled prediction $\hat{p}_{ij}(\delta)$. Figure 16 compares the distribution of connections by plotting the average simulated distribution alongside the data from 2007, with both distributions binned on the horizontal axis. The left panel displays the in-degree distribution, while the right panel shows the out-degree distribution. I include only one year for illustration, as other cross-sections exhibit similar patterns (Figures A2.6 and A2.7 present results for 2012 and 2018 respectively). The distribution for both in-degree and out-degree is qualitatively close to the observed data. However, quantitatively, the predicted distribution overestimates the number of sellers with few connections and is right-skewed, suggesting that it predicts some firms will have more connections than they actually do. Similar results are reported by Bernard and Zi (2022), who also found their balls-and-bins simulations for Norwegian data overpredicted matches for well-connected buyers and sellers but performed reasonably well in predicting matches for the middle of the distribution.

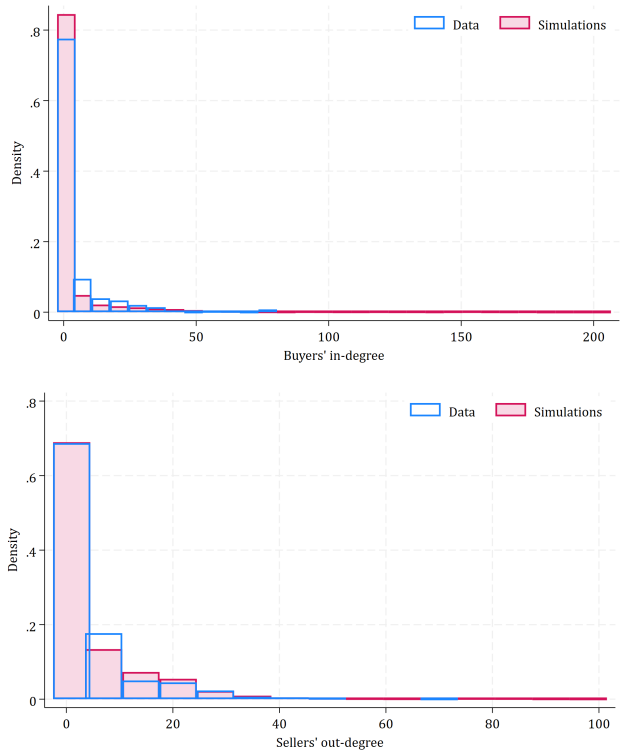
Figure A2.8 shows the ratio between simulation averages and the data for the 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles. This ratio can be used to assess how accurate predictions are at a given point in the distribution, meaning that the closer the ratio is to one, the more accurate the predictions. In-degree ratios between simulations and data are relatively low, with the 25^{th} percentile having the worst ratio and the 90^{th} percentile having the best ratio (0.8). Out-degree ratios are much higher at all percentiles and are closer to one in the 75^{th} and 90^{th} percentiles, but are about 0.5 for the bottom of the distribution. This means the model predicts connections for well-connected sellers more accurately but only predicts about half of the connections for the less well-connected sellers.

The balls-and-bins model performs qualitatively well but tends to overestimate the average number of connections in the tails of both the sellers' and buyers' distributions. While it may not accurately predict the precise number of connections, the model could still provide a reliable estimate of the relative distribution of connections within the network. To assess this, I estimate inter-quantile ratios for the in-degree and out-degree distributions, which measure the ratio between two percentiles of the degree distributions. Figure 17 displays the inter-quantile ratios from the simulation alongside those from the data for buyers, while Figure 18 shows the same for sellers. For buyers, the model accurately predicts the relative distribution between the 50^{th} and 75^{th} percentiles (p^{75}/p^{50}), but performs less well for the 90^{th} to 50^{th} percentile ratio (p^{90}/p^{50}). This result is expected, given that the model overestimates the number of connections at the 90^{th} percentile by nearly a factor of two.

The inter-quantile ratios for sellers are notably more accurate, with most years showing

[16] In other balls-and-bins models, the value of l is typically set rather than estimated. For instance, Atalay et al. (2011) and Bernard et al. (2018) use an average shipment value of \$36,000 US dollars per ball, while Bernard and Zi (2022) use various sizes ranging from \$100 to \$2,000 and \$36,000 US dollars.

Figure 16: In-degree and out-degree distributions for data and simulations

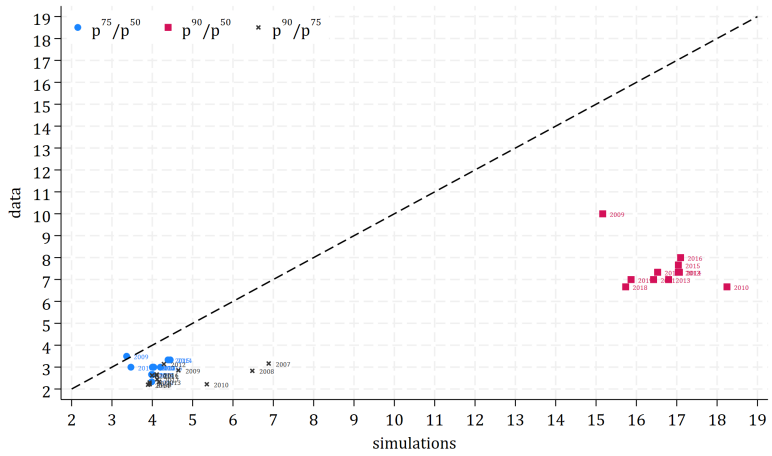


Notes: The figure shows the distribution of the balls-and-bins simulations and the observed data for the in-degree (top panel) and the out-degree (bottom panel). I split the bins in the horizontal axis using the maximum connections in the simulations divided by a factor of 10.

results that are relatively close to the 45-degree line. However, while the model performs better with respect to the inter-quantile ratios for sellers, it does not always accurately predict the direction of changes in these ratios. For example, between 2007 and 2011, the observed inter-quantile ratios p^{90}/p^{50} and p^{75}/p^{50} increase, indicating that well-connected sellers become more connected relative to the median. In contrast, the simulated ratios predict a decrease. In other periods, the model succeeds in capturing the trend of the inter-quantile ratios. For instance, between 2009 and 2015, both the observed and simulated inter-quantile ratios decrease, suggesting that well-connected sellers become less connected relative to median-connected sellers.

In summary, the balls-and-bins model is effective for capturing broad qualitative patterns in networks but falls short in accurately fitting more detailed features. Alternative matching models, which offer higher predictive power than the balls-and-bins approach, provide a better fit for nuanced aspects of network data. However, these models often involve high-dimensional fixed effects, making non-linear estimations computationally demanding. Despite its limitations in predicting extreme connections and the underlying assumption of independence, the balls-and-bins model remains valuable for understanding connections within the central ranges of the distributions.

Figure 17: In-degree inter-quantile ratio for data and simulations



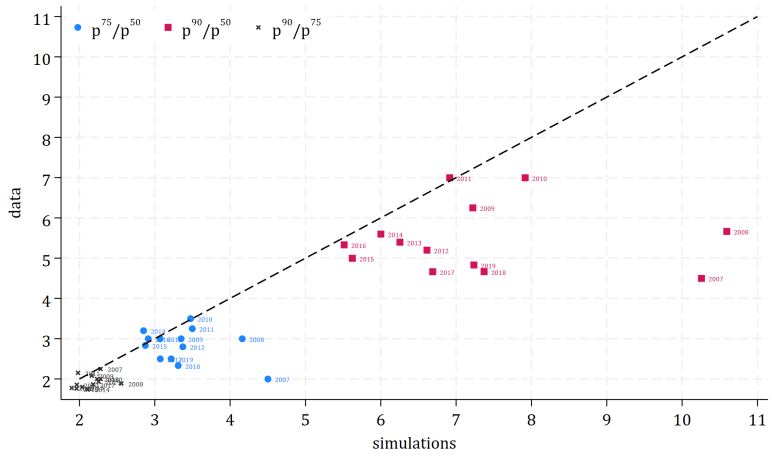
Notes: The figure shows the correlation between the inter-quantile ratios obtained from the simulated balls-and-bins model and the data for buyers' in-degree.

5 Conclusion

As access to administrative firm-to-firm data has expanded, empirical research has made significant strides in understanding buyer-seller relationships. With an increasing number of countries collecting detailed granular data, there is a growing need for comprehensive documentation of the limitations and discrepancies in data collection practices across different jurisdictions. While current documentation predominantly addresses discrepancies in trade values, often attributed to agent behaviour, there is limited analysis of discrepancies related to the number of firms. Using Colombian customs data, I document issues related to firm counting in datasets lacking official identifiers, which can result in an overestimation of firm numbers. Specifically, inconsistencies in name revisions disproportionately affect larger firms, which represent a substantial share of connections and trade, leading to a misrepresentation of key network statistics used in theoretical models of firm-to-firm allocations.

Colombian customs data provides a valuable opportunity to document systematic patterns due to its extensive use in firm-to-firm empirical research and its relative accessibility. By constructing a buyer-seller network and adjusting for corrections to foreign names, I replicate stylised facts of production networks over an extended period. Although these patterns are qualitatively similar to those observed in other networks, this analysis underscores the importance of documenting statistics that capture changes in the distribution of connections within the network over time. Finally, by evaluating the fit of a generalised balls-and-bins model to the network data, the results reveal that while random matching models can produce distributions qualitatively similar to the observed network, they often fail to accurately represent the tails of the in-degree and out-degree distributions. Moreover, these random matching models may inadequately capture other network features due to their reliance on assumptions such as independent link formation. Future research should consider reassessing some of these assumptions to better understand and model

Figure 18: Out-degree inter-quantile ratio for data and simulations



Notes: The figure shows the correlation between the inter-quantile ratios obtained from the simulated balls-and-bins model and the data for sellers' out-degree.

the redistribution of connections in production networks.

References

Allingham, M., & Sandmo, A. (1972). Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics*, *1*(3-4), 323–338.

Arkolakis, C. (2010). Market Penetration Costs and the New Consumers Margin in International Trade. *Journal of Political Economy*, *118*(6), 1151–1199.

Armenter, R., & Koren, M. (2014). A Balls-and-Bins Model of Trade. *American Economic Review*, *104*(7), 2127–2151.

Atalay, E., Hortaçsu, A., Roberts, J., & Syverson, C. (2011). Network Structure of Production. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 5199–202.

Benguria, F. (2022). Do US Exporters Take Advantage of Free Trade Agreements? Evidence from the USColombia Free Trade Agreement. *Review of International Economics*, *30*(4), 1148–1179.

Benita, F., & Urzúa, C. M. (2016). Mirror Trade Statistics Between China and Latin America. *Journal of Chinese Economic and Foreign Trade Studies*, *9*(3), 177–189.

Bernard, A., Bøler, E., & Dhingra, S. (2019). *Firm-to-Firm Connections in Colombian Imports* (Routledge-ERIA Series in Development Economics). World Trade Evolution – Growth, Productivity and Employment.

Bernard, A., & Dhingra, S. (2019). *Importers, Exporters and the Division of the Gains from Trade* (Working paper).

Bernard, A., Dhyne, E., Manova, K., & Moxnes, A. (2022). The Origins of Firm Heterogeneity: A Production Network Approach. *Journal of Political Economy*, *130*(7), 1765–1804.

Bernard, A., & Moxnes, A. (2018). Networks and Trade. *Annual Review of Economics*, *10*, 65–85.

Bernard, A., Moxnes, A., & Saito, Y. (2019). Production Networks, Geography, and Firm Performance. *Journal of Political Economy*, *127*(2), 639–388.

Bernard, A., Moxnes, A., & Ulltveit-Moe, K. (2018). Two-sided Heterogeneity and Trade. *Review of Economics and Statistics*, *100*(3), 424–439.

Bernard, A., & Zi, Y. (2022). *Sparse Production Networks* (Working paper).

Blum, B., Claro, S., & Horstmann, I. (2012). *Import Intermediaries and Trade Costs: Theory and Evidence* (Working paper).

Carballo, J., Ottaviano, G., & Volpe Martincus, C. (2018). The Buyer Margins of Firms' Exports. *Journal of International Economics*, *112*, 33–49.

Chaney, T. (2014). The Network Structure of International Trade. *American Economic Review*, *104*(11), 3600–3634.

Eaton, J., Eslava, M., Jinkins, D., Krizan, C., & Tybout, J. (2021). *A Search and Learning Model of Export Dynamics* (Working paper No. 29100). National Bureau of Economic Research.

Eaton, J., Kortum, S., & Kramarz, F. (2011). An Anatomy of International Trade: Evidence From French Firms. *Econometrica*, *79*(5), 1453–1498.

Eaton, J., Kramarz, F., & Kortum, S. (2019). *Firm-to-Firm Trade: Exports, Imports, and the Labor Market* (Working paper No. 9557). CESifo.

Fajgelbaum, P., & Khandelwal, A. (2021). The Economic Impacts of the US-China Trade War. *Annual Economic Reviews*, *14*, 205–228.

Feenstra, R. C., Hai, W., Woo, W. T., & Yao, S. (1999). Discrepancies in International Data: An Application to China-Hong Kong Entrepot Trade. *American Economic Review*, *89*(2), 338–343.

Ferrantino, M., Liu, X., & Wang, Z. (2012). Evasion Behaviors of Exporters and Importers: Evidence from the USChina Trade Data Discrepancy. *Journal of International Economics*, *86*(1), 141–157.

Fisman, R., & Wei, S. (2004). Tax Rates and Tax Evasion: Evidence from 'Missing Imports' in China. *Journal of Political Economy*, *112*(2), 471–500.

Hamanaka, S. (2012). Whose Trade Statistics Are Correct? Multiple Mirror Comparison Techniques: A Test Case of Cambodia. *Journal of Economic Policy Reform*, *15*(1), 33–56.

Helpman, E., Melitz, M., & Rubinstein, Y. (2007). Estimating Trade Flows: Trading Partners and Trading Volumes. *IO: Productivity*.

Herkenhoff, P., Krautheim, S., & Sauré, P. (2021). *A Simple Model of Buyer-Seller Networks in International Trade* (Working paper No. 9124). CESifo.

Hughes, D. (2022). *Estimating Nonlinear Network Data Models with Fixed Effects* (Working paper No. 2203.15603).

Javorcik, B., & Narciso, G. (2017). WTO Accession and Tariff Evasion. *Journal of Development Economics*, *125*(100), 59–71.

Kellenberg, D., & Levinson, A. (2019). Misreporting Trade: Tariff Evasion, Corruption, and Auditing Standards. *Review of International Economics*, *27*(1), 106–129.

Kramarz, F., Martin, J., & Mejean, I. (2020). Volatility in the Small and in the Large: The Lack of Diversification in International Trade. *Journal of International Economics*, *122*(103276).

Krizan, C., Tybout, J., Wang, Z., & Zhao, Y. (2020). *Are Customs Records Consistent Across Countries? Evidence from the US and Colombia* (Working paper No. 20-11). Center for Economic Studies, US Census Bureau.

Lim, K. (2018). *Endogenous Production Networks and the Business Cycle* (Working paper).

Liu, F., Wheiler, K., Ganguly, I., & Hu, M. (2020). Sustainable Timber Trade: A Study on Discrepancies in Chinese Logs and Lumber Trade Statistics. *Forests*.

Magerman, G., Dhyne, E., & Rubínová, S. (2015). *The Belgian Production Network 2002-2012* (Working paper No. 288). National Bank of Belgium.

Makhoul, B., & Otterstrom, S. (1998). Exploring the Accuracy of International Trade Statistics. *Applied Economics*, *30*(12), 1603–1616.

Mishra, P., Subramanian, A., & Topalova, P. (2008). Tariffs, Enforcement, and Customs Evasion: Evidence from India. *Journal of Public Economics*, *92*(10-11), 1907–1925.

Sen, R. (2000). Analysing International Trade Data in a Small Open Economy. *Journal of Southeast Asian Economies*, *17*, 23.

Sheveleva, L. (2019). Multi-product Exporters: Facts and Fiction. *SSRN Electronic Journal*.

Slemrod, J., & Yitzhaki, S. (2002). *Tax Avoidance, Evasion, and Administration* (A. J. Auerbach & M. Feldstein, Eds.; 1st ed., Vol. 3).

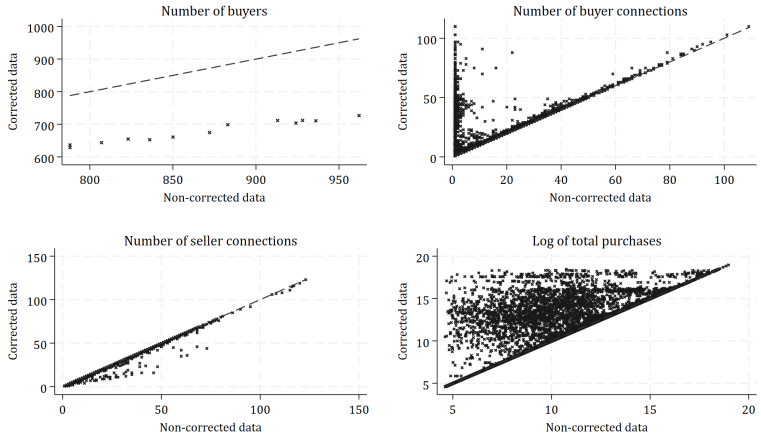
Stoyanov, A. (2012). Tariff Evasion and Rules of Origin Violations Under the Canada-US Free Trade Agreement. *Canadian Journal of Economics*, *45*(3), 879–902.

Sugita, Y., Teshima, K., & Seira, E. (2023). Assortative Matching of Exporters and Importers. *Review of Economics and Statistics*, *105*(6), 1544–1561.

Vincent, J. (2004). *Detecting Illegal Trade Practices by Analyzing Discrepancies in Forest Products Trade Statistics: An Application to Europe, with a Focus on Romania* (Working Paper No. 3261). The World Bank.

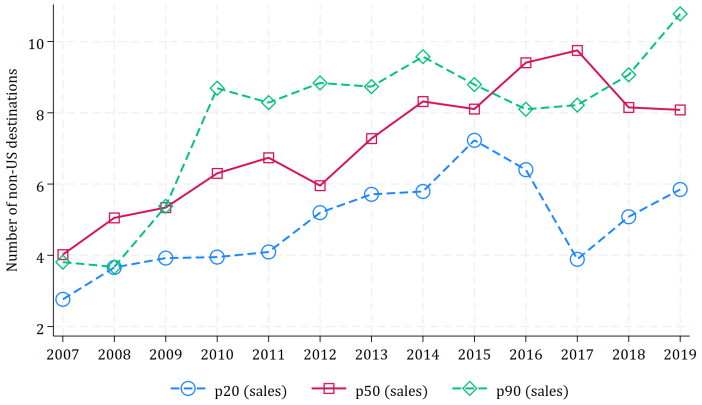
A Figure appendix

Figure A2.1: Correlation corrected and non-corrected buyers' names



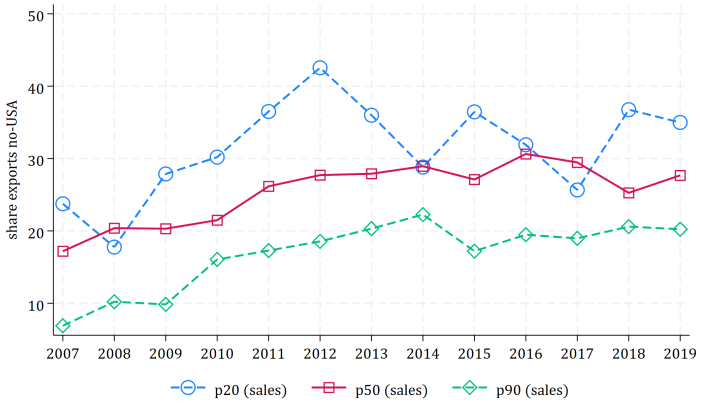
Notes: The figure displays the correlation between the values obtained from corrected and non-corrected buyers' names when accounting for duplicated buyer names in some transactions. The top-right panel illustrates the differences in the number of buyers, while the top-left panel shows the discrepancies in buyer connections, which are undercounted when using non-corrected data. The bottom-left panel highlights the number of seller connections, which are overcounted with non-corrected data, and the bottom-right panel depicts the total purchases, which are underestimated when using non-corrected data. The 'non-corrected' data uses all buyers up to filter 3 in Table 1, but counts as a unique buyer those names that account for more than one firm.

Figure A2.3: Number of non-US destinations



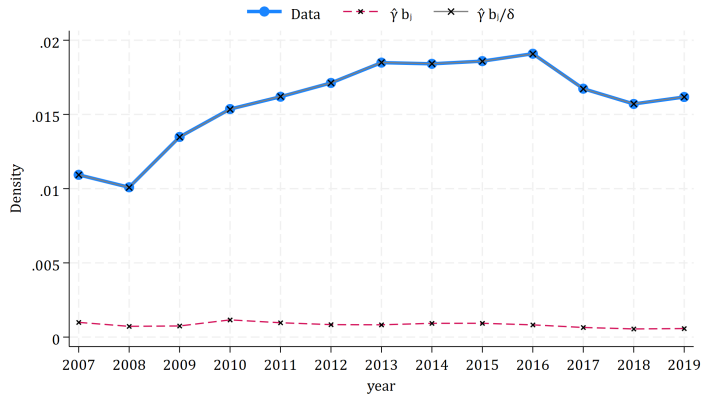
Notes: The figure plots the average number of non-US destinations for small, medium and large exporter by sales.

Figure A2.2: Share of non-US exports



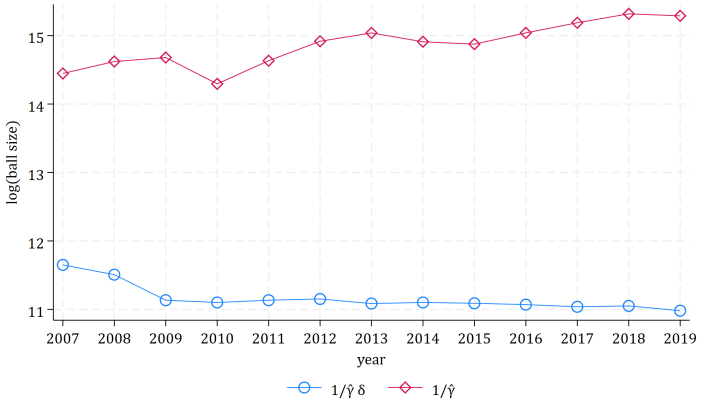
Notes: The figure plots the average share of exports to non-US destinations for small, medium and large exporter by sales.

Figure A2.4: Simulated density from generalised balls-and-bins model



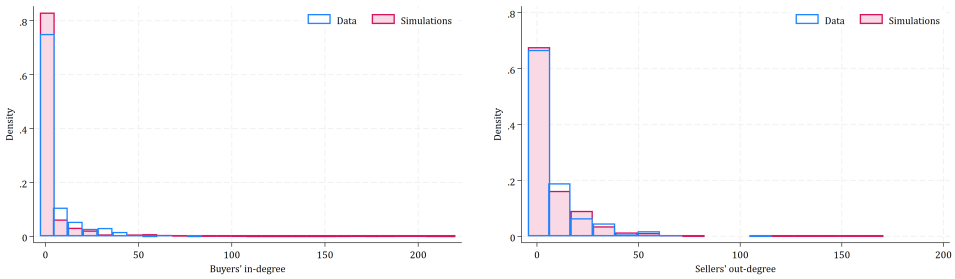
Notes: The figure plots the density of the data, the density from the estimations of the generalised bins-and-balls model with $\hat{\gamma}$, and the density after re-scaling the model by δ.

Figure A2.5: Estimated and re-scaled buyer's ball size



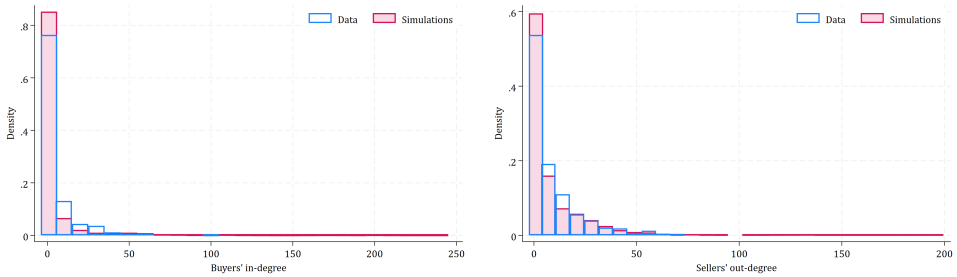
Notes: The figure plots the log (ball size) for the estimated balls-and-bins model in red and the re-scaled log(ball size) in blue.

Figure A2.6: In-degree and out-degree distributions for data and simulations (2012)



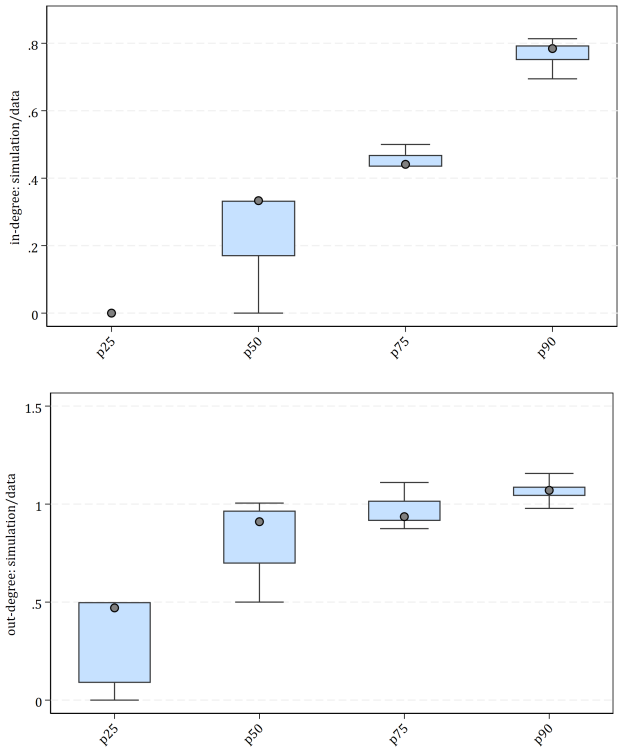
Notes: The figure shows the distribution of the balls-and-bins simulations and the observed data for the in-degree (left panel) and out-degree (right panel). I split the bins in the horizontal axis using the maximum connections in the simulations divided by a factor of 10.

Figure A2.7: In-degree and out-degree distributions for data and simulations (2018)



Notes: The figure shows the distribution of the balls-and-bins simulations and the observed data for the in-degree (left panel) and out-degree (right panel). I split the bins in the horizontal axis using the maximum connections in the simulations divided by a factor of 10.

Figure A2.8: In-degree and out-degree fit ratio by percentile



Notes: The figures plot the ratio between the in-degree (top panel) and out-degree (bottom panel) percentiles between the simulations and the observed data for the 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles. I use the average percentile from all the simulated samples.