Cutting Out the Middleman: The Structure of Chains of Intermediation*

Matthew Grant† and Meredith Startz‡

April 2022 [Please click for most recent version]

Abstract

Distribution of goods often involves multiple intermediaries sequentially buying and reselling. We show that multi-intermediary chains arise in response to internal economies of scale in trade costs, and that this can account for patterns in firm size and prices we document in original data on consumer goods in Nigeria. Interventions that shorten chains involve a fundamental welfare trade-off: they lower marginal cost but also the number of sellers, reducing competition, product availability, and access to retailers. We embed this general insight in a quantifiable model, and find that for apparel in Nigeria, cutting out middlemen often harms more remote consumers.

*We thank Treb Allen, David Atkin, Kyle Bagwell, Andy Bernard, Johannes Boehm, Lorenzo Casaburi, Arun Chandrasekhar, Shoumitro Chatterjee, Davin Chor, Swati Dhingra, Cecile Gaubert, Gene Grossman, Sam Kortum, Eduardo Morales, Ezra Oberfield, Nina Pavcnik, Steve Redding, Andres Rodriguez-Clare, Ina Simonovska, and Bob Staiger for their helpful comments. This work would not be possible without the excellent research assistance of Hakeem Bishi, Yemi Ogundare, Samuel Adesanya, Nicole Wilson, Sheah Deilami-Nugent, and the Lagos Trader Project enumerator team. Special thanks are due to Shelby Grossman for her collaboration on the Lagos Trader Survey, and to David Atkin and Dave Donaldson for sharing apparel price data. We thank the Tokyo Foundation for Policy Research and the Kauffman Foundation for financial support.

†Dartmouth College. Email: matthew.w.grant@dartmouth.edu

‡Dartmouth College and NBER. Email: meredith.l.startz@dartmouth.edu

1 Introduction

How do goods made in one place reach consumers in another? Models of trade typically abstract from the details, assuming that a producer sells directly to consumers. The reality is that goods may pass through the hands of multiple intermediaries. A shopkeeper selling mobile phones in a small town in Nigeria, for instance, is unlikely to source them directly from the manufacturer in southern China. Instead, she might buy from a wholesaler in a regional market town, who relies on an importer, and so on. At each step toward consumers, costs are incurred and markups may be charged. Understanding intermediation chains is therefore key to understanding the prices and product availability faced by consumers in different locations, and their potential to gain from globalization and trade.

Wholesale and retail firms account for a large share of economic activity all over the world, and there is a great deal of interest in the extent to which they hold market power and mediate price passthrough in both rich and poor countries. Political discourse has often taken a dim view of the trading sector, seeing it as merely driving a price wedge between producers and consumers. This is reflected in policies that restrict the role of middlemen, from limits on where agents can buy and sell (such as India's regulated agricultural marketplaces) to outright prohibition (such as Bangladesh's 2011 ban on delivery order traders). Recently, the idea that changes in trade costs might connect producers and consumers more directly even without heavy-handed policy intervention has gained traction – effectively, that middlemen will be cut out naturally as a result of falling transport costs, improving communication, and new matching technologies such as e-commerce platforms or programs that link farmers with buyers.

Are policies and technologies that cut out middlemen likely to help or harm consumers? Answering this question requires an understanding of why multiple layers of intermediation might arise in the first place. We show that internal economies of scale in the trade costs incurred by an intermediary who buys and resells a good — whether these costs arise from transport, search, or other functions that intermediaries serve — offer a general microfoundation for the existence of multi-intermediary chains. This insight allows us to consider how chain structure will respond endogenously to changes in policies and costs, and how this in turn will influence consumer welfare. Shorter chains *can* increase welfare but will not necessarily do so, regardless of whether they are induced by policy mandates or technology improvements. Instead, shorter chains imply both lower marginal costs and lower entry of intermediaries in downstream markets, and there is a fundamental welfare tradeoff between these two forces.

This paper has three main contributions. First, we introduce data that spans several links in a multi-intermediary distribution chain and document new stylized facts. Second, we provide conceptual framework for why such chains arise, and show that this also offers insight into the welfare implications of policy or technology changes that alter chain structure. Finally, we build these ideas into a tractable, quantifiable framework that can be taken to data, and show that it can account for the stylized facts we have laid out.

We begin by introducing data from an original survey of wholesale and retail traders in Lagos, Nigeria that highlights the ubiquity of multi-intermediary chains and their relevance for consumer prices. These traders source manufactured consumer goods like apparel and electronics from all over the world, and sell to customers throughout Nigeria. A novel feature of the data is that it allows us to observe multiple links along the distribution chain. Two-thirds of respondents' international suppliers are upstream wholesalers rather than manufacturers, and the majority of their sales are to downstream traders rather than final consumers. It is known that agricultural collection chains in developing countries often include multiple levels of intermediation (Barrett et al. (2020)). We document that this is also the case for the distribution of manufactured goods, and that firm size, sourcing costs, and markups co-vary systematically with chain position.

We next show that a simple mechanism gives rise to such multi-intermediary chains: internal economies of scale in sourcing. If individual buyers face fixed costs that vary across sellers or markets, it may be profit-maximizing to source a good from a reseller rather than directly from the producer, depending on the trade-off between fixed and variable costs. Intermediaries serve a range of functions, and fixed costs (or internal economies of scale more generally) are a feature many of them. Our framework describes how any task or trade cost that has this feature – regardless of the underlying source of the costs – leads to multi-intermediary chains. Examples are everywhere in wholesale and retail trade, ranging from time (e.g. spent searching for products), to transport (e.g. operating a truck or filling a shipping container), to financing (e.g. fees for wire transfers and letters of credit), to regulatory costs (e.g. licenses and port inspection fees).

Intuitively, this is why a rural shopkeeper in Nigeria may prefer to source goods from a local market town rather than directly from China. For a small firm, saving on port fees and the cost of a trip to China may be worth paying a higher unit price to a local wholesaler. Iterating this logic can lead to a whole chain of intermediaries who source from other intermediaries. A distribution chain is, essentially, an endogenous set of points at which downstream demand is aggregated up in response to economies of scale in sourcing. The structure of the chain, including the number of links and the degree of competition at each link, will depend on the specifics of demand and trade costs. We show that places with features characteristic of developing country markets, such as small equilibrium firm size and high barriers to accessing production locations, are more likely to be served via long chains. Our framework may therefore be particularly relevant to policy in developing countries, and helpful in understanding cross-country differences in distribution and collection structures.

This framework is also useful for understanding the welfare implications of intermediation chains. The availability of resale markets where the same good can be sourced at different costs can be understood as analogous to a choice over production technologies. This constitutes a new dimension of a classic question about optimal levels of entry when firms have market power. We consider the problem of a planner who cares about welfare in a destination market, where there are consumers who demand a good produced elsewhere. The planner takes entry decisions and imperfectly competitive behavior of sellers in wholesale and retail markets as given, but manipulates the menu of sourcing options to achieve second-best outcomes. The planner can, for instance, prohibit sourcing from particular markets or from resellers, or set policy that affects trade costs. Policies that shorten chains implicitly move the destination market toward sourcing strategies with lower variable and higher fixed costs. Shorter chains benefit consumers by reducing marginal costs, but harm them by reducing entry. The distribution structure selected by the market equilibrium will not generally be efficient. Because of these second best considerations, even pure cost reductions – for instance, the removal of bureaucratic barriers to accessing certain markets, or introduction of platforms that reduce communication costs – can shift the market toward sourcing strategies that lower welfare. Broadly, cutting out middlemen can either help or harm consumers.[1]

To pin down these theoretically ambiguous welfare effects, we build a quantifiable model that relates the equilibrium chain structure serving consumers in many locations to fundamentals of geography and demand. The model connects to a common quantitative trade framework with monopolistically competitive firms facing CES demand, but innovates on a number of dimensions in order to capture resale and source-specific internal scale economies – and the resulting endogenous distribution chains – in a tractable way.[2] We consider distribution of a single good produced in a single origin market. Heteroge-

[1] Manipulation of resale sourcing technologies is never first-best. Our conclusion is not that policymakers necessarily *ought* to regulate intermediation in this way – that depends on the set of policy instruments available and the underlying sources of trade costs in a specific setting. Rather, we offer a way of thinking about how distribution structures will change in response to policies that are undertaken in the real world, as well as actual changes in trade costs and technologies.

[2] In modeling chains, our work is complementary to emerging literatures on endogenous production networks, global value chains, and transportation routing. We share some technical considerations with these literatures, but examine a distinct empirical phenomenon driven by a different underlying mechanism. Intermediation involves buying and reselling without transformation – there is no set number of tasks that need to be completed, and links describe transaction points, rather than routing through physical locations. We attribute the existence of chains with multiple links to internal economies of scale, rather than task- or match-specific productivities that vary across firms or locations (as with production), or to external economies of scale in trade costs (as with transport).

neous wholesale and retail traders can enter in each location to serve local consumers, and also to resell onward to downstream traders. Traders at each link charge markups, and so multi-intermediary chains feature double marginalization. We provide a tractable approach to modeling wholesale pricing: elasticities of demand pass back along the chain in a way that fully incorporates the endogenous decisions of downstream sellers but aggregates to a closed-form solution for the elasticity at each step.

We quantify the model for distribution of Chinese-made apparel throughout Nigeria, combining our original survey data with secondary data on prices, consumer spending, and road travel times. Our estimates are able to match a nuanced set of observed empirical patterns. First, Nigerian consumers are served by long chains. Second, traders that are further downstream from producers are smaller, and pay higher prices and lower fixed costs to source apparel. Traders that are more upstream from final consumers are larger, and source more directly from producers. Finally, elasticities of demand are passed back along the chain, and traders that are more upstream face more elastic demand and charge lower markups. Consumers in smaller and more remote markets are served by longer chains.

Using these estimates, we consider several counterfactuals related to ongoing policy discussions in developing countries. We first consider what would happen if Nigerian retailers were only able to source directly from China. This exercise captures the spirit of less extreme mandatory disintermediation policies that are used in practice, such as India's mandis.[3] At baseline, most retailers source from a local wholesaler, and so this implies large increases in their equilibrium fixed cost of sourcing. As a consequence, the number of retail outlets carrying Chinese apparel plummets, consumer spending shifts to other goods, and welfare throughout Nigeria falls. Second, we consider the role of investments or technologies that target fixed costs of trade – how much would a business-to-business e-commerce platform need to reduce fixed costs of sourcing from China (perhaps via a reduction in search costs) in order to benefit consumers in Nigeria? Small cost reductions actually decrease welfare, as the extensive margin shift to higher fixed cost sourcing strategies causes a reduction in the number of sellers that outweighs the gain from lower variable costs. Lagos benefits from modest fixed cost reductions due to international e-commerce, but fixed costs of sourcing from China have to fall by at least 75 percent before more remote parts of the country start to see a net benefit.

We join a small set of papers focused on how wholesale and retail distribution influence consumer welfare in developing countries (Atkin and Donaldson (2015); Lagakos (2016); Atkin, Faber and Gonzalez-Navarro (2018)). This relates more broadly to a literature on traders in agricultural value chains, and their role in determining price gaps

[3] Regulated agricultural marketplaces outside of which trading is restricted (Chatterjee and Mahajan (2021)).

between farmers and consumers (Fafchamps and Hill (2008); Dillon and Dambro (2017); Krishna and Sheveleva (2017); Casaburi and Reed (2019); Chatterjee (2019); Dhingra and Tenreyro (2020); Bergquist and Dinerstein (2020)). Although our empirical application is to manufactured goods, our framework is equally applicable to collection chains, such as those in agriculture. These literatures have typically studied prices and market power of intermediaries taking the structure of the chain as given. Three papers have documented explicit attempts to cut out a layer of intermediation – an NGO-driven effort to help Amazonian fishermen bring their catch to market themselves in Bartkus et al. (2022), a policy banning a type of intermediary in Bangladeshi oil distribution in Emran et al. (2021), and a Colombian start-up circumventing a wholesale market to sell directly to small fruit and vegetable vendors in Iacovone and McKenzie (2019). We offer a general approach to microfounding chains, to understand how structures respond endogenously to policy or technology changes and the resulting welfare consequences.

Wholesaling accounts for a substantial share of international trade, and analysis of customs microdata from a range of countries has documented that selling via an intermediary is more common in smaller transactions and when selling to destinations with higher access barriers. This has motivated models of exporting via a wholesaler as a way for manufacturers to reach consumers at lower fixed cost (Blum, Claro and Horstmann (2009); Bernard et al. (2010); Ahn, Khandelwal and Wei (2011); Crozet, Lalanne and Poncet (2013); Akerman (2018); Ganapati (2020)). More or less explicit in this literature is the idea that wholesalers serve an aggregation role, pooling exports from multiple firms or products to cover fixed costs. Aggregation in response to economies of scale is also at the heart of our view of intermediaries. We extend this logic to yield an endogenous structure of aggregation points, showing that chains with multiple intermediaries may naturally arise and how this matters for consumer welfare.

Conceptually, we build on a classic literature exploring the efficiency of free entry equilibria in the presence of internal scale economies (Spence (1976); Dixit and Stiglitz (1977); Mankiw and Whinston (1986)), and work considering these issues in new trade models (Dhingra and Morrow (2019); Matsuyama and Ushchev (2020)). In order to capture the nature of intermediation and resale, we extend this problem to consider the choice over a menu of resale "technologies" (i.e. source markets) featuring different combinations of fixed and variable costs. This introduces another margin along which social and private incentives can diverge, and there is room to achieve second-best outcomes by using policy to constrain these sourcing options.

The existence of intermediaries in goods trade has been attributed to a variety of specific mechanisms, often focused on information problems or the provision of financing or transport services (Rubinstein and Wolinsky (1987); Biglaiser (1993); Antras and Costinot

(2011); Bardhan, Mookherjee and Tsumagari (2013)). Economies of scale are a feature of a whole class of these intermediation tasks, regardless of the exact mechanism – the ways intermediaries actually solve these problems and the types of costs they incur in doing so often feature internal economies of scale. This can encompass information, regulatory, financial, bureaucratic, and transportation costs of trade. Earlier work shows that travel to overcome search and contracting problems is an important source of fixed costs of trade in this particular Nigerian context (Startz (2021)), but emprical evidence on the substance of trade costs and the existence of economies of scale comes from a wide range of work, including in trucking (Teravaninthorn and Raballand (2009)), container shipping (Cosar and Demir (2018)), trade finance (Niepmann and Schmidt-Eisenlohr (2017)), and licensing and non-tariff barriers.

We offer a framework for understanding the endogenous determination of distribution and collection structures, to enable both positive and normative predictions about changes in these structures. We begin in Section 2 by describing wholesale and retail trade in consumer goods in Nigeria, and introducing data that highlights the importance of multi-intermediary chains and documents novel empirical patterns associated with chain position. In Section 3, we introduce the main theoretical insights about economies of scale and chain formation in a simple geography. In Section 4, we show how this theoretical insight can be implemented in a quantitative equilbrium model in a more realistic geography, and in Section 5 we quantify that model in the Nigerian setting and consider policy counterfactuals.

2 Distribution chains in Nigeria

2.1 Wholesale and retail trade in Nigeria

The trading sector in Nigeria is large and economically important. As of 2013, wholesale and retail trade accounted for 17% of GDP and 25% of total employment, and was the largest contributor to recent GDP growth (Nigerian National Bureau of Statistics). The sector is highly decentralized, and composed mostly of small-scale informal enterprises—some estimates suggest that as much as 98% of spending on consumer packaged goods in Nigeria goes through small traditional outlets (Nielsen (2015)).[4] Nigeria's trading sector is not exceptional in its economic importance or form relative to other developing

[4]This stands in contrast to the trading sector in rich countries, characterized by larger formal firms and higher labor productivity (Lagakos (2016)). However, wholesaling is still economically important – Ganapati (2020) shows that wholesaling in the United States is increasingly consolidated in response to large fixed cost investments, but that the share of manufactured goods transactions in the United States that were intermediated by wholesalers actually rose from 32% to 50% between 1992 and 2012. We are not aware of data that allows for systematic comparisons of the structure of intermediation across countries.

countries; for instance, the comparable fraction of sales through traditional outlets is 96% in Ghana and 92% in India. In spite of interest in the very large and growing Nigerian consumer goods market among international firms, the presence of large-scale modern retailers is extremely limited. Even genuine branded goods tend to make their way to consumers via an "informal and fragmented" (Leke et al. (2014)) distribution system in which manufacturers exercise "little control over the rest of the distribution chain" (Nielsen (2015)).

2.2 Lagos Trader Survey data

We collected original data that captures part of the distribution chain for manufactured consumer goods imported into Nigeria. The Lagos Trader Survey (LTS) is a panel survey of wholesale and retail traders in Lagos, Nigeria, and includes information about their international and domestic transactions.[5]

LTS participants were identified through a census of over 50,000 shops in commercial areas of Lagos. Interviews were conducted with traders whose shops were randomly sampled from the census.[6] The sample includes any trader dealing in manufactured consumer goods (excluding food), which we group into six product categories: apparel (including shoes, bags, and textiles), electronics, toiletries and beauty products, hardware, home goods, and miscellanous other products. These goods account for roughly 17 percent of consumer spending in Nigeria (Euromonitor (2010)). We use data from the second wave of the LTS, which covered the activities of a sub-sample of 595 traders in 2015, and focus mainly on importers, for whom we have the most detailed data about transactions.

Table 1 shows summary statistics about the wholesale and retail trading businesses captured in the LTS data, split by importer status. Consistent with the general picture of Nigeria's trading sector as fragmented, we find that traders' businesses are small, owner-operated wholesale and retail firms. The median firm has been in operation for nine years, and has one shop and employs one worker in addition to the owner. Ninety-three percent of traders have just one or two shops, and these account for 73 percent of employment (not including owners) and 82 percent of the value of goods purchased. On average, traders purchased approximately US $60,000 in stock to resell in 2015.

The data captures traders' purchases from international suppliers at the transaction level, and identifies where suppliers are located and whether they are manufacturers or wholesalers.[7] Over thirty different source countries are represented in the data; the largest

[5] More details about the Lagos Trader Survey can be found in Startz (2021).

[6] The listing focuses on commercial and wholesaling areas of the city, and does not include most residential or manufacturing areas or traditional food markets.

[7] All information is reported by the interviewees, not through direct contact with suppliers. On net, we expect that this leads to overestimates of the extent to which traders buy directly from producers, as they may assume any distributor of a known brand is the "manufacturer" of that brand.

Table 1: Descriptive statistics

		All	Non-Importer	Importer
Years in operation	Mean	10.36	10.32	10.40
	(Median)	(9)	(8)	(10)
Number of shops	Mean	1.42	1.33	1.51
	(Median)	(1)	(1)	(1)
Number of workers	Mean	1.10	0.89	1.31
	(Median)	(1)	(0)	(1)
Cost of stock purchased ($US/year)	Mean	61,097	54,924	67,252
	(Median)	(19,134)	(12,344)	(33,947)
Wholesale fraction of total sales	Mean	0.54	0.45	0.61
	(Median)	(0.50)	(0.50)	(0.60)

Notes: All variables refer to 2015. There are 595 observations for all variables except cost of stock purchased, for which N=475 due to item non-response in the survey. Observations are reweighted by sampling probabilities.

is China, but the United Arab Emirates, Turkey, Hong Kong, Benin, India, and the United States are also common. Overall, 63% of international suppliers are wholesalers rather than manufacturers. For each transaction, we observe the product type and quantity purchased, the cost paid to the supplier, and the average price the trader charged when selling that particular good. For each shipment[8], we observe various components of trade costs, including transportation, clearing the port, hiring agents, travel to the source location, and the total value of assorted other costs, such as financial fees. These transaction- and shipment-level data enable us to construct measures of unit costs, trade costs, and markups, and to relate these to supplier type.

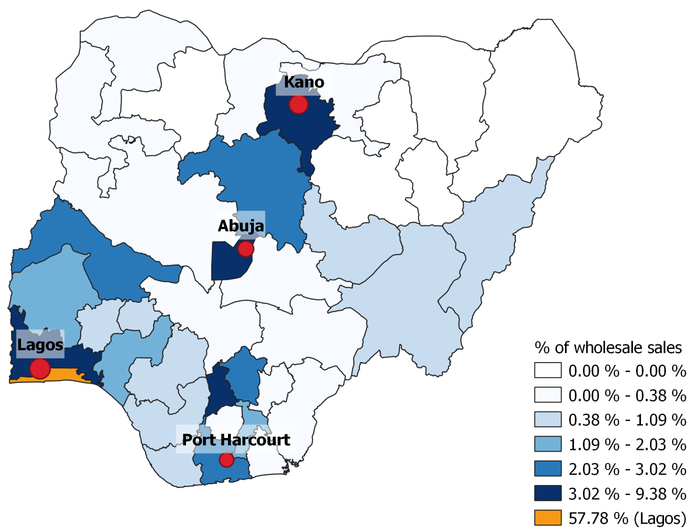
Looking downstream, we observe the fraction of traders' sales that are retail versus wholesale, and, starting from 2016, whether they have sales to locations outside of Lagos. In total, 87% of traders in the LTS do some wholesaling (96% of importers), and on average 54% of traders' sales are wholesale (61% for importers).[9] A little over half of traders' wholesale sales are within Lagos state[10], but the remainder go to other destinations, consistent with Lagos' role as the commercial capital and main port of entry for Nigeria. Figure 1 shows the fraction of wholesale sales that go to each state in Nigeria. Many sell to downstream traders in locations that serve as commercial hubs for other parts of the

[8] Shipments may include multiple transactions, involving more than one product or supplier.

[9] The preponderance of wholesaling in the sample is unsurprising due to the focus on commercial areas of the city and restriction to businesses in permanent physical premises. The tens of thousands of small shops in residential areas and mobile hawkers throughout the city are excluded from the sample, and are likely to purchase their supplies from the type of wholesalers in the LTS data.

[10] Lagos state accounts for roughly a third of Nigeria's national GDP (Lagos Bureau of Statistics).

Figure 1: Downstream sales in Nigeria



country such as Kano, which is the commercial center in the north, or Port Harcourt in the southeast.

2.3 Stylized facts about Nigerian distribution chains

The LTS data follows goods across multiple links in the distribution chain, conditional on passing through the hands of a Lagos trader. This allows us to construct a measure of chain length. We define purchases from manufacturers as one end point on the chain, and retail sales to final consumers as the other. Purchases from or sales to other traders are additional points of intermediation. The data allows us to observe chains with a minimum length of three and maximum length of five. When all of a trader's suppliers are manufacturers and all of her sales are retail, this a chain of length three with one intermediary, the trader herself. A chain in which all suppliers are wholesalers and all sales are wholesale has a length of at least five, with at least three intermediaries – the trader and intermediaries she buys from and sells to.[11]

In this section, we make use of the fact that the data allow us to track multiple vertical links to document a set of novel facts about the intermediaries and distribution chains that bring internationally manufactured goods to Nigerian consumers.

[11]Our measures of chain length are truncated at length four and five, becase we do not observe the supplier's supplier or the buyer's buyer. The chains that we observe are therefore actually a lower-bound on the true length of intermediation chains passing through Lagos.

Table 2: Chain Length for Importers

	Chain length (mean)	Steps upstream of trader	Steps downstream of trader
All	4.23	1.63	1.62
Apparel	4.20	1.65	1.56
Electronics	4.29	1.78	1.51
Beauty	4.30	1.50	1.81
Hardware	4.20	1.62	1.60
Homewares	4.21	1.60	1.66

Note: Observations are at the transaction level, with averages reweighted by sampling probabilities. Only upstream import transactions are included. Steps upstream are the (truncated) number of agents on the chain above the Lagos trader, and are observed at the transaction level. Steps downstream are the (truncated) number of agents on the chain below the Lagos trader, and are observed at the trader level and allocated to all transactions conducted by that trader.

Fact #1: Nigerian consumers are served by multi-intermediary chains

The first fact to come out of the data is that Nigerian consumers are served by multi-intermediary chains. Table 2 shows average chain length for importers, and separately for each product category. Chains have, on average, at least two or three independent intermediaries between a foreign producer and a domestic Nigerian consumer. This reflects an approximately even split between intermediaries upstream and downstream of the Lagos-based trader. Although there is some variation across products – traders dealing in electronics are somewhat further downstream from manufacturers than the average, while those in beauty and cosmetics are less so – there are multiple layers of intermediation in all product categories.

Fact #2: Smaller traders are more likely to buy in resale markets, and pay higher prices and lower fixed costs

Disaggregating the details of traders' sourcing decisions, we observe patterns suggestive of scale economies that vary systematically across source markets. Overall, two-thirds of suppliers are resellers rather than manufacturers, but this total reflects substantial source market-level variation. For instance, while Lagos traders report large trade flows of clothing purchased from Benin and electronics from Dubai, it seems unlikely that those goods were originally produced in those locations. Consistent with a broad characterization of these

Table 3: Relationship between supplier type and firm size and sourcing costs

	(1) Log unit cost ($US)	(2) Fixed trade costs ($US)	(3) Log revenue ($US)	(4) Number of workers	(5) Number of shops
% of purchases from wholesaler	0.49**	-454.14	-0.66**	-0.69***	-0.17
	(0.22)	(322.78)	(0.32)	(0.24)	(0.11)
Obs.	650	650	178	346	346
Mean of dependent variable	1.99	1160.47	9.99	1.36	1.54
Product category FEs	x	x	x	x	x

Note: In each column, observations are at the level at which the outcome variable is measured. Columns (1) - (3) have one observation per trader, while columns (4) and (5) have one observation per imported shipment. Observations with variable profit ratios (sale price divided by purchase price) below the 10th or above the 90th percentiles are trimmed in (4) and (5). Standard errors are clustered at the firm level.

markets as entrepôts, suppliers there are heavily reported as wholesalers.[12]

Table 3 shows the relationship between buying from a wholesaler (versus a manufacturer) and traders' firm size and trade costs incurred in sourcing. Column (1) shows that, within a product category, traders pay higher unit costs when buying from a wholesaler. Column (2) shows that they also pay lower fixed costs, measured here as reported costs of traveling to the source market and fees paid to clear the port and hire agents.[13] If buying from a wholesaler allows for a tradeoff between lower fixed and higher variable costs, we should expect larger firms to source more directly and smaller firms to source indirectly.[14] Consistent with this, we see in Columns (3), (4) and (5) that traders who source from wholesalers have smaller businesses, whether measured in terms of total sales, number of workers, or number of shops.

Fact #3: Upstream firms charge lower markups

Traders vary not only in their sourcing strategies, but also in the extent to which they directly serve retail customers versus selling onward to other resellers. Column (1) of

[12] Even in major manufacturing locations such as China and Germany, a substantial fraction of suppliers are wholesalers. This is consistent with evidence from customs data that a large fraction of exports in many countries, including China, are via wholesalers (Ahn, Khandelwal and Wei (2011)).

[13] There is no clear rule for which trade cost categories can be considered fixed versus variable. However, the overall pattern shown is robust to the inclusion or exclusion of different cost categories, or to simply allowing the type of seller to modify the relationship between value of goods purchased and total trade costs, as shown in Appendix Table 7.

[14] Note that the causality driving this relationship should run both ways – a firm that is smaller for reasons unrelated to trade costs should select into a lower fixed cost sourcing strategy, and that strategy, because it also involves higher variable costs, should cause the firm to sell less and be smaller.

Table 4: Relationship between chain position and pricing

	(1) % of purchases from wholesaler	(2) Log unit cost ($US)	(3) Markup (%)
% of sales that are wholesale	-0.14	-0.79**	-0.33***
	(0.11)	(0.35)	(0.12)
% of purchases from wholesaler		0.12	0.01
		(0.17)	(0.06)
Log unit cost ($US)			-0.09***
			(0.02)
Obs.	715	714	714
Mean of dependent variable	1.62	2.07	0.67
Product category FEs	x	x	x

Note: All columns have observations at the import transaction level. Observations with variable profit ratios (sale price divided by purchase price) below the 10th or above the 90th percentiles are trimmed. Standard errors are clustered at the firm level.

Table 4 shows that traders' upstream and downstream positions are negatively correlated. Traders who are further upstream from final consumers – i.e. those who sell more wholesale relative to retail – are more likely to source directly from manufacturers themselves.[15]

Column (2) of Table 4 shows that traders who are further upstream pay lower prices to their suppliers, conditional on their downstreamness. They also charge lower markups to their customers, even when controlling for the fact that they face lower unit costs on average. Column (3) implies that going from entirely retail sales to entirely wholesale sales is associated with a 33 percentage point reduction in average markups. This is large, but the magnitude is roughly consistent with traders' self-reports that their wholesale prices are on average 10% lower than their retail prices. Taking these two facts together – that more upstream traders pay lower prices to their suppliers and charge lower markups – is suggestive of demand elasticities that vary systematically along the chain, increasing as you move upstream away from final consumers.

In sum, Nigerian traders are small and independent, and chains serving Nigerian consumers are long, with on average at least two or three middlemen. The characteristics of these traders and their transactions vary systematically with their position on the distribu-

[15]Note that a negative correlation between upstreamness and downstreamness is not mechanical. If the dominant pattern were that some traders are involved in longer chains than others, we could observe a positive correlation, as Antras and Chor (2018) show for value chains.

tion chain, and the empirical patterns we observe are consistent with both source-specific economies of scale and more elastic demand (and therefore lower markups) at upstream points along the chain. In the following sections, we show that the presence of internal economies of scale in trade costs can account for all of these facts.

3 A Simple Model of Intermediation

To understand why and how chains with multiple intermediaries might form, we build a framework in which consumers in a destination market demand a good that is produced elsewhere. That good can either be sourced directly from a producer, or indirectly through a resale market. We begin by showing that economies of scale in trade costs can lead to equilibria in which intermediaries maximize profits by sourcing indirectly. We then ask how these indirect sourcing equilibria affect consumers in the destination market, and whether consumers benefit from policies or technologies that induce a shift to a direct sourcing equilibrium. We work in a stylized setting with a broad demand specification in order to highlight the main forces involved. In the following section, we complicate the setting and impose more structure on the model in order to bring it to data.

3.1 Model setup

3.1.1 Consumer demand

We consider the distribution of a single homogeneous good. This good is produced in one location, the "origin", and is demanded by consumers in a destination market. Consumers purchase the good from an endogenous number of local traders, N.[16] Each trader v sells a quantity q_v. We assume a partial equilibrium approach is justified and normalize the marginal utility of income to one, so that the consumer payout from this good is given by

$$U = G\left(\sum_{v=1}^{N} f\left(q_v\right)\right) \tag{1}$$

where $G'(\cdot) > 0$, $G''(\cdot) < 0$, $f(0) = 0$, $f'(\cdot) > 0$ and $f''(\cdot) \leq 0$. We also assume that preferences are such that firm choice of quantity has the usual second order condition, and that equilibria are stable in the sense of Hahn (1962).[17][18] These preferences nest commonly

[16]The assumption that consumers buy only in their home market is simplifying but not necessary, and is equivalent to assuming parameters take values such that it is not cost-effective for a single consumer to source from other markets.

[17]The stability condition requires that marginal revenue is weakly decreasing in the competition (as defined in Section 3.1.3) faced by an individual firm. We provide formal statements in the Appendix.

[18]This form of utility follows Spence (1976) and Mankiw and Whinston (1986), with the addition of the stability condition. This ensures that equilibria are unique, although uniqueness is relevant only in Section

used demand specifications, including linear and CES, and constant passthrough demands more broadly. They also allow for sellers to be either small or large with respect to the market, and for consumers to perceive those sellers as either perfectly or imperfectly substitutable. Imperfect substitutability across sellers implies that consumers may experience gains from *seller* variety even though the good itself is homogeneous. These gains may arise from sellers' different locations within a market, or other factors that affect individual consumers' preferences over purchasing from one seller versus another.[19]

3.1.2 Sourcing technology

Traders are intermediaries who source the good and resell it without transformation. They can purchase the good in the origin, where it is produced. However, the same good can also be sourced from other markets, where it is being resold. The costs of sourcing the good in market i and selling it to consumers in the destination are (c_i, F_i), where c_i represents a constant variable cost per unit of the good and F_i is a strictly positive fixed cost that does not vary with the number of units purchased.[20]

The variable cost includes the price paid to a supplier in the source market, as well as any variable trade costs incurred to get the good to the destination market and sell it there, such as transportation costs, insurance, tariffs, and so on. The fixed cost captures economies of scale that empirical evidence suggests are often present in trade costs. For instance, this may include the cost of traveling to a source market to search for a supplier, fees paid at borders or to access letters of credit, time spent gathering information, the cost of operating a truck, and so on. For the time being, we assume that traders take all costs as exogenous. Implicitly, this assumes that the destination market is small and does not affect upstream prices, but this is only to focus attention on the key forces; we endogenize upstream pricing in the following section of the paper.

This setting implies that traders face alternative technologies for serving the destination market with the same good, which feature different combinations of variable and fixed costs. Therefore, we can think of a technology frontier of source markets indexed by j, consisting of the origin, o, and the set of resale markets that have non-dominated cost

3.3.2, when we consider the welfare implications of discrete changes in sourcing technology. Seade (1980) explores some of the (unusual) demand structures that are ruled out by this type of assumption.

[19] For example, consider an address model in which consumers have ideal variety preferences across sellers that cause them to prefer their nearest retailer, but with some willingness to trade off price against time to reach other sellers. This is consistent with the findings of Atkin, Faber and Gonzalez-Navarro (2018) and Lagakos (2016), both of which suggest that developing country consumers consider retail outlets to be imperfect substitutes, and the latter of which connects this explicitly to tradeoffs between price and time/distance.

[20] We assume this form of economies of scale for simplicity of exposition, but the results extend readily to more general decreasing marginal costs of sourcing.

combinations. So long as markups in resale markets are positive (i.e. the good is sold for more than variable cost, which is necessary for firms to cover their fixed sourcing costs) and a triangle inequality holds on any variable trade costs,[21] then variable cost must be higher in all resale markets than in the origin, $c_{j\neq o} \geq c_o$. This in turn implies that the fixed costs for all resale markets on the frontier are lower than the origin, $F_{j\neq o} < F_o$. If this were not the case for any resale market, it would be cost-dominated by the origin, and would not be on the frontier. Buying directly from the producer is therefore a higher fixed cost, lower variable cost sourcing technology compared to buying through a resale market on the frontier.

3.1.3 Traders optimization and destination market equilibrium

The number of traders serving the destination market is endogenous. There is a large pool of identical potential traders, who can pay a fixed cost $f^e \geq 0$ to enter the market. After entry, each trader v chooses a source market j and a quantity q to buy and resell, in order to maximize their profits:

$$\pi_v\left(q_v, q_{-v}, j, N\right) = \left(p_v - c_j\right) q_v - F_j - f^e \tag{2}$$

where $p_v \equiv p_v\left(q_v, q_{-v}, N\right)$. The profits, π_v, and price, p_v, of a firm v depend on the quantity choices of all other firms and the number of competitors. However, the form of utility means that we can capture the impact of all other firms' decisions on the profits of a given firm via a single variable, θ_v, which we will refer to as the level of competition faced by firm v. Formally, we define $\theta_v \equiv \sum_{v'\neq v} f\left(q_{v'}\right)$ when firms are large. [22] This then permits us to express profits when sourcing from j as

$$\pi_v\left(q_v, \theta_v, j\right) = \left(p_v - c_j\right) q_v - F_j - f^e$$

[21]The triangle inequality says that it is not less expensive to trade via a third location than to trade directly, e.g. $\tau_{jk} \leq \tau_{jl}\tau_{lk}$ if variable trade costs are denoted by a multiplicative τ_{jk} between locations j and k. While the triangle inequality is not essential – if it does not hold, then there is simply an additional reason that goods may be routed through resale markets – it is an especially innocuous assumption in our setting since we do not consider physically indirect routing that does not involve a transaction to be a chain link. Therefore, most cases that might appear to be violations of the triangle inequality need not be as long as τ_{ij} is defined as the cost over the least-cost actual trade route between i and j.

[22]When firms are small, $\theta_v = \int_0^N f\left(q_{v'}\right) dv'$.

where $p_v = p_v(q_v, \theta_v)$.[23]

Traders enter up to zero profits. We focus on symmetric equilibria in which $q_v = q$ for all v. Ignoring integer constraints on the number of firms, an equilibrium in entry, sourcing, and quantity choices is described by the following conditions:

$$\pi(q_j, \theta_j, j) = 0 \tag{3}$$

$$\pi(q_j, \theta_j, j) \geq \pi\left(q'_{j'}, \theta_j, j'\right) \ \forall j' \tag{4}$$

$$\frac{\partial}{\partial q_j}\pi(q_j, \theta_j, j) = 0 \tag{5}$$

where q_j and θ_j are the equilibrium per-firm quantity and level of competition when sourcing from j. Furthermore, $q'_{j'}$ is the quantity that would be chosen by a trader if it deviated to sourcing from market j', holding all other traders' choices fixed, and $\frac{\partial}{\partial q_j}\pi(q_j, \theta_j, j)$ is the partial derivative of profits holding θ_j fixed.

There are two types of symmetric equilibria: "direct sourcing", in which all traders in the destination source from the origin, and "indirect sourcing", in which they source from a resale market. In a direct sourcing equilibrium, the destination is served via a chain with one intermediary – the trader in the destination – who sources from the producer. In an indirect sourcing equilibrium, the chain has at least two intermediaries – the destination trader and a seller in the resale market – and may have more depending on where the resale market sources the good. We have not yet modeled upstream markets explicitly, but one can imagine that there are traders in the resale market who face an analogous choice between sourcing from the origin and some other set of resale markets. If the equilibrium there also features indirect sourcing, then chains can grow to include more than two intermediaries. In Section 4, we follow this logic through to model upstream wholesale markets and measure the full length of the distribution chain.

[23] To see why θ_v captures the impact of other firms' entry and quantity decisions on the price (and hence profits) of firm v, note that given the form of demands assumed in 3.1.1,

$$\begin{aligned} p_v &= \frac{\partial U(q_v, q_{-v}, j, N)}{\partial q_v} \\ &= G'\left(\sum_{v'} f(q_{v'})\right) f'(q_v) \\ &= G'(f(q_v) + \theta_v) f'(q_v) \end{aligned}$$

3.2 Indirect sourcing

When will the destination market be served by indirect sourcing? The choice across sources is fundamentally a trade-off between fixed and variable costs. In order for there to be an indirect sourcing equilibrium, it must be that traders find the lower fixed cost associated with indirect sourcing to outweigh the losses from paying a higher variable cost.

A necessary condition for an indirect sourcing equilibrium to exist is that no trader wants to deviate to direct sourcing.[24] To build intuition, we compare the profits from indirect sourcing to those from direct sourcing (holding other firms' choices constant) using a second-order approximation to profits with respect to variable costs. An indirect sourcing equilibrium can exist when:

$$\frac{F_o - F_{j\neq o}}{c_{j\neq o} - c_o} \geq q_j \left[1 + \frac{1}{2}\left(\frac{c_{j\neq o} - c_o}{c_{j\neq o}}\right)\left(\varepsilon_p^q - 1\right)\rho\right] \tag{6}$$

where q_j and c_j are per-firm quantity and unit cost, ε_p^q is the price elasticity of demand, and ρ is the passthrough rate, all under a symmetric indirect sourcing equilibrium from location j.[25]

If the destination is served via indirect sourcing, then Equation 6 must hold for the resale market in question. The first order forces driving sourcing are the relative cost advantages of the origin versus resale markets, and the quantity being sourced. Indirect sourcing is more attractive when the fixed cost advantages of the resale market are large $(F_o - F_{j\neq o})$ relative to the additional variable costs incurred $(c_{j\neq o} - c_o)$ – or, conversely, when the fixed cost barriers to accessing production locations are relatively high. Indirect sourcing is also more attractive when equilibrium firm size is small (i.e. q_j is low), because the total value of the variable cost disadvantage is small relative to the quantity-invariant fixed cost advantage. The second-order forces driving indirect sourcing are the price elasticity of demand facing individual traders and the passthrough rate. When the price elasticity of demand or the passthrough rate are lower, individual firms get less of a payout from deviating to a lower variable cost strategy.

Equation 6 is a statement about equilibrium relationships,[26] but has substantive empirical content. Small firm size and high barriers to accessing production locations are

[24] This condition is necessary but not sufficient. In order for there to be an equilibrium sourcing from any given resale market j, it must also be that no firm wants to deviate to any other resale market, but these sufficient conditions do not provide any additional insight into the question of direct versus (any) indirect sourcing.

[25] Formally, we define the passthrough rate as $\rho = \frac{\partial p}{\partial c}$ and the elasticity of demand as $\varepsilon_p^q = -\frac{p}{q}\frac{\partial q}{\partial p}$.

[26] We keep the form of demand and competition intentionally general in this section, but under narrower assumptions, similar statements can be made in terms of comparative statics with respect to parameters, rather than equilibrium relationships.

characteristics of many markets in developing countries. This suggests that indirect sourcing is more common in these places. This result is not necessarily driven by any assumption that there are market failures specific to developing countries – only that the characteristics of demand and costs are such that firms are small, and that this is consistent with the existence of indirect sourcing equilibria. However, if there *are* other market failures that further limit the size of firms operating in developing countries – such as span of control problems or credit constraints – these will also push toward more indirect sourcing and longer chains by further lowering q.

3.3 Welfare implications of cutting out the middleman

How are consumers affected by sourcing goods via a resale market? We consider the problem of a planner who values welfare in the destination market, and who manipulates sourcing options but takes traders' entry and quantity decisions as given. We show that the sourcing strategy chosen by the market equilibrium may be different from the (second-best) planner's solution, and policy interventions that induce more direct sourcing can either benefit or harm consumers at the end of the chain.

3.3.1 Continuous sourcing technology

Sourcing technologies represent markets, which are inherently discrete objects. However, it is useful to start by imagining that there is a continuous technology frontier, so that we can use derivatives to characterize the margins on which private incentives diverge from social ones. We denote the frontier in terms of variable cost as a function of fixed cost, $c(F)$, and assume that $c'(F) < 0$ and $c''(F) > 0$ and $F > 0$. The market equilibrium (ME) is defined as in Section 3.1.3, except that due to the continuity of the cost function, Equation 4 is now:

$$\frac{\partial \pi(q, \theta, F)}{\partial F} = 0 = c'(F)\, q + 1 \tag{7}$$

The planner can select the sourcing technology used by traders in the destination market. What happens if the planner moves the market along the frontier to a slightly higher level of fixed cost, starting from the market equilibrium? (Because we focus on identical traders in a symmetric equilibrium, in which $q_v = q$ for all v, we suppress v subscripts from this point forward.)

Under any technology, traders will enter up to zero profits and choose quantity to maximize profits. There is therefore a constrained equilibrium, defined by the technology chosen by the planner in lieu of 7, while Equations 3 and 5 continue to hold. In this equilibrium, the marginal cost, the number of firms and per-firm quantity, and the level of competition are functions of the F chosen by the planner, i.e. $c(F)$, $N(F)$, $q(F)$,and $\theta(F)$.

Lemma 1: A small change in the fixed cost around the market equilibrium yields the following comparative statics: $\frac{d}{dF}q(F_{ME}) > 0$, $\frac{d}{dF}N(F_{ME}) < 0$, and $\frac{d}{dF}\theta(F_{ME}) = 0$, where F_{ME} is the level of fixed cost chosen in the market equilibrium.

This result follows non-trivially from the form of demand described in Equation 1, and we leave a complete derivation to the Appendix. The intuition, in brief, is that under free entry the overall level of competition must be unchanged due to an envelope condition on the technology choice under the ME. If competition is unchanged, but marginal cost falls, the profit-maximizing quantity must increase in order for marginal revenue to continue to equal marginal cost. Finally, if quantity per firm increases and competition is unchanged, then entry must fall.[27]

With these comparative statics in hand, we can turn to how welfare changes in response to this change in technology. Destination market welfare is consumer surplus, as profits are zero in equilibrium:

$$W = U - Npq$$

Will the planner's preferred technology lie above or below the market equilibrium along the frontier?

Proposition 1. If the assumptions on the form of demand from Section 3.1.1 hold, and there is a continuous sourcing cost frontier, the planner's preferred sourcing technology may have either a higher or lower fixed cost than the market equilibrium technology.

Proof. See Appendix.

To understand the intuition behind this result, consider the following decomposition of the welfare effect of a move to a higher F technology (the derivation of which is shown in the Appendix):

$$\frac{dW}{dF} = \left(\frac{\partial U}{\partial N} - pq\right)\frac{dN}{dF} + N(p-c)\frac{dq}{dF} - N\left(c'(F)q + 1\right) \tag{8}$$

The three terms on the right side of Equation 8 describe the change in welfare from a move to a higher fixed cost technology resulting from the change in the number of

[27]Note that this resembles a more familiar problem in which there is a single technology and we take a comparative static with respect to F while holding c constant. In that case, the result is different and more trivial – an increase in F decreases profits, and so there must be exit under free entry, and the level of competition falls. In contrast, in this setting, F and c move together, and traders have already optimized with respect to the technology choice. The combination of the envelope condition on technology and the zero profit condition imply that competition does not change in response to local technology changes around the market equilibrium.

sellers ($\left(\frac{\partial U}{\partial N} - pq\right) \frac{dN}{dF}$), the change in the quantity supplied per seller ($N\left(p-c\right)\frac{dq}{dF}$), and the change in cost conditional on total quantity supplied ($N\left(c'\left(F\right)q+1\right)$), respectively. While the decomposition is general, in order to sign terms it is useful to focus on a small change in technology starting from the ME, as we do in Lemma 1. We take each term in turn:

Term 1 – number of sellers: The first term captures the welfare gain or loss from the change in the number of sellers. The the social benefit of an additional seller is $\frac{\partial U}{\partial N}$, while the social cost is pq. Under our assumptions on preferences, utility is weakly increasing in seller variety, so that $\frac{\partial U}{\partial N} - pq > 0$.[28] We showed in Lemma 1 that $\frac{dN}{dF} < 0$, so the number of sellers is decreasing. Therefore, this term is negative, and an increase in fixed cost reduces welfare due to a reduction in the number of sellers.

Term 2 – quantity per seller: The second term corresponds to the welfare gain or loss from the change in quantity per seller. Markups, $p-c$, capture the difference between the social benefit and social cost of an additional unit. Since markups are strictly positive and $\frac{dq}{dF} > 0$, the entire second term is positive. Therefore, welfare increases due to a reduction in the quantity distortion arising from market power.

Term 3 – cost minimization: The third term describes the change in firms' costs conditional on the number of sellers and the quantity per firm. This term must be zero because of the envelope condition on the cost-minimizing technology choice at the ME, shown in Equation 7.

The fact that the third term is zero tells us that the planner's technology choice and the market technology choice will not generally coincide. We can understand the third term as capturing the part of consumer welfare that is aligned with firms' profit maximization decisions – minimizing total cost conditional on the total quantity supplied in the market – while the first and second terms capture aspects of consumer welfare that are not internalized by firms. The first-order condition of the planner's problem will set the entirety of Equation 8 to zero, while the third term alone is set to zero in the ME. Therefore, the ME sourcing technology will not be efficient, except in special cases where the first two terms are both zero or cancel one another out. The former arises in the limit as the market tends toward perfect competition, and the latter under CES demand with monopolistic competition, which is a knife-edge case in which changes in the variety and quantity distortions are perfectly offsetting.[29] Otherwise, the planner is willing to accept some distortion of cost minimization in order to counterbalance net variety and quantity

[28] $\frac{\partial U}{\partial N} - pq = G'\left(Nf\left(q\right)\right)\left[f\left(q\right) - qf'\left(q\right)\right]$ where $f\left(q\right) - qf'\left(q\right) \geq 0$ due to the concavity of $f\left(\cdot\right)$ and $G'\left(\cdot\right) > 0$ by assumption.

[29] The market equilibrium under CES demand with monopolistic competition features insufficient entry, as in Spence (1976). A planner who could adjust the number of firms directly would want to do so, but changes to the sourcing technology are not an equivalent instrument.

distortions that arise in the market equilibrium.

The total welfare effect of a movement along the technology frontier is the net of the first two terms on the right side of Equation 8, which have opposite signs when starting from the market equilibrium. Moving to a higher fixed cost reduces seller variety and mitigates the quantity distortion. The planner prefers a higher fixed cost sourcing technology if the quantity distortion dominates, or a lower one if the variety distortion dominates. The concavity of welfare with respect to fixed costs means that if the planner wishes to lower (or raise) the fixed cost locally, then the planner's solution must lie below (or above) the market equilibrium, as stated in Proposition 1.

The decomposition in Equation 8 is in terms of quantity, but we can derive an equivalent result in terms of the (symmetric) price charged by all firms. The sum of the second and third terms of Equation 8 is $-Nq\frac{dp}{dF}$ and is therefore directly related to the change in the equilibrium price, which is:

$$\frac{dp}{dF} = \underbrace{\frac{\partial \mu}{\partial q}\frac{dq}{dF}}_{\substack{=-\frac{(p-c)}{q}\frac{dq}{dF} \\ \propto \text{ Eqn 8 term 2}}} + \underbrace{\frac{\partial \mu}{\partial F} + c'(F)}_{\substack{=0 \\ \propto \text{ Eqn 8 term 3}}} \tag{9}$$

The change in the markup due to the change in per firm quantity ($\frac{\partial \mu}{\partial q}\frac{dq}{dF}$) is negative – Lemma 1 established that $\frac{dq}{dF} > 0$ and under the free entry condition, $\mu = \frac{F}{q}$, and so $\frac{\partial \mu}{\partial q} < 0$. The change in the markup due to the change in fixed cost holding quantity constant ($\frac{\partial \mu}{\partial F}$) is positive (again, due to the free-entry condition), and the change in the variable cost ($c'(F)$) is negative given the definition of the technology frontier. These last two terms combined correspond to the third term of the quantity decomposition in Equation 8 and add up to zero for the same reason: free entry implies that price is equal to average cost, and the envelope condition on technology choice at the ME that implies there is no change in average cost from a small change in technology. Taken together with the negative first term, this implies that the equilibrium price is decreases with a small change to higher fixed cost technology around the ME.[30]

The fact that we can sign the direction of the price change, however, holds only around the ME on a continuous technology frontier – we will see in the following section that the result becomes ambiguous in the policy-relevant case with discrete technologies. Furthermore, the fact that price is decreasing does not imply that market power is decreasing. As noted in Lemma 1, the overall level of competition is unchanged. By grouping the μ terms

[30] In fact, there is an even simpler argument: per Lemma 1, competition is unchanged but per-firm quantity is increasing. Therefore prices must fall.

together in Equation 9, we can see that the effect on markups is ambiguous.

$$\frac{d\mu}{dF} = \frac{\partial\mu}{\partial q}\frac{dq}{dF} + \frac{\partial\mu}{\partial F}$$

where $\frac{\partial\mu}{\partial F} > 0$, $\frac{\partial\mu}{\partial q} < 0$, and the net depends on the magnitude of $\frac{dq}{dF}$.[31] That magnitude depends on the passthrough rate, and so markups may rise or fall.

3.3.2 Discrete sourcing technology

In reality, the technology frontier is made up of a finite number of potential source markets, and the relevant changes in sourcing are discrete, rather than continuous. The planner can dictate sourcing (or not sourcing) from a particular market, which may involve a large change in fixed and variable costs relative to the source chosen by the market equilibrium. We are particularly interested in whether the planner should cut out middlemen by requiring direct sourcing, or conversely, by discouraging sourcing from resale markets. Does this type of policy benefit consumers in the destination?

As in the continuous setting of Section 3.3.1, there is a constrained equilibrium that arises under the technology dictated by the planner, with entry and quantity-setting conditions continuing to hold.

Lemma 2: A discrete change to a higher fixed cost technology starting from the market equilibrium leads to a constrained equilibrium with higher quantity per trader, fewer traders, and lower competition relative to the market equilibrium: $\Delta q > 0$, $\Delta N < 0$, and $\Delta\theta \leq 0$.

The intuition comes in three parts, and we provide a proof in the Appendix. First, if firms are forced to adopt a different level of fixed cost than they would have chosen with the level of competition held fixed, then they must earn (weakly) negative profits by revealed preference. As per-firm profits are strictly falling in the level of competition, it must be that the level of competition in the new constrained equilibrium is (weakly) lower so that the free entry condition continues to hold. Second, at a new, higher level of fixed cost, variable costs will be strictly lower, which implies marginal revenue must also fall. However, the stability condition implies that the fall in competition pushes marginal revenue higher, so that per-firm quantity must increase. And third, since competition falls and per-firm quantity is higher, it must be that there are fewer firms.

The distinction between Lemma 1 and Lemma 2 is that in the former case, we consider

[31] Under free entry, $\frac{dq}{dF} = \frac{q}{F} \cdot qc'(F) \cdot \frac{\partial p}{\partial c}$. If $\frac{\partial p}{\partial c} > -\frac{1}{qc'(F)}$ then $\frac{d\mu}{dF} > 0$, and otherwise $\frac{d\mu}{dF} \leq 0$. Since passthrough can be arbitrarily large and is bounded below by 0, it is easy to find examples where markups increase or decrease in F.

small changes around the market equilibrium, so that an envelope condition on the level of fixed cost holds. In the latter case, the fixed cost frontier is discrete, and so this envelope condition does not hold. The conclusions of both lemmas are very similar, although the different impacts on the level of competition are important for the welfare consequences. We can now turn to the planner's problem in the discrete case, and ask whether the planner's preferred technology will be above or below the level the market equilibrium.

Proposition 2. If the assumptions on the form of demand from Section 3.1.1 hold and the cost frontier is discrete, a change to a higher fixed cost technology starting from the market equilibrium may increase or decrease welfare.

Proof. See Appendix.

To understand Proposition 2, consider the change in welfare due to a shift from an equilibrium with sourcing from j to one with sourcing from j', which can be decomposed as:[32]

$$\Delta W = \int_{N_j}^{N_{j'}} \left[\frac{\partial}{\partial N} U(\nu, q_j) - p_j q_j \right] d\nu + N_{j'} \int_{q_j}^{q_{j'}} \left[p\left(N_{j'}, \vartheta\right) - c_{j'} \right] d\vartheta - N_{j'} \left(q_j \Delta c + \Delta F \right) \tag{10}$$

where a j subscript denotes the equilibrium value of the given variable when firms source from j, and Δ denotes the difference in the variable when sourcing from j versus j' (i.e. $\Delta N = N_{j'} - N_j$). This is equivalent to the decomposition in equation (8), but accounts for the fact that a change in sourcing technology will involve discrete jumps in fixed and variable costs. The three terms on the right side of the equation – describing changes in welfare due to entry, quantity per firm, and cost conditional on quantity, respectively – have the same interpretation as those in the continuous version.[33]

What happens to welfare if the market equilibrium selects indirect sourcing, but the planner mandates direct sourcing? Since all resale markets on the frontier have lower fixed costs than the origin market, direct sourcing therefore involves an increase in fixed cost and decrease in variable cost, so that $\Delta F > 0$ and $\Delta c < 0$. The signs of the first two terms on the right side of Equation 10 follow from reasoning similar to the continuous case, but the third terms does not.

Term 1 – number of sellers: Following Lemma 2, a higher fixed cost / lower variable

[32] A derivation is presented in the Appendix.

[33] Here, we are presenting price as a function of the symmetric per-firm quantity and the number of firms. This is equivalent to earlier expressions of price as a function of per-firm quantity and the level of competition given the symmetry assumption.

cost source implies a fall in entry, $\Delta N < 0$. Furthermore, $\frac{\partial}{\partial N} U(\nu, q_j) - p_j q_j \geq 0$ for all $\nu \in [N_{j'}, N_j]$, as $\frac{\partial}{\partial N} U(\nu, q_j) \geq \frac{\partial}{\partial N} U(N_j, q_j)$ (there is diminishing social valuation of variety)[34] and $\frac{\partial}{\partial N} U(N_j, q_j) \geq p_j q_j$ due to the concavity of $f(\cdot)$. Thus, this term is weakly negative – there will be fewer sellers and this makes consumers worse off.

Term 2 – quantity per seller: Following Lemma 2, a higher fixed cost / lower variable cost source implies an increase in per-firm quantity, $\Delta q > 0$. Furthermore, $p(N_{j'}, \vartheta) - c_{j'} > 0$ for all $\vartheta \in [q_j, q_{j'}]$, as $\frac{\partial}{\partial q} p(N_{j'}, q) < 0$ and $p(N_{j'}, q_j) > p(N_{j'}, q_{j'}) \geq c_{j'}$ due to the concavity of $G(\cdot)$. Thus, this term is strictly positive – quantity per firm will be larger and this makes consumers better off.

Term 3 – cost minimization: Finally, the third term is weakly negative. In contrast to the continuous case, where small changes in technology around the market equilibrium had zero effect on the cost of supplying a fixed quantity due to an envelope condition on the technology choice, large changes will have an effect. By definition, the technology chosen by the market minimized total cost conditional on quantity. Therefore, by revealed preference, any change in the technology will increase total cost and reduce welfare.

The sum of the three terms – the losses from the reduction in the number of sellers and the increase in total cost given quantity, versus the gain from the increase in per-firm quantity – may be either positive or negative. Recall that the third term reflects the part of welfare that is aligned with firms' individual incentives to minimize cost. Proposition 2 says that switching to a higher fixed cost technology involves a distortion of cost minimization that may help offset the net variety and quantity distortions at the market equilibrium, or may not. If it does not, this can either be because the technology change goes "too far" – the planner would like to increase the fixed cost, but not by that much – or because it actually goes in the wrong direction – the planner would rather decrease the fixed cost. On net, therefore, consumers may either gain or lose from policy that requires direct sourcing.

As in the continuous case, there is an analog to Equation 10 in terms of the change in price. Again, Δp is related to the sum of the second and third terms of Equation 10. In contrast to the continuous case, however, price may go up or down.[35] Whereas there was no change in average cost at the equilibrium quantity under Proposition 1, under Proposition 2, average cost increases when F increases. Price rises for that reason (corresponding to the third term of Equation 10), while it decreases with the increase in quantity per seller (corresponding to the second term of Equation 10). The total change in price moves with the net of the two terms. Moving to a higher fixed cost sourcing strategy increases prices

[34] The social gain from variety is equal to $G' \cdot (f - qf')$; thus, the derivative of the social gain from variety is $G'' \cdot (f - qf') \leq 0$.

[35] This implies that even when consumers place no value on seller variety, the total welfare effect of a higher fixed cost sourcing strategy is ambiguous.

if an increase in markups arising from the decrease in local competition outweighs the reduction in marginal cost arising from the switch to a source market with lower upstream markups and variable trade costs. Thus, there is a welfare trade-off between cost and competition associated with more direct sourcing.

3.4 Discussion of welfare considerations

We have just shown that cutting out upstream middlemen can either help or hurt consumers at the end of a distribution chain, due to a fundamental trade-off between costs and entry of intermediaries. In this section, we draw out a few additional considerations that are useful for understanding how this fundamental trade-off relates to real policies and empirical contexts.

3.4.1 Product variety

In Section 3.3.2, we showed how a policy-induced direct sourcing equilibrium may be better or worse for consumers than indirect sourcing under a market equilibrium. It is worth noting, however, than a direct sourcing equilibrium may not exist. This is the case when the potential profits for a monopolist are insufficient to cover the fixed costs of entry and origin sourcing. If indirect sourcing is prohibited and a monopolist cannot earn positive profits under direct sourcing[36], then the destination market will not be served at all and consumer surplus goes to zero.

The fact that the destination market may not be served at all in our single good set-up means that chain structure has implications for product variety as well as seller variety. Several recent papers find that smaller, more remote markets in developing countries have less product variety (Atkin and Donaldson (2015); Gunning, Krishnan and Mengistu (2018)). This empirical pattern is predicted by any trade model that has product-level scale economies. Our framework implies that the existence of intermediation chains can ameliorate this tendency. Goods for which local demand is not large enough to support procurement from the production location may still be available when they can be bought in resale markets that feature lower fixed costs of sourcing. Therefore, policies that cut out middlemen can harm remote consumers by eliminating this alternative channel for accessing products.[37]

[36]Even if we ignore the integer constraint on the number of firms, this point holds if entry is zero.

[37]Note that this is not a distinct channel from those discussed in the welfare decomposition in Section 3.3. Rather, it is a description of the case where entry goes all the way to zero.

3.4.2 Technology improvements

The welfare analysis of Section 3.3 considered policy interventions that mandate movements along a technology frontier. However, the "theory of the second-best" considerations driving those results also imply that changes to the shape of the technology frontier – arising from innovations in technology, or from policies that effectively tax or subsidize some trade costs – also have ambiguous effects on welfare. This is true even when the changes are ones that strictly reduce the costs of some sourcing options, shifting the technology frontier inward.

Consider a market that is in an indirect sourcing equilibrium, and then experiences an exogenous reduction in the fixed cost of direct sourcing. This cost reduction could arise from removal of bureaucratic barriers, like reducing the red tape involved in getting a visa to visit the source market. Or, it could arise from actual technology improvements, like the introduction of a platform that facilitates communication between destination traders and origin sellers. Either way, suppose that this cost reduction is large enough that the market equilibrium switches from indirect to direct sourcing. Even though direct sourcing involves a lower fixed cost than it would have before the change, it may still involve a higher fixed cost than indirect sourcing. If this happens, the new equilibrium may feature lower entry and lower welfare. (Of course, it may also increase welfare.) Importantly, these potentially perverse effects of technology improvements can only arise through changes on the extensive margin of source market choice – if the costs of accessing a market decreases and traders continue to use that source, those inframarginal gains can only improve welfare.

3.4.3 Mechanisms driving gain and loss

Conventional wisdom often assumes implicitly that shortening distribution and collection chains will help consumers and small producers in developing countries. Our analysis suggests the need for more caution. Even under what might appear to be the circumstances least favorable to the need for intermediaries – when the destination market is always served, when consumers do not care about seller variety, and when middlemen have market power and do not transform the good itself – cutting out middlemen can reduce welfare by decreasing competition in the destination market.

Of course, consumers will not *necessarily* be hurt by such a policy, and can in fact be helped. It may be equally surprising to some readers that the sourcing structure selected by the market equilibrium is not generally efficient, and that this offers a potential rationale for intervention. However, even when there are gains from cutting out middlemen, the mechanism is not the one highlighted by the conventional wisdom, which often focuses on reducing the market power of intermediaries. Instead, in this framework, there are net gains to consumers when variable cost savings outweigh *increases* in market power.[38]

Saying more about the conditions under which policy is likely to help versus hurt requires imposing more structure on demand and competition. We have shown that this is fundamentally an empirical question, and so in the next section we turn to building these theoretical insights into a quantitative framework that can be taken to data.

4 Quantitative Model

The core insights of the framework in Section 3 are quite general: intermediation chains will arise in response to internal economies of scale in trade costs, and shorter chains can either help or harm consumers in destination markets. We now turn to quantifying these forces through a more narrowly specified model in a realistic geography. Doing so requires three steps, relative to the general framework. First, we provide a specific functional form for demand. Second, we model wholesale markets explicitly, endogenizing prices along the whole chain. Third, we assume traders are small relative to markets, but allow them to face heterogenous costs and match values with buyers and sellers. This imbeds the broad insights of the previous section into a quantifiable framework, which captures the empirical patterns in firm size and pricing described in Section 2. As we demonstrate in Section 5, this framework is extremely computationally tractable and can be taken to data with a large number of locations.

4.1 Environment

4.1.1 Geography and endowments

The economy consists of a set of locations $i \in \{1, \ldots, J\}$, each of which has a measure D_i of identical consumers. Consumers are endowed with Y_i units per capita of a numeraire good, which is freely and costlessly traded.

There is also an "intermediated" good, produced in an origin location, o, by perfectly competitive firms using a constant returns to scale production technology. Although it is only produced in the origin, the intermediated good can be re-sold by traders in other locations. Buying the good in a market j and selling it in i incurs both fixed ($F_{ij} > 0$) and multiplicative variable ($\tau_{ij} > 1$) trade costs. Trade costs are denominated in units of

[38]Direct sourcing may reduce variable costs in the destination in part through elimination of markups charged by intermediaries in upstream resale markets, as we show in Section 4. A more complete argument, therefore, is that shortening chains can reduce the role of market power of held by intermediaries who are cut out of the chain, while increasing that of those who remain, and that the latter effect may be larger than the former. However, there can be gains to consumers in our framework even if there are no upstream markups.

the numeraire and conform to triangle inequalities $F_{ij} \leq F_{ik} + F_{kj}$ and $\tau_{ij} \leq \tau_{ik}\tau_{kj}$.

A distribution chain is a sequence of links (i.e. transactions at particular locations) through which the intermediated good moves from the origin to final consumers in a given location. A chain is denoted z and is defined by a vector of length L_z whose elements are the ordered locations at which the good is bought and sold. The location $z(1)$ is always where the good is consumed, $z(L_z)$ is always the origin, and $z(l)$ is the location of the transaction that is l steps upstream from final consumers.

4.1.2 Preferences

Utility for a representative consumer[39] is given by:

$$U = q_0 + AV^{\alpha} \tag{11}$$

where q_0 is consumption of the numeraire and V is the payout from consumption of the intermediated good. The parameter $A > 0$ controls the relative weight of the intermediated good and the outside sector, while α captures the elasticity of expenditure on the intermediated good with respect to the price index (defined below).[40] We assume that consumers in all locations have a sufficiently large endowment that they consume a positive amount of the numeraire.

Consumers in location i purchase the intermediated good from from a measure of local retailers, N_{ri}, indexed by ω. The payout from the intermediated good is a constant elasticity of substitution (CES) aggregator modified as in Benassy (1996):

$$V = N_{ri}^{\gamma - \frac{1}{1-\sigma_r}} \left(\int\limits_{\omega \in N_{ri}} q(\omega)^{\frac{\sigma_r - 1}{\sigma_r}} \, d\omega \right)^{\frac{\sigma_r}{\sigma_r - 1}} \tag{12}$$

where $q(\omega)$ is the quantity purchased from retailer ω and σ_r is the elasticity of substitution across retailers.[41] The Benassy (1996) system breaks the mechanical relationship between taste for variety and market power that is built in to a standard CES aggregator. It allows for the possibility that private incentives may over- or under-provide variety, governed by the γ parameter. As in Section 3, variety describes imperfect substitutability of sellers, not products. There is a consumer price index for the intermediated good:

$$\tilde{P}_{ri} \equiv N_{ri}^{-\frac{1}{1-\sigma_r} - \gamma} P_{ri}$$

[39] These preferences can alternatively be microfounded as arising from a discrete choice across retailers with whom individual consumers have idiosyncratic match values, as described in the Appendix.

[40] We assume $\alpha \in (0, 1)$ to ensure that firm profits will decline in the price index.

[41] We assume $\sigma_r > 1$ so that retailers are substitutes.

where $P_{ri} = \left(\int_{\omega \in N_{ri}} p(\omega)^{1-\sigma_r} d\omega\right)^{\frac{1}{1-\sigma_r}}$ is the conventional CES price index and $p(\omega)$ is the price charged by retailer ω. In the case where $\gamma = \frac{1}{\sigma_r - 1}$, these preferences collapse to standard CES. If $\gamma > \frac{1}{\sigma_r - 1}$, then the social valuation of seller variety will be higher than the private one, and the opposite if $\gamma < \frac{1}{\sigma_r - 1}$. The size of γ will therefore govern whether there is insufficient or excess entry under the market equilibrium.[42]

4.1.3 Intermediation

There is a large set of potential traders that can enter into either wholesaling (w) or retailing (r) in every location. Firms can enter and sell only in a single location,[43] and sell to other traders if wholesaling or to consumers if retailing. Traders pay a sunk cost f_i to enter as type $u \in \{w, r\}$.

After entering, traders make sequential choices about where to source, which chain to serve, which supplier to buy from, and what price to charge. Traders are heterogenous, facing idiosyncratic fixed costs of sourcing from each location, as well as individual match values with each possible chain and supplier they could choose within a location. Thus, the overall payout for a trader in location i, of type u, sourcing from location j, and sourcing from wholesaler ω to serve step l on chain z will have the form:[44]

$$\Pi_{uijzl\omega} = \zeta(z,l)\,\varepsilon_{zl}(\omega)\,\pi_u(p, c(\omega), z, l) - F_{ij} - \xi(j) - f_i$$

where $\pi_u(p, c(\omega), z, l)$ denotes variable profits given own price p, unit cost $c(\omega)$, and choice to serve step l on chain z.

Realizations of the idiosyncratic component of fixed cost, $\xi(j)$, are observed after entering, and are drawn independently from a Gumbel distribution with scale parameter s. The individual step-chain and supplier match values, $\zeta(z,l)$ and $\varepsilon_{zl}(\omega)$, are observed only after choosing a source location and after choosing a step-chain pair, respectively, and are drawn independently from Frechet distributions with shape parameters $\frac{1}{\beta}$ and $\frac{1}{\mu_w}$. We normalize the Frechet scale parameters (as described in the Appendix) so that the expected

[42] We assume that $\gamma \in [0, \frac{1-\alpha}{\alpha})$, where the lower bound implies consumers (weakly) gain from variety, and the upper bound delivers uniqueness of the equilibrium in Proposition 3.

[43] This rules out traders serving consumers in multiple locations and making joint sourcing decisions across those locations. Because there is double marginalization along the chain, there is a motive to integrate, albeit one that is slightly different from the standard vertical integration problem in a production setting. Abstracting from this is reasonable simplification in the context of Nigerian consumer goods, because as discussed in Section 2, we observe very few firms that enter in multiple markets. However, endogenizing the extent of cross-location integration in this framework may be important for understanding differences in distribution structure across countries with different income levels, and is an interesting subject for future work.

[44] We subscript by both step and chain as the same chain may have multiple steps on the same origin-destination pair, but zl is a single choice – i.e. firms do not choose their step separately from the chain.

value of the maximized shock from each source location is one. This implies that traders have idiosyncratic match values with individual chains and suppliers, but the expected value of sourcing from a location is not increasing in the number of sellers there.[45]

4.2 Profit maximization by traders

We now describe traders' sequential optimization decisions over entry, source locations, chains, suppliers, and pricing.

4.2.1 Entry and sourcing choices

The measure of wholesale and retail traders for every location i is pinned down by a zero profit condition:

$$f_i = s \ln \left(\sum_j \exp \left(\frac{\pi_{uij} - F_{ij}}{s} \right) \right) \tag{13}$$

where π_{uij} are the expected variable profits of trader type u when serving route ij.

After entry, traders observe the realizations of their fixed cost shocks, and choose market to source from. The share of type-u traders in i choosing to source from j is:

$$\chi_{uij} = \frac{\exp \left(\frac{\pi_{uij} - F_{ij}}{s} \right)}{\sum_{j'} \exp \left(\frac{\pi_{uij'} - F_{ij'}}{s} \right)} \tag{14}$$

The χ_{uij} are standard logit shares because the fixed sourcing cost shocks, $\xi(j)$, follow a Gumbel distribution. The combination of small, heterogeneous traders and the lack of choke prices in the demand system ensures that there will be non-zero trade flows along every possible chain in equilibrium. This allows the equilibrium to be characterized using first-order conditions, and avoids an intractable combinatorial optimization problem.[46]

Once in j, each trader observes realizations of the chain match values $\zeta(z)$ and chooses

[45] In the absence of the normalization, an increase in the number of wholesalers in a market increases the expected match value for a trader sourcing there. This generates an agglomeration force – larger markets attract more wholesale sourcing, which further increases their size as more downstream demand flows through them and induces entry. We shut this down in order to focus attention on the consumer welfare tradeoffs, which arise regardless of agglomeration in wholesaling, and to ensure a unique sourcing equilibrium. However, agglomeration in wholesale markets may be empirically relevant in some settings, and can be allowed for within this framework at the cost of some additional complication.

[46] While this is familiar roadbloack in the international trade literature, the standard solutions are not compatible with modeling resale in response to economies of scale. For instance, assuming there are only constant variable trade costs would generate non-zero flows on all routes but is directly inconsistent with an economies of scale mechanism, while relying on a productivity distribution to achieve the same result would imply that some intermediaries must resell below marginal cost.

which chain z to serve.[47] The share of traders serving route j to i who choose link l on chain z is

$$\psi_{uijzl} = \frac{\pi_{uijzl}^{\frac{1}{\beta}}}{\sum_{z'} \pi_{uijz'l'}^{\frac{1}{\beta}}} \tag{15}$$

where π_{uijzl} are the expected variable profits from choosing step l on chain z and $\frac{1}{\beta}$ describes the dispersion of the chain-specific match values. Choosing one chain to serve means that pricing is separable across chains, which ensures that there is a pure strategy pricing equilibrium.[48] This assumption is stylized, but generates realistic aggregate behavior while yielding standard markup rules and well-behaved demands.[49]

4.2.2 Demand and pricing

Retail demand in our framework is simple and takes a modified CES monopolistic competition form. Profit maximizing retail prices follow a constant markup $m_r \equiv \frac{\sigma_r}{\sigma_r - 1}$ over marginal cost. Marginal cost, of course, will depend on the prices set by wholesalers in each location.

Wholesale pricing is more complicated, because wholesalers' buyers are themselves resellers. Their elasticity of demand will be governed not only by the distribution of their match values with sellers, but also the demand elasticity they in turn face as they pass changes in their purchase price through to their sale price. The problem at any stage of the chain, therefore, depends on the full sequence of decisions by downstream agents. We show that when buyer-seller match values are distributed Frechet there are closed-form wholesale demand and pricing rules that account for this downstream dependence at all points on the chain.

Result 1: If retail demand is CES and the assumptions on traders' optimization problem described in Section 4.1.3 hold, then wholesale demand at every step on the chain also follows a CES form and the price elasticity of demand faced by a trader at $z(l)$ is:

$$\sigma_l = (\sigma_r - 1)\left(\frac{\mu_w + 1}{\mu_w}\right)^{l-1} + 1$$

[47] This choice is from within the set of feasible chains: either wholesale or retail (depending on the trader's type) with steps from j to i. Technically, this is actually a choice of step and chain in the case of chains with multiple wholesale links with the same source and destination, e.g. a chain like (o, i, i, i).

[48] If this were not the case, sellers would face demand from multiple types of traders who arrive to make wholesale purchases in a given location, who may represent different types of downstream demand. When a given seller changes her price, she would expect the composition of traders who choose to buy from to her to shift, and there will not generally be an equilibrium in pure strategies.

[49] An assumption that traders serve all chains that go from their chosen source j to their home location and can perfectly price discriminate across chains yields similar aggregate behavior.

where $\frac{1}{\mu_w}$ is the shape parameter of the buyer-seller match value distribution.[50]

To see why this is the case, consider the following argument by induction. The price charged by a trader ω is $p_{uijzl}(\omega)$, and for simplicity, we suppress $uijz$ subscripts since all transactions are by definition within a specific chain. Suppose that the agent at step l is a trader who faces CES demand of the form

$$q_l = b_l p_l^{-\sigma_l}$$

That trader's marginal cost will be $c_l(\omega) = \tau_{l+1} p_{l+1}(\omega)$ where $p_{l+1}(\omega)$ is the price charged by supplier ω one step upstream. Profits when buying from supplier ω can be expressed as

$$\pi_l(\omega) = \left(\frac{\varepsilon(\omega) b_l}{\sigma_l}\right)\left(\frac{\sigma_l}{\sigma_l - 1}\tau_{l+1} p_{l+1}(\omega)\right)^{1-\sigma_l}$$

The trader will choose the upstream seller to maximize these profits, and since $\varepsilon(\omega)$ is distributed Frechet with shape parameter $\frac{1}{\mu_w}$, the probability of choosing seller ω from the set of potential sellers, N_{l+1}, is

$$\Pr(\omega) = \frac{p_{l+1}(\omega)^{\frac{1-\sigma_l}{\mu_w}}}{\int_{N_{l+1}} p_{l+1}(\omega')^{\frac{1-\sigma_l}{\mu_w}} d\omega'}$$

This implies that the sales of upstream sellers at $z(l+1)$ can be expressed as

$$q_{l+1} = b_{l+1}\left(p_{l+1}(\omega)\right)^{\frac{1-\sigma_l}{\mu_w} - \sigma_l}$$

where

$$b_{l+1} \equiv b_l \frac{\left(\frac{\sigma_l}{\sigma_l - 1}\tau_{l+1}\right)^{-\sigma_n}}{\int_{N_{l+1}} p_{l+1}(\omega')^{\frac{1-\sigma_l}{\mu_w}} d\omega'}$$

Note that if $\sigma_{l+1} \equiv \frac{\sigma_l - 1}{\mu_w} + \sigma_l$ and demand at step n has the hypothesized form $q_l = b_l p_l^{-\sigma_l}$, then we have shown that demand at step $l+1$ also has this form. Since consumer demand at the final step of the chain is CES, by induction, demand will follow this form at every stage of the chain.

This formulation implies that elasticities of demand are increasing – and therefore, markups are decreasing – as you move upstream along the chain away from final consumers. This captures the new stylized facts we laid out in Section 2 about how pricing and markups vary along the chain for consumer goods in Lagos. While neat closed-form

[50] More generally, if μ_w varies across locations, the elasticity will follow the recursive form $\sigma_l = (\sigma_{l-1} - 1)\left(\frac{\mu_{wl}+1}{\mu_{wl}}\right) + 1$.

aggregation relies on the assumption that buyer-seller match values are distributed Frechet, the prediction that elasticities of demand increase as you move upstream is more general and does not depend on the distribution. The assumption that the parameters governing wholesale and retail elasticities are the same across locations and steps in the chain is convenient for empirical tractability, but is also not essential.[51]

4.3 Equilibrium

Having described the setting and optimization decisions of traders, we now characterize the sourcing equilibrium that arises to serve consumers in all destinations.

4.3.1 Retail prices and chain length

The ultimate goal of the quantitative model is to characterize the structure of the distribution chains serving consumers in different locations, and the welfare implications of that structure. Both of these are related to how consumer prices depend on the delivery chain. We obtain substantial tractability from the fact that, as we show in this section, the price on a given chain is pinned down by parameters and all possible chains have positive flows in equilibrium. The retail price of the good at the end of a specific chain is simply the product of the price in the origin, the variable trade costs between each link on that chain, and the constant multiplicative markups at each link.

Recall that there is a consumer price index for the intermediated good:

$$\tilde{P}_{ri} \equiv N_{ri}^{-\frac{1}{1-\sigma_r}-\gamma} P_{ri}$$

where P_{ri} is the conventional CES price index. This can be expressed as a function of the entry and sourcing decisions of local retailers and the price index of goods brought back from a given source:

$$P_{ri}^{1-\sigma_r} = N_{ri} \sum_j \chi_{rij} P_{ij}^{1-\sigma_r}$$

where χ_{rij} denotes the share of retail traders in i sourcing from j (so that $N_{ri}\chi_{rij}$ is the measure of retail traders sourcing from j) and P_{ij} is a price index of chains brought back to i from source j.

A measure of retailers with length 1 sourcing from j and selling to consumers in i would offer a price index:

[51]In fact, the pattern will also hold in a non-CES system, as long as passthrough isn't too incomplete. However, the prediction is across steps *within* a chain, and we should not necessarily expect it to hold in cross-sectional comparisons of more up- or downstream steps across different chains if there are differences in μ_w or σ_r.

$$P_{ij} = m_r \tau_{ij} \mathcal{P}_j$$

where m_r is the retail markup, τ_{ij} is the variable trade cost, and $\tilde{P}_j$ is a wholesale price index in market j. Given the setting we have defined, all of these objects depend on parameters only, and not on any equilibrium outcomes. Note that the only i-specific component of the price is the variable trade cost incurred by retailers at the last step of the chain.

The wholesale price index in the source location is defined by:

$$\mathcal{P}_{j\neq o}^{1-\sigma_r} \equiv \frac{1}{\phi_j^{1-\sigma_r}} \left[\left(\frac{M_2}{m_r} \tau_{jo} p_o \right)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + \sum_{j'} \left(\frac{M_3}{m_r} \tau_{jj'} \tau_{j'o} p_o \right)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + \sum_{j''} \sum_{j'} \left(\frac{M_3}{m_r} \tau_{oj'} \tau_{j'j''} \tau_{j''o} p_o \right)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + \dots \right] \tag{16}$$

$$\mathcal{P}_o^{1-\sigma_r} \equiv \frac{1}{\phi_o^{1-\sigma_r}} \left[p_o^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + \left(\frac{M_2}{m_r} \tau_{oo} p_o \right)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + \sum_{j'} \left(\frac{M_3}{m_r} \tau_{oj'} \tau_{j'o} p_o \right)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + \dots \right]$$

and where M_L denotes the total markup charged over all links at the end of a chain length L. The expressions weight the price from each chain based its share in consumer expenditure and the propensity of retailers to serve it. The propensity of retailers to serve a particular chain depends on an index of profits across all chains from location j, ϕ_j, which is defined by:

$$\phi_{j\neq o}^{1-\sigma_r} \equiv \left(\frac{M_2}{m_c} \tau_{jo} p_o \right)^{\frac{1}{\beta}(1-\sigma_r)} + \sum_{j'} \left(\frac{M_3}{m_r} \tau_{jj'} \tau_{j'o} p_o \right)^{\frac{1}{\beta}(1-\sigma_r)} + \sum_{j''} \sum_{j'} \left(\frac{M_4}{m_r} \tau_{jj'} \tau_{j'j''} \tau_{j''o} p_o \right)^{\frac{1}{\beta}(1-\sigma_r)} + \dots \tag{17}$$

$$\phi_o^{1-\sigma_r} \equiv p_o^{\frac{1}{\beta}(1-\sigma_r)} + \left(\frac{M_2}{m_r} \tau_{oo} p_o \right)^{\frac{1}{\beta}(1-\sigma_r)} + \sum_{j'} \left(\frac{M_3}{m_r} \tau_{oj'} \tau_{j''o} p_o \right)^{\frac{1}{\beta}(1-\sigma_r)} + \dots$$

Importantly, both ϕ_j and $\mathcal{P}_j$ are independent of i – despite the chain-specific pricing, each source market has a single wholesale price index that it presents to all downstream locations. The cost components of Equations 16 and 17 are the price in the origin, p_o, and upstream variable trade costs τ that were incurred to get the good to location j. These do not depend on the downstream destination. Markups do depend on the downstream chain, but under the assumption that the retail and wholesale elasticities of substitution are shared across locations, they depend only on the downstream length of the chain rather than the specific destination.

Because all chains of all lengths are active between all ij pairs, the set of downstream chain lengths and therefore markups are invariant to i. Taken together, this implies that prices are a direct function of parameters, and do not depend on the full system of equations defining equilibrium. This gives the model substantial empirical tractability, making it easy to compute counterfactuals and to consider a large set of locations in the trading network.

The average chain length serving consumers in a particular location influences the retail price index, but is also of independent interest as an empirical object describing the sourcing equilibrium. The formula for this outcome, which we denote by λ_i for average chain length of chains ending in location i is

$$\lambda_i = \sum_j N_{ri}\chi_{rij}\left(\frac{P_{ij}}{P_{ri}}\right)^{1-\sigma_r}\Lambda_{ij}$$

where Λ_{ij} is the weighted average chain length on routes arriving from j.

In turn, we show in the Appendix that

$$\Lambda_{ij} = \frac{\Lambda_j}{\left(\phi_j \mathcal{P}_j\right)^{1-\sigma_r}}$$

where, depending on whether or not j is the origin,

$$\Lambda_{j\neq o} \equiv 2\left(\frac{M_2}{m_c}\tau_{jo}p_o\right)^{\frac{1}{\beta}(1-\sigma_r)} + 3\sum_{j'}\left(\frac{M_3}{m_r}\tau_{jj'}\tau_{j'o}p_o\right)^{\frac{1}{\beta}(1-\sigma_r)} + 4\sum_{j''}\sum_{j'}\left(\frac{M_4}{m_r}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o\right)^{\frac{1}{\beta}(1-\sigma_r)} + \ldots$$

$$\Lambda_o \equiv (p_o)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + 2\left(\frac{M_2}{m_r}\tau_{oo}p_o\right)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + 3\sum_{j'}\left(\frac{M_3}{m_r}\tau_{oj'}\tau_{j'o}p_o\right)^{(1-\sigma_r)\left(1+\frac{1}{\beta}\right)} + \ldots$$

Note that adding an additional link onto an existing chain necessarily increases the retail price at the end. This is because we assume that variable trade costs are weakly positive and conform to a triangle inequality, and because markups are constant under a CES structure at both wholesale and retail links. This is in contrast to the more general framework of Section 3, under which prices could actually be lowered by shortening a chain, due to variable markups. However, price is not necessarily lower at the end of a short chain compared to any longer chain – if the short chain has fewer links but goes from China to Lagos by way of Antarctica, it may have a higher price than a chain with more links that follows a route with lower trade costs.

4.3.2 Equilibrium definition and uniqueness

In equilibrium, consumers maximize their utility, traders maximize their profits through pricing and sourcing decisions, and traders enter up to zero expected profits in every location.

An equilibrium is defined by prices, sourcing shares, and measures of wholesale and retail traders such that:

1. Consumers choose a retailer and quantity to maximize their utility given in 11.

2. Traders enter up to zero expected profits, as given in 13.

3. Traders choose source locations to maximize expected profits, as given in 14.

4. Traders choose to participate in the chain that maximizes expected profits given their source location, as described by 15.

5. Traders pick the seller on that chain that maximizes their profits given their idiosyncratic draws across all relevant sellers.

6. Traders set prices that maximize their expected profits given their choice of source location, chain, and seller.

Proposition 3. If the assumptions on the setting and consumer and trader payouts of Section 4.1 hold, then the equilibrium is unique.

Proof. See Appendix.

The proof establishes that for any given number of retailers, the sourcing shares will be unique. And, the distribution of consumer prices generated by a measure of retailers which source from a given location is only a function of model parameters, so that for a given number of retailers it is possible to solve for the consumer price index and consumer expenditure. The final piece is to show that (under the parameter restrictions we have made), per-retailer profits will be strictly decreasing in the number of retailers – and thus, the measure of retailers which satisfies the zero-profit condition is unique. This implies a unique measure of retailers, sourcing decisions, and consumer price index and consumer expenditures in every location.

Our normalizations of the trader match shocks mean that the number of wholesalers do not affect the choice of routes by retailers and thus the retail problem is separable. We can nonetheless solve for the wholesaling equilibrium after first solving the retail problem. This, in turn, implies the value of goods flowing along every chain, which determines the profits for wholesalers of serving every chain (as a function of the endogenous number of wholesalers also serving that chain) and hence from sourcing from every source j for every location i. This leads to a unique wholesaling equilibrium. Full details of the wholesaling solution and a formal proof of its uniqueness are provided in the Appendix.

4.3.3 Welfare

Consumer welfare in i is a function of the variety-adjusted consumer price index, $\tilde{P}_{ri} \equiv N_{ri}^{-\frac{1}{1-\sigma_r}-\gamma} P_{ri}$. The expression for consumer surplus is:

$$CS_i = \left(\frac{1-\alpha}{\alpha}\right) (\alpha A_i)^{\frac{1}{1-\alpha}} \tilde{P}_{ri}^{-\frac{\alpha}{1-\alpha}}$$

As noted in Section 4.3.1, the price charged by traders along any particular chain is fixed. However, the measure of traders serving a given chain is endogenous. Given the total measure of traders in location i, consumers are better off when sourcing is weighted toward locations with lower P_{ij}. Traders are more inclined to source in low cost places when the fixed costs of sourcing there are low.

We show in the Appendix that the results from Section 3 still hold in this demand system.[52]The form of demand does rule out pro-competitive pricing effects, because markups do not vary in response to entry.[53] However, the overall welfare conclusions remain the same – the market equilibrium can feature either excess or insufficient entry, and the planner may wish to lengthen or shorten distribution chains. This is due to the same fundamental trade-off between costs and entry as in Section 3, although the benefits of entry operate through the variety channel alone. Increasing the sourcing share from a high fixed cost, low variable cost source location lowers the average delivered price but also reduces the total measure of retail traders via the free-entry condition. Whether this raises or lowers the true consumer price index depends on the value of γ, which captures whether firms' private decisions over- or under-value the social gain from variety.

[52]The key is the ability to summarize the decisions of other firms into a single level of competition θ. These preferences fit a utility from the intermediated good of

$$U = \frac{1}{h(N)} G\left(\sum_{v=1}^{N} h(N) f(q_v)\right)$$

and utility functions of this form also admit a level of competition. Once we do so, it is straightforward to extend our results from Section 3 into this context; a full proof is provided in the Appendix. To see that our utility fits in this form, let $h(N) = N^{\gamma\left(\frac{\alpha\sigma}{(\alpha-1)\sigma+1}\right)}$, $G(X) = AX^{\alpha\frac{\sigma}{\sigma-1}}$, and $f(q_v) = q_v^{\frac{\sigma-1}{\sigma}}$. The resulting expression can be simplified to $U = X_0 + A\left(N^{-\frac{1}{1-\sigma}-\gamma}\left(\sum_v q_v^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}\right)^{\alpha}$ which is exactly our utility (note this expression is in terms of the direct utility form from Benassy (1996) instead of the indirect utility form used elsewhere in the text).

[53]It may be possible to extend the model to capture this force under a variable elasticity of substitution demand system without finite choke prices.

5 Quantification for Nigeria

We now return to the Nigerian context to quantify the welfare implications of distribution chains. We calibrate the model from Section 4 to capture the distribution of Chinese-made apparel to locations across Nigeria, simulate the intermediation chains serving each location under those parameters, and then compare relevant simulated moments to those from our Lagos data. Our goals are, first, to show that the model is able to capture nuanced empirical patterns in the part of the trade network that we observe, at reasonable parameter values. Second, we provide a sense of the potential magnitude of implications for consumer welfare, while highlighting the key parameters driving welfare results and how they might vary across settings. Finally, we demonstrate the computational tractability of the framework, which can be applied in settings with a large number of locations and varying data inputs.

5.1 Calibration of parameters

We consider the distribution of Chinese-made apparel to consumers in Lagos and the other 36 states in Nigeria (including the Federal Capital Territory) via China and Dubai. We treat apparel as a single good (alternatively, a homogeneous bundle), and take each of the 39 locations to be a single market. We work with the level of aggregation that corresponds to the data we have, but the quantification is sufficiently computationally tractable to allow for a larger number of locations. [54] Our estimation of parameters relies mainly on the Lagos Trader Survey (LTS) data described in Section 2. We supplement this with similarly structured original survey data on apparel traders in the capital city of Oyo state, and with secondary administrative and survey data on the Nigerian economy.

5.1.1 Intermediaries' costs

Intermediaries face cost fundamentals related to both entry and sourcing the good. We estimate fixed costs of entry $f_i \in \{f_{Lagos}, f_{Other}\}$, which we constrain to be the same for wholesale and retail but allow to differ between Lagos and other locations in Nigeria. The cost of the good in all locations also depends on the manufacturer price at the origin, $p_o = \$8.70$, which we take from average cost paid to manufacturers in China in the LTS data. All intermediaries face multiplicative variable trade costs $\{\tau_{ij}\}_{\forall\ i,j}$ and fixed trade costs $\{F_{ij}\}_{\forall\ i,j}$ to buy in market j and sell in market i. For both fixed and variable trade costs, we assume that costs are symmetric from j to i and i to j, and that the triangle

[54] The level of aggregation chosen to define a market is not neutral, since it has implications for the ability of buyers arriving at a location to benefit from variety.

inequality holds with equality for importing from China or Dubai directly to locations within Nigeria. This implies, for instance, that the trade cost to move goods from China to Kano is exactly the cost to move them from China to Lagos and then Lagos to Kano.[55] The fixed trade costs are subject to an additive, intermediary-specific idiosyncratic shock from a Gumbel distribution with scale parameter s.

Variable trade costs

We take variable trade costs between China, Dubai, and Lagos directly from the average values reported in the survey data.[56][57]However, we only directly observe domestic trade costs in two locations – Lagos and Ibadan, which are the capitals of Lagos and Oyo states. Therefore, we impose the following functional form on variable trade costs within Nigeria:

$$\tau_{ij} = \phi_\tau + \upsilon_\tau d_{ij}$$

where d_{ij} is the travel time in hours by road between locations. We calculate travel times between the capital cities of each Nigerian state from Google Maps queries.[58] Using the reported values of trade costs within Lagos and from Lagos to Oyo state from the supplementary survey data, we exactly fit ϕ_τ and υ_τ given $d_{i(\text{Lagos})}$. We estimate $\phi_\tau = 1.0307$,implying a 3.1% ad-valorem trade cost for within-location sourcing, and $\upsilon_\tau = 0.0121$, which is an additional 1.2% ad-valorem trade cost per hour of road travel. Although these estimates are based on very limited data, our framework can make use of trade costs estimated using any method, and it is reassuring that our implied distance costs lie between those for Nigeria in Atkin and Donaldson (2015) and for West Africa in Teravaninthorn and Raballand (2009). Using our estimates of ϕ_τ and υ_τ and measured travel times, we calculate the τ_{ij} between all other locations in Nigeria.

Fixed trade costs

Equation 14 implies the following relationship between the fixed cost of sourcing from any location j relative to the home market and variable profits and sourcing shares:

[55]This assumption is consistent with the fact that Lagos is the main port of entry for goods from overseas which would then be transported by road within Nigeria. Air travel from anywhere in Nigeria to China or Dubai is also likely to go through the main national airport in Lagos.

[56]We do not observe China-Dubai trade, and therefore set $\tau_{(\text{China})(\text{Dubai})}$ to make the triangle inequality hold with equality from Lagos.

[57]Note that it is not necessary to use reported trade costs. Instead, what is required is any two of the following three: retail and wholesale prices, some (observed or estimated) measure of variable trade costs, and markups (or the elasticity of substitution across retail outlets).

[58]By using actual travel time rather than distance, we at least partly account for the fact that the road network is sparser and lower quality in more remote parts of the country.

$$F_{ij} - F_{ii} = \pi_{ij} - \pi_{ii} - s \ln \frac{\chi_{ij}}{\chi_{ii}}$$

In principle, this allows for estimation of fixed costs for any locations where sourcing shares and variable profits are observed. In our case, this means fixed trade costs from China and Dubai to Lagos, and from Lagos to Oyo state.

The scale parameter s governing the distribution of idiosyncratic components of fixed costs is set to \$128.72 to match the variance of the distribution of fixed costs of travel for West Africa from Startz (2021). We set F_{ii} based on the non-transport costs of sourcing reported by traders in Lagos who source from within Lagos.[59]

Because we do not observe sourcing shares for other locations, we assume that fixed costs within Nigeria follow:

$$F_{ij} = F_{ii} + \upsilon_F d_{ij}$$

where we can fit υ_F exactly based on our Lagos and Oyo estimates. We estimate $F_{ii} = \$563$ and $\upsilon_F = \$199.8$. For a trader based in Oyo and purchasing \$30,000/year in stock from Lagos, for instance, this implies a 37/63 ratio of fixed to variable trade costs. Using the estimate of υ_F, and measured travel times, we calculate estimated F_{ij} between all other locations in Nigeria.

Entry costs

Equation 13 implies that we can estimate the fixed cost of entry, inclusive of any fixed cost of sourcing from the home market via:

$$f_i + F_{ii} = \pi^{v}_{rij} - F_{ij} - s \ln \chi_{ij}$$

We observe χ_{ij} and π_{ij} for Lagos and Oyo, F_{ii} for Lagos, and estimated F_{ij} in the previous step. Therefore, we can calculate entry costs for Lagos and Oyo using this approach, and find $f_{Lagos} = \$5,029.48$ and $f_{Oyo} = \$1,652.52$. We assume that the estimate for Oyo holds for all other non-Lagos locations within Nigeria.[60]

[59] This includes both financial costs incurred by the trader, and an estimate of the average time cost of the reported one day per week spent on purchasing. However, it is not strictly necessary to identify F_{ii} separately from the entry cost f_i, and so the level of this estimate is not critical.

[60] We normalize entry costs in China and Dubai since they do not influence consumer outcomes in Nigeria.

5.1.2 Utility function parameters

The model includes parameters that govern the utility of final consumers. We allow A_i, which shifts expenditure on the intermediated good relative to the outside sector, to be a function of features of each location i. The elasticity of expenditure on the intermediated good with respect to its price index is α. Finally, γ parameterizes a difference between the private and social value of variety.

Equations 11 and 12 imply that per capita expenditure on the intermediated good (apparel) can be log-linearized to yield:

$$\ln\left(\frac{E_i}{D_i}\right) = \delta_i + \beta_1 \ln \bar{p}_{ri} + \beta_2 \ln N_{ri} \tag{18}$$

where $\bar{p}_{ri}$ is the average retail price and N_{ri} is the number of sellers in location i. The terms δ_i, β_1, and β_2 are functions of the demand parameters.[61]

We estimate a regression based on Equation 18 using secondary data for each state in Nigeria. Full details on these data and the measures we use are provided in the Appendix. We take per capita expenditure on apparel from the Living Standards Measurement Survey (LSMS) for Nigeria, and the average unit price of apparel from the Nigerian CPI.[62] Data on the number of traders in each state comes from the Small and Medium Enterprises Development Agency of Nigeria (SMEDAN).

We allow A_i, which describes the fraction of spending that goes to apparel versus other goods, to be a function of GDP per capita and the density of formal markets in each state. Data on GDP per capita comes from the Nigerian National Bureau of Statistics (NBS), and the fraction of localities that have a permanent market is from the LSMS data. Because prices and the number of sellers may be correlated with unobserved factors that also influence consumer spending, we use the travel time from Lagos (based on Google Maps queries) and the average cost of land (from the LSMS) as instruments for $\bar{p}_{ri}$ and N_{ri}.

Results are shown in Appendix Table 8. We estimate $\alpha = 0.435$ and $\gamma = 0.858$. Because γ is larger than $\frac{1}{\sigma_r - 1} = 0.645$, the social value of seller variety exceeds the private value. Relative to a lower γ, this makes interventions that shift traders toward higher fixed cost sourcing strategies less valuable from a consumer welfare perspective, because they exacerbate insuffient entry. Our welfare results are therefore driven by the responsiveness of expenditure to the number of sellers relative to prices.

[61] Per capita expenditure is $\frac{E_i}{D_i} = (\alpha A_i)^{\frac{1}{1-\alpha}} \left(\frac{N_{ri}^{\gamma}}{\bar{p}_{ri}}\right)^{\frac{\alpha}{1-\alpha}}$ if all retailers charge the same price, and is an approximation if they do not. We calculate $\alpha = \frac{\beta_1}{\beta_1 - 1}$, $\gamma = -\frac{\beta_2}{\beta_1}$, and $A_i = \frac{\beta_1 - 1}{\beta_1} \exp\left(\frac{\delta_i}{1-\beta_1}\right)$.

[62] We thank David Atkin and Dave Donaldson for sharing state-level apparel prices with us.

5.1.3 Match values and elasticities

Finally, we estimate parameters describing the willingness of buyers to substitute across sellers. The dispersion of match values across chains is governed by β_r, the dispersion of buyer-seller match values between traders is determined by μ_w, and σ_r describes the price elasticity of demand faced by each retail trader.

We set $\sigma_r = 2.55$ to match markups charged by retailers in the LTS data,

$$m_r = \frac{\sigma_r}{\sigma_r - 1}$$

where m_r is the average observed retail markup.

The cross-chain elasticity parameter, $\beta = 0.19085$, is calibrated to match the share of Lagos traders sourcing from China that we observe in the LTS data. We show in the Appendix that under a second order approximation to log markups, $m_w = m_r \exp\left(-\frac{1}{\mu_w \sigma_r}\right)$, where m_w is the wholesale markup. This implies that:

$$\mu_w = \left(\frac{1}{\sigma_r}\right) \frac{1}{\ln m_r - \ln m_w}$$

We estimate $\mu_w = 3.19$ to match average observed markups charged by wholesalers in the LTS data.

5.2 Baseline outcomes

Using the parameters estimated in Section 5.1, we quantify the model for all states in Nigeria. This yields equilibrium measures of sellers, sourcing shares, and prices, allowing us to describe the distribution structures serving consumers in each location.[63]

We begin by comparing calibrated baseline outcomes in Lagos to the raw data patterns from the Lagos Trader Survey in Table 5. In column (1) we show data patterns related to the stylized facts introduced in Section 2, restricting the sample to apparel traders sourcing in China, Dubai, or Nigeria. We show average firm size, steps upstream and downstream of the Lagos trader, shares of traders sourcing from China and Dubai, and coefficients from regressions similar to those in Section 2 relating downstreamness, size, and markups to chain position. In column (2), we show the corresponding moments from the baseline calibration for Lagos. However, since our LTS data measures of upstreamness

[63] Because in theory there are infinitely many chains serving each location, there are not closed form solutions to our expressions for the wholesale price indices $\mathcal{P}_j$. In practice, we make use of the fact that, under the approximation to wholesale markups described in Section 5.1, there are simple closed form solutions as described in the Appendix. Alternatively, we could have simply truncated the resulting expressions. It is computationally feasible to trace out even very long chains, and the contribution of long chains to the price index drops off fairly rapidly.

Table 5: Comparison between baseline simulation and data patterns in Lagos

Outcomes for Lagos		(1) LTS data	(2) Calibration	(3) Truncated calibration
Average firm revenue ($US)		28,344	27,141	27,141
Steps upstream		1.82	1.72	1.58
Steps downstream		1.52	1.31	1.29
Sourcing share	From China	0.422	0.430	0.430
	From Dubai	0.090	0.010	0.010
OLS coefficients	Steps upstream & downstream	-0.221	-0.871	-0.384
	Revenue ($US) & steps upstream	-21,650	-5,054	-27,957
	Markup & steps downstream	-0.186	-0.160	-0.190

and downstreamness are truncated at one step away from the Lagos trader, we consider the comparably truncated outcomes from the calibration in column (3) to be the right comparison for judging model fit.

Table 5 demonstrates that the implementation of the model successfully captures the main empirical patterns we laid out in Section 2. The only moment in the table that is explicitly targeted is the share sourcing from China. The model matches the scale of Lagos firms and their approximate position along the full Nigerian distribution chain, as well as nuanced patterns in the relationships between chain position and scale and pricing. Traders who are closer on the chain to the manufacturer are also further from final consumers (i.e. upstreamness and downstreamness are negatively related), traders who are closer to the manufacturer are larger, and markups are decreasing in upstreamness. The model fits both the sign and the approximate magnitude of these patterns. The only moment the model does not seem to fully capture is the tendency to source from Dubai – perhaps because, in reality, there are procompetitive or one-stop-shopping forces that make Dubai an attractive source market that are not captured in the baseline model. If anything, the model appears to slightly undersstate the length of chains observed in the data.

The calibration yields a rich set of descriptives about the structure of distribution chains serving the rest of Nigeria, outside of Lagos. The expenditure-weight average length of chains serving consumers in Nigeria is 4.5, implying that there are on average two or three intermediaries between the manufacturer and the final consumer. The largest wholesaling centers under the calibration are in Lagos, Kano, Bauchi, Kaduna, and Rivers states, which include the major commercial cities of Lagos, Kano, and Port Harcourt, as well as the main north-south corridor between Abuja and Kano, all highlighted in Figure 1 in Section 2.

Table 6 shows that equilibrium chain length is increasing in remoteness, proxied by

Table 6: Relationships between distribution structures and location features

	(1) Chain length	(2) Average price ($US)	(3) Ratio of total markup to variable trade cost	(4) Ratio of total markup to variable trade cost
Travel time from	0.023***	0.765***	0.003	-0.004**
Lagos (hours)	(0.004)	(0.119)	(0.002)	(0.002)
State population	-0.028***	-0.474		
(millions)	(0.010)	(0.293)		
Chain length				0.279***
				(0.049)
Obs	37	37	37	37

Note: The unit of observation is a state within Nigeria. Travel time from Lagos is measured in hours based on Google Maps, as described in Section 2. Chain length is the expenditure-weighted average of the number of agents on the chain serving final consumers in each state. The total markup is expenditure-weighted average of the product of the multiplicative markups incurred at each step on the chain, and the total variable trade cost is the analagous object for trade costs.

hours of travel time from Lagos, and decreasing in market size, proxied by state population. The opposite holds for the average retail price. Smaller, more remote markets within Nigeria are served by longer chains with higher prices and less seller variety. Column (3) and (4) suggest that the relationship between remoteness and chain structure is likely to complicate inferences about trade costs, prices, and competitiveness. While column (3) finds a positive relationship between the ratio of the total markup incurred between the origin and the final destination market to the total variable trade cost, the sign of this relationship flips when accounting for distribution structure in column (4).

5.3 Counterfactuals

5.3.1 Direct sourcing from the origin

We next turn to comparing outcomes under our baseline calibration to outcomes under a variety of alternative scenarios. We begin with the most basic – what happens if indirect sourcing is prohibited, so that retailers in all locations must source directly from China? Although complete prohibition of wholesaling isn't a common policy, this counterfactual has the flavor of more realistic disintermediation policies that prohibit or discourage wholesaling in particular places along the chain. For instance, in Ethiopia, farmers are required to sell some products (such as coffee) directly through the national commodity exchange. India's policy on Agricultural Produce Market Committee markets, or mandis, makes it difficult for traders to buy at farmgates or in more local, unregulated marketplaces. There has been a great deal of debate recently over whether deregulation of these markets would benefit farmers or consumers, or indeed, whether it would lead to more or less intermediation (Chatterjee and Mahajan (2021)).

In practice, we arrive at a direct-sourcing-only equilibrium by setting trade costs between consumer locations and all sources other than China to high levels so that the share of firms sourcing from China approaches one hundred percent. By construction, the result is that all consumers are served by chains of length three with only one intermediary, their local retailer. The equilibrium fixed sourcing cost paid per retailer increases almost 10-fold, as most sourced domestically in the baseline. As a consequence, the number of retail outlets carrying Chinese apparel crashes throughout Nigeria, falling by 94.3% overall. Consumers reduce their spending on the intermediated good by 69.3% on average across all states, shifting toward the outside sector in response to the lack of access to retailers, in spite of a substantial reduction in the average delivered price. Consumer surplus from Chinese apparel falls to 27.6% of its baseline level.

While this is a stylized example, it reflects realistic forces in response to mandatory disintermediation policies that restrict access to low fixed cost sources, such as disallowing farm gate trade, non-licensed local wholesale markets, or resale of imported goods. Small retail shops cannot aggregate enough demand to cover the fixed costs of direct sourcing, and so there is exit and the surviving firms are larger. In our setting, the loss of retail variety and corresponding reduction in expenditure represents the idea that consumers are not willing to travel to the couple of larger stores in the state capital that do carry the imported good.

5.3.2 International e-commerce innovation

Even in circles in which heavy-handed regulation is not popular, the idea that falling trade costs might connect producers and consumers more directly is often viewed favorably. One approach that has gained particular traction is the development of platforms to improve buyer-seller matching by reducing search costs or information asymmetries. These take a variety of forms, ranging from business-to-business e-commerce platforms—for instance, the 2018 expansion of Taobao in China, studied in Couture et al. (2021), or ConnectAmericas, studied in Peru in Carballo et al. (2022)—to public exchanges, such as the Ethiopian commodity exchange established in 2008 and since emulated in other countries, such as Ghana (Nyarko and Pellegrina (2021)). The last five years has seen a wave of public and private investment in platforms, both to link farmers with buyers in agricultural markets, in places like India, Ghana, and Uganda, or to link small retailers with larger distributors.

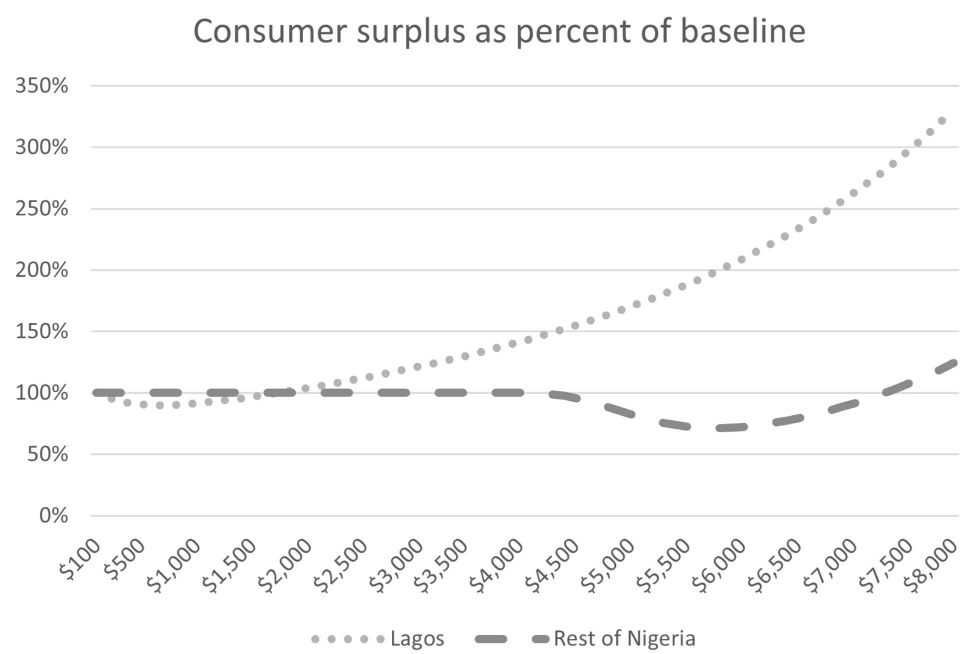
It is implicitly assumed that such technologies will benefit small producers and consumers, and in a framework with direct trade and efficient levels of firm entry, that would indeed be the case. However, in a world with intermediation chains and potentially inefficient sourcing equilibria, it is important to take endogenous restructuring of distribution into account. As noted in Section 3, even an improvement in technology that creates a pure reduction in trade costs can potentially reduce welfare in the presence of the second-best considerations related to the trade-off between variable costs and entry.

We model the introduction of such a business-to-business platform as a reduction in the fixed cost of sourcing from China, to any location in Nigeria. While we don't assume any particular mechanism, one could imagine the fixed cost reduction arises from an something like an improvement in search technology – the information available on the platform reduces the time spent searching or reduces the need to travel to the source market in person to search, thus cutting down on time and travel components of fixed sourcing costs. Rather than take a stand on how much such a technology would reduce costs, we run the counterfactual simulation at different levels of fixed cost reductions and show how welfare changes across destinations in Nigeria. Figure 2 shows the welfare impacts of increasingly large reductions in the fixed cost of sourcing from China. The baseline fixed cost of sourcing from China in Lagos is \$10,201, and we consider reductions ranging from 1% up to 80% of that cost.

Figure 2 illustrates several key aspects of the welfare implications of distribution chains. First, reductions in the fixed cost of sourcing from China have non-monotonic effects on consumer surplus. In Lagos, 43% of traders already source from China at baseline. Small reductions in the fixed cost come with large increases in the fraction who choose to source there, and reductions in the total number of retailers. At first, the exit of sellers outweighs the fall in price resulting from lower variable costs, and consumer surplus goes down. At around a 20% reduction in the baseline fixed cost, the value of the cost reductions starts to outweigh the loss of seller variety, and consumer surplus increases relative to the baseline. As fixed cost reductions get large enough, entry increases, and consumer surplus increases further.

The pattern in the rest of Nigeria, outside Lagos, is more nuanced. Reductions in the fixed cost of sourcing from China have essentially no impact at low levels because they are too small to induce traders to switch to sourcing from China and so generate no noticeable shift in sourcing shares. This is consistent with one of the patterns that has emerged so far

Figure 2: Changes in consumer surplus due to fixed trade cost reductions



Note: The x-axis shows the size of the reduction in the fixed cost of sourcing from China. The baseline fixed cost of sourcing from China in Lagos is $10,201, so a $100 reduction is approximately a 1% reduction in the Lagos-China component of the fixed cost. The y-axis shows the change in consumer surplus relative to the baseline value under each cost reduction scenario.

from some of the new agricultural matching platforms – they seem to have little impact, in part because few of the intended "direct" transactions end up taking place. If such platforms don't reduce costs by enough to make it worth a discrete switch in strategy for a substantial number of agents, then we should not expect to see much in the way of gains or losses. In our counterfactual scenario, traders outside Lagos start shifting toward direct sourcing from China around a 45% reduction in the Lagos-China cost. This initially has a net negative welfare impact, as the marginal shifts in sourcing and resulting decreases in entry outweigh the gains from lower passed through costs, just as they did in Lagos. Consumers in most of Nigeria don't start to see gains until the fixed cost reduction is almost 75% of the baseline Lagos-China cost.[64]

[64] The quantitative model rules out destination market gains from increased wholesale entry in upstream markets. In reality, intermediaries elsewhere in Nigeria might gain from wholesale entry in Lagos, through gains from seller variety or procompetitive effects on wholesale prices. To the extent that this happens, it would mitigate the losses we see from small cost reductions.

6 Conclusion

Wholesale and retail firms play a major role in the distribution of goods all over the world. Both the literature on agriculture in developing countries and our survey data on manufactured goods in Nigeria document that, contrary to the default assumption in the trade literature, long intermediation chains in which goods pass sequentially through the hands of more than one reseller are prevalent. We show that economies of scale in trade costs faced by individual firms are sufficient to give rise to such chains when goods can be resold by agents other than the original producer. While simple, this conceptual insight yields a rich microfoundation for the endogenous structure of distribution chains and generates substantive empirical predictions about what chains will look like across locations, and what firms and transactions will look like along chains.

Thinking about intermediaries as facing a menu of source markets with different fixed and variable costs is also the key to understanding how policies or technologies that shorten chains will impact consumers. We show that the market equilibrium does not generally select the efficient distribution structure, and that the second best distribution equilibrium may involve either longer or shorter chains. These second best considerations arise from a trade off between minimizing the variable costs of serving a particular set of consumers and offsetting entry distortions in the presence of market power and economies of scale. Quantifying these forces in the context of Chinese-made apparel sold in Nigeria suggests that shortening chains may indeed reduce consumer welfare in remote locations, whether it is due to regulatory intervention or technology improvements.

Price gaps between producers and consumers seem to be large on average in developing countries, and policymakers are extremely interested in reducing them. International organizations frequently fixate on reduction of "marketing costs" as a win-win solution to the "classic food price dilemma" (World Bank 2009): how to raise prices for poor producers without raising them for poor consumers. It is common for policy makers to assume that cutting out intermediaries will reduce their influence on prices and eliminate extra costs, but we demonstrate that it cannot be assumed that more direct connections are *necessarily* good for small producers or consumers. Accounting for the endogenous structure of intermediation chains is key to understanding how policy levers targeting the trading sector or trade costs are likely to affect consumers.

It may be tempting for economists to take the opposite stance, assuming that the market equilibrium delivers efficient distribution structures, and that regulation of intermediation can only cause harm and reductions in trade costs can only do good. This is also not the case – we show that, in principle, there can be a role for policy interventions that increase welfare, and that even pure technology improvements can have perverse effects.

This highlights the importance of not building efficiency into models of intermediation or agricultural trade by construction, for instance through the reliance on a common form of CES demands with monopolistic competition. It also implies that allowing for and documenting the extent of endogenous chain restructuring should be important for empirical work on intermediation, rather than simply studying changes in prices and passthrough while assuming that the chains themselves remain fixed.

We provide rich empirical evidence on a specific set of manufactured consumer goods in Nigeria, but hope that this paper highlights the need for much more systematic empirical documentation of distribution structures across settings, products, and countries. Casual empiricism suggests that consumers in rich countries may, on average, be served by shorter chains than those in poor ones. This would be consistent with the predictions of our model. New data collection may be needed, especially in developing countries, but there is also potential to use linked customs and VAT microdata as these become available across an increasingly wide range of countries.

Explaining the cross-country patterns that we suspect will arise from such data may also require additions to our modeling framework. We considered distribution of a single good, by firms that only serve a single location. These simplifications are realistic in the context we study in Nigeria, but we have abstracted from several forces that are likely to be important for understanding differences in the structure of distribution chains across rich and poor countries. The first is economies of scope that could be achieved if one intermediary can source and sell multiple goods at a total cost that is less than the cost of dealing in each separately. Scope decisions also introduce an additional "one stop shopping" role for entrepôt locations. For instance, a trader might be able to go to Dubai to buy goods from both China and India, rather than having to pay the costs to source from each separately.

A second force that will be key to understanding, for instance, the supermarket revolution in some parts of the developing world, is the motive to integrate by serving consumers in multiple locations. Integration allows intermediaries to both take advantage of greater economies of scale in sourcing, and to eliminate double marginalization at at least one step of the chain. If intermediaries in the developing world face more constraints on scale or the ability to have multiple outlets than those in the rich world – for instance due to differences in credit constraints or span of control – then allowing for scope and integration decisions becomes particularly important for explaining differences in chain structure across countries. We think this is likely to be a fruitful direction for future research, particularly if improved data also makes empirical comparisons across developing and developed countries possible.

References

Ahn, JaeBin, Amit K. Khandelwal, and Shang-Jin Wei. 2011. "The role of intermediaries in facilitating trade." *Journal of International Economics*, 84(1): 73–85.

Akerman, Anders. 2018. "A theory on the role of wholesalers in international trade based on economies of scope." *Canadian Journal of Economics*, 51(1): 156–185.

Antras, Pol, and Arnaud Costinot. 2011. "Intermediated Trade." *The Quarterly Journal of Economics*, 126(3): 1319–1374.

Antras, Pol, and Davin Chor. 2018. "On the Measurement of Upstreamness and Downstreamness in Global Value Chains." NBER NBER Working Paper 24185.

Atkin, David, and Dave Donaldson. 2015. "Who's Getting Globalized? The Size and Implications of Intra-National Trade Costs." SSRN Scholarly Paper ID 2642224.

Atkin, David, Benjamin Faber, and Marco Gonzalez-Navarro. 2018. "Retail Globalization and Household Welfare: Evidence from Mexico." *Journal of Political Economy*, 126(1): 1–73.

Bardhan, Pranab, Dilip Mookherjee, and Masatoshi Tsumagari. 2013. "Middlemen Margins and Globalization." *American Economic Journal: Microeconomics*, 5(4): 81–119.

Barrett, Christopher B., Thomas Reardon, Johan Swinnen, and David Zilberman. 2020. "Agri-food Value Chain Revolutions in Low- and Middle-Income Countries." *Journal of Economic Literature.*

Bartkus, Viva Ona, Wyatt Brooks, Joseph P. Kaboski, and Carolyn Pelnik. 2022. "Big fish in thin markets: Competing with the middlemen to increase market access in the Amazon." *Journal of Development Economics*, 155: 102757.

Benassy, Jean-Pascal. 1996. "Taste for variety and optimum production patterns in monopolistic competition." *Economics Letters*, 52(1): 41–47.

Bergquist, Lauren Falcao, and Michael Dinerstein. 2020. "Competition and Entry in Agricultural Markets: Experimental Evidence from Kenya." *American Economic Review*, 110(12): 3705–3747.

Bernard, Andrew B., J. Bradford Jensen, Stephen J. Redding, and Peter K. Schott. 2010. "Wholesalers and Retailers in US Trade." *American Economic Review*, 100(2): 408–413.

Biglaiser, Gary. 1993. "Middlemen as Experts." *The RAND Journal of Economics*, 24(2): 212–223.

Blum, Bernardo S., Sebastian Claro, and Ignatius J. Horstmann. 2009. "Intermediation and the Nature of Trade Costs: Theory and Evidence."

Carballo, Jeronimo, Marisol Rodriguez Chatruc, Catalina Salas Santa, and Christian Volpe Martincus. 2022. "Online business platforms and international trade." *Journal of International Economics*, 137: 103599.

Casaburi, Lorenzo, and Tristan Reed. 2019. "Interlinked Transactions and Competition: Experimental Evidence from Cocoa Markets."

Chatterjee, Shoumitro. 2019. "Market Power and Spatial Competition in Rural India." Faculty of Economics, University of Cambridge Cambridge Working Papers in Economics.

Chatterjee, Shoumitro, and Aprajit Mahajan. 2021. "Why Are Indian Farmers Protesting the Liberalization of Indian Agriculture?"

Cosar, A. Kerem, and Banu Demir. 2018. "Shipping inside the box: Containerization and trade." *Journal of International Economics*, 114: 331–345.

Couture, Victor, Benjamin Faber, Yizhen Gu, and Lizhi Liu. 2021. "Connecting the Countryside via E-Commerce: Evidence from China." *American Economic Review: Insights*, 3(1): 35–50.

Crozet, Matthieu, Guy Lalanne, and Sandra Poncet. 2013. "Wholesalers in international trade." *European Economic Review*, 58: 1–17.

Dhingra, Swati, and John Morrow. 2019. "Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity." *Journal of Political Economy*, 127(1): 196–232. Publisher: The University of Chicago Press.

Dhingra, Swati, and Silvana Tenreyro. 2020. "The Rise of Agribusiness and the Distributional Consequences of Policies on Intermediated Trade." Social Science Research Network SSRN Scholarly Paper ID 3535470, Rochester, NY.

Dillon, Brian, and Chelsey Dambro. 2017. "How Competitive Are Crop Markets in Sub-Saharan Africa?" *American Journal of Agricultural Economics*, 99(5): 1344–1361.

Dixit, Avinash K., and Joseph E. Stiglitz. 1977. "Monopolistic Competition and Optimum Product Diversity." *The American Economic Review*, 67(3): 297–308. Publisher: American Economic Association.

Emran, M. Shahe, Dilip Mookherjee, Forhad Shilpi, and M. Helal Uddin. 2021. "Credit Rationing and Pass-Through in Supply Chains: Theory and Evidence from Bangladesh." *American Economic Journal: Applied Economics*, 13(3): 202–236.

Euromonitor. 2010. "Euromonitor Passport Database."

Fafchamps, Marcel, and Ruth Vargas Hill. 2008. "Price Transmission and Trader Entry in Domestic Commodity Markets." *Economic Development and Cultural Change*, 56(4): 729–766.

Ganapati, Sharat. 2020. "The Modern Wholesaler: Global Sourcing, Domestic Distribution, and Scale Economies."

Gunning, Jan Willem, Pramila Krishnan, and Andualem T. Mengistu. 2018. "Fading Choice: Transport Costs and Variety in Consumer Goods." Centre for the Study of African Economies, University of Oxford 2018-05-2. Publication Title: CSAE Working Paper Series.

Hahn, F. H. 1962. "The Stability of the Cournot Oligopoly Solution." *The Review of Economic Studies*, 29(4): 329–331.

Iacovone, Leonardo, and David McKenzie. 2019. "Shortening Supply Chains: Experimental Evidence from Fruit and Vegetable Vendors in Bogota." *World Bank Policy Research Working Papers.*

Krishna, Kala, and Yelena Sheveleva. 2017. "Wheat or Strawberries? Intermediated Trade with Limited Contracting." *American Economic Journal: Microeconomics*, 9(3): 28–62.

Lagakos, David. 2016. "Explaining Cross-Country Productivity Differences in Retail Trade." *Journal of Political Economy*, 124(2): 579–620.

Leke, Acha, Reinaldo Fiorini, Richard Dobbs, Fraser Thompson, Aliyu Suleiman, and David Wright. 2014. "Nigeria's renewal: Delivering inclusive growth." McKinsey & Company Global Institute.

Mankiw, N. Gregory, and Michael D. Whinston. 1986. "Free Entry and Social Inefficiency." *The RAND Journal of Economics*, 17(1): 48–58.

Matsuyama, Kiminori, and Philip Ushchev. 2020. "When Does Procompetitive Entry Imply Excessive Entry?" Social Science Research Network SSRN Scholarly Paper ID 3650105, Rochester, NY.

Nielsen. 2015. "Africa: How To Navigate the Retail Distribution Labyrinth." Nielsen.

Niepmann, Friederike, and Tim Schmidt-Eisenlohr. 2017. "International trade, risk and the role of banks." *Journal of International Economics*, 107: 111–126.

Nyarko, Yaw, and Heitor S Pellegrina. 2021. "From Bilateral Trade to Centralized Markets: A Search Model for Commodity Exchanges in Africa."

Rubinstein, Ariel, and Asher Wolinsky. 1987. "Middlemen." *The Quarterly Journal of Economics*, 102(3): 581–593.

Seade, Jesus. 1980. "The stability of cournot revisited." *Journal of Economic Theory*, 23(1): 15–27.

Spence, Michael. 1976. "Product Selection, Fixed Costs, and Monopolistic Competition." *The Review of Economic Studies*, 43(2): 217–235.

Startz, Meredith. 2021. "The value of face-to-face: Search and contracting problems in Nigerian trade."

Teravaninthorn, Supee, and Gael Raballand. 2009. *Transport Prices and Costs in Africa: A Review of the Main International Corridors.* World Bank Publications.

A Data Appendix

A.1 Lagos Trader Survey data

This section provides a brief description of the methodology of the Lagos Trader Survey, as well as the treatment of the survey data used in the analysis. More details about the survey are provided in Startz (2021).

A.1.1 Survey methodology

The Lagos Trader Survey (LTS) aimed to collect data from a sample of traders in Nigeria selling manufactured consumers goods. The sample frame was created from a listing of all shops located in major commercial and wholesaling areas of the city of Lagos. This listing exercise covered the markets and plazas (indoor markets in multi-story buildings) served by LAWMA, the Lagos State waste collection agency, supplemented with market areas located on Nigerian federal government land that were not included in the LAWMA list. It does not include residential or manufacturing areas, or traditional markets where vendors mainly sell food and household consumables, or several markets selling primarily used goods. Between October 2014 and April 2015, a team of research assistants counted 52,830 shops throughout these areas of Lagos, enumerating them by location (building and floor numbers) and the type of products being sold.

In the first round of the LTS (LTS1) in 2015, interviews were conducted with 1,179 respondents randomly sampled from this sample frame, reflecting a response rate of 82%. A second round of the LTS survey (LTS2) was conducted in 2016, with a 100% resample of respondents who imported in the previous round, and a 30% resample of non-importers. The response rate was 85% overall, with 75% of original respondents completing the full survey, and an additional 10% (most of whom had gone out of business or relocated) completing a shorter phone survey. A third round of the LTS was completed in 2018. The data in this paper draws mainly from LTS2; all statistics and analysis are reweighted to account for sampling probabilities.

A.1.2 Survey data

LTS2 asks retrospectively about traders' activities in the calendar year 2015. In the survey, respondents were asked questions at five different levels of observation: 1) the business, 2) source countries, 3) shipments, 4) suppliers, and 5) individual purchases. Questions about the size and tenure of the business as well as the fraction of total sales that are wholesale were collected at firm level. For imports, questions about trade costs were asked at the level of a specific shipment, from a particular country in a particular month. Questions about the cost and sale price of a particular product were asked at the transaction level, from a particular supplier within a shipment. For domestic purchases, it was not possible to recover this level of detail due to the much higher frequency of transactions. Instead, respondents are asked about a typical purchase from their two biggest suppliers, and less information is available about trade costs. For this reason, most of the analysis in Section 2 of the paper focuses on importers. Throughout the paper, analysis is done at the level of observation at which the outcome variable is observed. For instance, when the outcome is annual revenue, other variables are aggregated up to the firm level. When the outcome is the price of a specific product, other variables are distributed down to the transaction level.

Trade costs

Data on trade costs (for imports) were collected with reference to a specific shipment from a particular country. We collected information on the following categories of costs: visas, airfare, other travel costs, transportation, clearing the port, and payments to agents (e.g. to inspect the product). Importantly, respondents are asked to report these costs paid for the entire shipment, including all products purchased from all suppliers. For instance, the question about transportation asks, "How much did you pay in total for transporting/shipping ALL the products you bought on this trip?". Transaction level details (e.g. about the cost paid to a supplier) were only asked for the top two products purchased from each supplier in that shipment, and therefore may not cover the full value of items included in the shipment. Instead, we also asked about the total value of all goods purchased, e.g. "What was the total cost of ALL the goods you bought on this [date] trip to [country]? Please include both the products you just told me about and any others you purchased." Response rates on this question were much lower, and so we are only able to calculate ad valorem equivalent transport or total trade costs for a subset of shipments.

Prices and markups

Data on import purchases were collected at the level of individual transactions, for a specific product, from a specific supplier, in a specific country, on a particular date. This data includes the type of product, the quantity purchased, the cost paid to the supplier, and the average price the trader charged to customers who bought that product. In cases where a trader bought more than two products from a supplier in a particular shipment, the survey asked for details about the top product. The data may therefore not be representative of lower value purchases in cases where the trader buys many different things from a single supplier at a time.

Markups are calculated in two different ways for the purposes of the analysis in the paper. Our baseline measure of markups is simply sale price divided by purchase price, minus one. For instance, if a trader bought a pair of shoes for $5 and sold them on average for $7.50, we calculate the markup to be 50%. There are a limited number of cases in which purchases and sales are reported in different units (e.g. "bales" versus "sacks") and we are not able to convert between these units and so cannot calculate the markup. Our second measure of markups includes the ad valorem equivalent of transportation costs in the denominator, i.e. sale price divided by unit cost, where unit cost is purchase price plus unit transportant costs. Because we observe transportation costs at the shipment rather than the transaction level, we are only able to calculate this measure for shipments where the total cost of all products purchased in the shipment is reported. The sample for any analysis using this measure is therefore much smaller, and relies on the assumption that all transportation costs are variable costs.

There is substantial variation in the raw ratio of revenue to stock cost. In all shipment and transaction-level analysis, we trim values below the 10th and above the 90th percentile of this ratio distribution. In effect, this leaves transactions for which this ratio is over 330% or below 5% out of the analysis. Given the structure of the survey, we think that these very large and very small values are extremely likely to reflect entry errors – for instance, extra zeros in monetary values, or errors in the quantity unit recorded. However, results with trimming at the top and bottom 5% or 2% are qualitatively similar, if noisier, and are available on request.

A.2 Other data sources

The analysis in Section 5 of the paper relies on several additional data sources, beyond the Lagos Trader Survey. In this section, we describe those data sources and the way they are used.

Original survey data from Oyo state

For some of the analysis in Section 5, we use survey data that we collected from apparel traders in Ibadan, the capital of Oyo state, located approximately 130 kilometers from Lagos. This survey was conducted in 2019, and covered a representative sample of traders in used clothing markets. The survey was structured similarly to the LTS survey described above, and asked retrospectively about transactions over the preceding 6 months. We use information on the retail fraction of sales, total annual revenue, the location of suppliers, and source-specific variable trade costs. These data enable the estimation of domestic variable and fixed trade costs within Nigeria, as well as firm entry costs outside Lagos, as described in Section 5.1.1.

Living Standards Measurement Survey (LSMS) data

We use data from Wave 2 of the General Household Survey-Panel (GHS-P) covering 2012/2013. The GHS-P is a representative household panel survey covering all of Nigeria and conducted by the Nigerian National Bureau of Statistics (NBS) and the World Bank as part of the Living Standards Measurement Survey (LSMS) program. From the post-planting and post-harvest household questionnaires, we add up expenditures on apparel items to obtain total annual apparel expenditure, which we divide by household size to calculate expenditure per capita. This is the outcome variable in the regression used to estimate utility function parameters in Section 5 of the paper.

When estimate utility function parameters, we allow A_i to be a function of GDP per capita and the density of markets. We take the empirical measure of market density from the GHS-P community questionnaire. Each community in the data is asked whether they have a regular market in the community, and we average this response across communities within each state to get a measure of the fraction of localities that have a market.

We also use a measure of land cost from the GHS-P post-planting community questionnaire as an instrument for the number of sellers in the same regression. As a proxy for land costs, we use the reported price to obtain an acre of land for which the owner has full property rights (as opposed to cultivation rights or sharecropping, for instance), and take an average across all communities within each state.

CPI microdata

We are grateful to David Atkin and Dave Donaldson for sharing apparel price data from the Nigerian consumer price index data collected by the Nigerian National Bureau of Statistics. For details on data collection protocols and treatment, see Atkin and Donaldson (2015).

We use price data from July 2010, as the period closest to our survey data collection that is available. Prices were collected monthly at multiple outlets in the capital city and two other urban locations in each state. We use the prices of 63 reported apparel goods, and within that limit consideration to the subset of 28 products for which prices are available in all Nigerian states in order to ensure that differences are not driven by differential product availability that may be related to the intermediation chain structure we study. limit taken from CPI microdata. For the purposes of Section 5, we use the average price across these common apparel goods at the state level.

SMEDAN data

We use estimates of the number of wholesale and retail traders in each state from a report on the Small and Medium Enterprises Development Agency of Nigeria (SMEDAN) and National Bureau of Statistics (NBS) Collaborative Survey on small enterprises conducted in 2013. The report provides a count of micro and small enterprises in each state in Nigeria. It also shows the percent of businesses in each category that are engaged in wholesale and retail trade nationwide, which is 55% of microenterprises and 21% of small and medium enterprises. We assume this nationwide sector breakdown holds in each state. We take the fraction of wholesale and retail traders that sell apparel from our LTS data, and also assume that this holds in each state. These calculations allow us to estimate the number of apparel traders in each state in Nigeria.

Domestic travel times

We take average travel times by road between the capital cities of each state in Nigeria from Google Maps. Google Maps queries were made and average travel times recorded by research assistants over the course of a week in February 2019. To the extent that the sparsity or quality of the road network may covary with remoteness in Nigeria, travel time is likely to offer a better proxy for trade costs than simple distance.

A.3 Supplementary tables

Table 7: Relationship between supplier type and firm size and sourcing costs

	(1) Fixed trade cost ($US) excl. agent & "other" costs	(2) Total trade costs ($US)	(3) Total trade costs ($US)
% of purchases from wholesaler	-459.90	-746.86	-683.84
	(314.42)	(657.47)	(625.48)
Total value of shipment ($US)		2.48	22.97
		(1.58)	(15.63)
Square of total value of shipment ($US)			-11521.14
			(7697.81)
Obs.	650	213	213
Mean of dependent variable	993.55	2970.87	2970.87
Product category FEs	x	x	x

Note: All columns have observations at the shipment level. Columns (2) and (3) have fewer observations because many shipments do not have the total value of purchases from all suppliers reported. Observations with variable profit ratios (sale price divided by purchase price) below the 10th or above the 90th percentiles are trimmed in (4) and (5). Standard errors are clustered at the firm level.

Table 8: Utility function parameter estimates

	Log apparel spending/capita ($US)		
	(1)	(2)	(3)
	OLS	OLS	IV
Log apparel price ($US)	0.35	0.06	-0.77
	(0.27)	(0.38)	(0.79)
Log sellers	0.04	0.03	0.66*
	(0.11)	(0.11)	(0.37)
Log GDP/capita ($US)		0.11	0.15
		(0.15)	(0.28)
% localities with market		0.24	0.62
		(0.52)	(0.45)
Obs.	33	33	33

Note: Data on the number of sellers is not available for 3 out of 37 states (Adamawa, Borno, and Yobe), and data on average land cost is missing for an additional state (Anambra). Data is drawn from the LSMS, SMEDAN, NBS, and Google Maps.

B Material from Section 3

B.1 Restrictions on utility function

In addition to imposing that $G'(\cdot) > 0$, $G''(\cdot) < 0$, $f(0) = 0$, $f'(\cdot) > 0$ and $f''(\cdot) \leq 0$, we assume that a second order condition on firm profits holds for firm quantity setting, and that equilibria are stable. In this subsection of the appendix, we provide conditions on $f(\cdot)$ and $G(\cdot)$ such that these additional assumptions hold. The notation is simplified by defining $Q \equiv \sum_{v'=1}^{N} f(q_{v'})$. Note that $\frac{\partial Q}{\partial q_v} = 0$ when firms are "small" with respect to the market and $\frac{\partial Q}{\partial q_v} = f'(q_v)$ otherwise.

The second order condition for firm quantity setting requires for any Q and q_v that $2\left[G''f'\frac{\partial Q}{\partial q_v} + G'f''\right] + q_v\left[G'''f'\left(\frac{\partial Q}{\partial q_v}\right)^2 + 2G''f''\frac{\partial Q}{\partial q_v} + G''f'\frac{\partial^2 Q}{\partial q_v^2} + G'f'''\right] \leq 0$. (Note that we have omitted arguments of the functions to shorten the expression).

Stability requires that marginal revenue for all traders is weakly falling in Q, i.e. $\frac{\partial MR}{\partial Q} \leq 0$. Formally, this requires that, for any Q and q_v, $G''f' + q_v\left(G'''f'\frac{\partial Q}{\partial q_v} + G''f''\right) \leq 0$. (Note that we have omitted arguments of the functions to shorten the expression). In the event that traders are "small" with respect to the market, this condition will hold with the assumptions made on the first two derivatives of $f(\cdot)$ and $G(\cdot)$, while if firms are large with respect to the market, this places an additional constraint on $G'''(\cdot)$.

B.2 Derivation of equation 6

To derive Equation 6, we take a second order approximation to profits around the equilibrium level of variable costs, c, and fixed costs, holding all other firms' behavior fixed

$$\pi\left(c', F'\right) \approx \pi(c, F) + \frac{\partial \pi(c, F)}{\partial c}\left(c' - c\right) + \frac{1}{2}\frac{\partial^2 \pi(c, F)}{\partial c^2}\left(c' - c\right)^2 + F' - F$$

It follows immediately from the envelope theorem that $\frac{\partial \pi}{\partial c} = -q$, and thus

$$\begin{aligned} \frac{\partial^2 \pi}{\partial c^2} &= \frac{\partial q}{\partial p}\frac{\partial p}{\partial c} \\ &= \frac{q}{c}\left(\varepsilon_p^q - 1\right)\rho \end{aligned}$$

where ρ is the equilibrium passthrough rate and ε_p^q is the price elasticity of demand. The second line follows from firms' profit-maximizing quantity choices.

Re-arranging the original second order approximation

$$\pi\left(c'\right) - \pi(c) = -q\left(c' - c\right)\left[1 + \frac{1}{2}\frac{\left(c' - c\right)}{c}\left(\varepsilon_p^q - 1\right)\rho\right] + F' - F$$

Thus, for an indirect sourcing equilibrium to hold, it must be that a change from indirect to direct sourcing yields (weakly) decreasing profits, giving the condition presented in the text.

B.3 Proof of Lemma 1

Lemma 1 A small change in the level of fixed cost around the level in the free market equilibrium yields the following comparative statics: $\frac{d}{dF}q\left(F_{ME}\right) > 0$, $\frac{d}{dF}N\left(F_{ME}\right) < 0$, and $\frac{d}{dF}\theta\left(F_{ME}\right) = 0$, where F_{ME} is the level of fixed cost chosen in the free market equilibrium.

Proof As discussed in the text, the free market equilibrium is defined by

$$\begin{aligned}
\pi\left(q,\theta,F\right) &= 0 \\
\frac{\partial}{\partial q}\pi\left(q,\theta,F\right) &= 0 \\
\frac{\partial}{\partial F}\pi\left(q,\theta,F\right) &= 0
\end{aligned}$$

where in the second and third terms, the partial derivatives are holding the other two arguments of the profit function fixed. The first two conditions hold even when F is chosen by the planner, while the third condition characterizes the free market equilibrium.

Notably, when the planner chooses F, the first two equations determine q and N as a function of F. Taking the derivative of the free entry condition with respect to F, we find

$$\begin{aligned}
0 &= \frac{d}{dF}\pi\left(q,\theta,F\right) \\
&= \frac{\partial\pi}{\partial q}\frac{dq}{dF} + \frac{\partial\pi}{\partial\theta}\frac{d\theta}{dF} + \frac{\partial\pi}{\partial F}
\end{aligned}$$

At the free market equilibrium, $\frac{\partial\pi}{\partial F} = 0$, and $\frac{\partial\pi}{\partial q} = 0$ via optimal choice of quantity (i.e. it is an envelope condition). And $\frac{\partial\pi}{\partial\theta} = q\frac{\partial p}{\partial\theta} < 0$. Thus, it must be that $\frac{d}{dF}\theta\left(F_{ME}\right) = 0$.

Next, we take the derivative of the condition for optimal choice of quantity with respect to F around the free market equilibrium:

$$\begin{aligned}
0 &= \frac{d}{dF}\left[\frac{\partial}{\partial q}\pi\left(q,\theta,F\right)\right] \\
&= \frac{\partial^2\pi}{\partial q^2}\frac{dq}{dF} + \frac{\partial^2\pi}{\partial q\partial\theta}\frac{d\theta}{dF} + \frac{\partial^2\pi}{\partial q\partial F}
\end{aligned}$$

And $\frac{\partial^2\pi}{\partial q\partial F} = -c'\left(F\right) > 0$, $\frac{\partial^2\pi}{\partial q^2} < 0$ by the second order condition, while we just showed

$\frac{d\theta}{dF} = 0$. Consequently

$$\frac{d}{dF} q\left(F_{ME}\right) > 0$$

Finally since $\frac{d}{dF}\theta\left(F_{ME}\right) = 0$ and $\frac{d}{dF} q\left(F_{ME}\right) > 0$, it must be that $\frac{d}{dF} N\left(F_{ME}\right) < 0$.

B.4 Equation 8

Equation (8) follows immediately from the first order condition of consumer surplus with respect to fixed cost and free entry. We start from the expression for welfare

$$W = U - Npq$$

We then take the derivative of this expression with respect to F to obtain

$$\begin{aligned} \frac{dW}{dF} &= \frac{\partial U}{\partial N}\frac{dN}{dF} + \frac{\partial U}{\partial N}\frac{dq}{dF} - pq\frac{dN}{dF} - Np\frac{dq}{dF} - Nq\frac{dp}{dF} \\ &= \left(\frac{\partial U}{\partial N} - pq\right)\frac{dN}{dF} - Nq\frac{dp}{dF} \end{aligned}$$

and from the zero-profit condition

$$\begin{aligned} p &= c + \frac{F}{q} \\ \frac{dp}{dF} &= c'\left(F\right) + \frac{1}{q} - \frac{F}{q^2}\frac{dq}{dF} \\ q\frac{dp}{dF} &= \left(c'\left(F\right) q + 1\right) - \left(p - c\right)\frac{dq}{dF} \end{aligned}$$

so that

$$\frac{dW}{dF} = \left(\frac{\partial U}{\partial N} - pq\right)\frac{dN}{dF} + N\left(p - c\right)\frac{dq}{dF} - N\left(c'\left(F\right) q + 1\right)$$

B.5 Proof of Proposition 1

Proposition 1 If the assumptions on the form of demand from Section 3.1.1 hold, and there is a continuous sourcing cost frontier, the planner's preferred sourcing technology may have either a higher or lower fixed cost than the market equilibrium technology.

Proof The proof is by example. For brevity, we use Equation 8 in the text (derived earlier in this Appendix) rather than re-deriving it here.

We start with expressions for $\frac{dq}{dF}$ and $\frac{dN}{dF}$ at the market equilibrium. In the proof of

Lemma 1, we established that around the market equilibrium

$$\frac{dq}{dF} = \frac{c'}{\frac{\partial^2 \pi}{\partial q^2}}$$

so that if we define N_{-v} as the number (or measure) of competitors, depending on whether firms are large or small with respect to the market, then it follows from $\frac{d\theta}{dF} = 0$ at the market equilibrium that

$$\begin{aligned} 0 = \frac{d\theta}{dF} &= N_{-v} f' \frac{dq}{dF} + f \frac{dN}{dF} \\ \frac{\frac{dN}{dF}}{\frac{dq}{dF}} &= -N_{-v} \frac{f'}{f} \frac{dq}{dF} \end{aligned}$$

We then turn to Equation 8 in the text. Using $U = G(Nf(q))$ (and dropping the arguments of functions for brevity), we can write

$$\begin{aligned} \frac{dW}{dF} &= G' \cdot (f - qf') \frac{\partial N}{\partial F} + N(G'f' - c) \frac{\partial q}{\partial F} \\ &= \left[-\frac{N_{-v}}{N} p \left(1 - q\frac{f'}{f}\right) + (p - c) \right] N \frac{\partial q}{\partial F} \end{aligned}$$

We will focus on the sign of $-\frac{N_{-v}}{N} p \left(1 - q\frac{f'}{f}\right) + (p - c)$ as $N\frac{\partial q}{\partial F} > 0$.

First, consider a cournot setting where firms are large, $G(Q) = A\left(Q - \frac{B}{2}Q^2\right)$, and $f(q) = q$. The particular $c(F)$ is arbitrary – it does not matter as long as it satisfies the conditions laid out in the text. In this setting, $1 - q\frac{f'}{f} = 0$, while $p - c > 0$ as fixed costs are strictly positive. Consequently, $\frac{dW}{dF} > 0$.

Alternatively, consider a setting where firms are small with respect to the market, $G(Q) = \ln Q$, $f(q) = q^{\frac{1}{2}} + q$, and $c(F) = \exp(-F)$. There is no cost of entry. It can be shown numerically that at the equilibrium $q \approx 1.2039$, $N \approx 0.6424$, and with these values and setup that

$$\begin{aligned} -\frac{N_{-v}}{N} p \left(1 - q\frac{f'}{f}\right) + (p - c) &= p \cdot q\frac{f'}{f} - c \\ &\approx -0.0806 \end{aligned}$$

Consequently, $\frac{dW}{dF} < 0$.

B.6 Lemma 2

Lemma 2 A discrete change to a higher fixed cost technology starting from the market equilibrium leads to a constrained equilibrium with higher quantity per traders, fewer traders, and lower competition: $\Delta q > 0$, $\Delta N < 0$, and $\Delta\theta \leq 0$.

Proof First, we establish that $\Delta\theta \leq 0$. This follows from revealed preference. Holding θ fixed, if traders chose to source from j, then

$$\pi\left(q_j, \theta_j, j\right) \geq \pi\left(q'_{j'}, \theta_j, j'\right)$$

where, as in the text, $q'_{j'}$ denotes the quantity which maximizes profits given θ_j and the cost structure of j'. In turn, by revealed preference, since when deviating firms choose $q'_{j'}$ and not $q_{j'}$ (which is the quantity which would maximize profits given $\theta_{j'}$ and the cost structure of j')

$$\pi\left(q'_{j'}, \theta_j, j'\right) \geq \pi\left(q_{j'}, \theta_j, j'\right)$$

wConsequently,

$$\pi\left(q_j, \theta_j, j\right) \geq \pi\left(q_{j'}, \theta_j, j'\right)$$

However, due to the zero profit condition, $0 = \pi\left(q_j, \theta_j, j\right)$, and it must be that $0 = \pi\left(q_{j'}, \theta_{j'}, j'\right)$. Since profits are strictly decreasing in θ, it must be that $\theta_{j'} \leq \theta_j$, and so $\Delta\theta \leq 0$.

Second, we establish that $\Delta q > 0$. In both equilibria, firms choose quantities so that

$$c\left(F\right) = p\left(q, \theta\right) + q\frac{\partial p\left(q, \theta\right)}{\partial q}$$

In changes, then (as before, using j' and j subscripts to denote equilibrium levels of variables under the respective sources)

$$\begin{aligned}
\Delta c &= \Delta\left(p\left(q, \theta\right) + q\frac{\partial p\left(q, \theta\right)}{\partial q}\right) \\
&= \left[p\left(q_{j'}, \theta_{j'}\right) + q_{j'}\frac{\partial p\left(q_{j'}, \theta_{j'}\right)}{\partial q} - p\left(q_j, \theta_{j'}\right) - q_j\frac{\partial p\left(q_j, \theta_{j'}\right)}{\partial q}\right] + \ldots \\
&\ldots\left[p\left(q_j, \theta_{j'}\right) + q_j\frac{\partial p\left(q, \theta_{j'}\right)}{\partial q} - p\left(q_j, \theta_j\right) - q_j\frac{\partial p\left(q_j, \theta_j\right)}{\partial q}\right] \\
&= \int_{q_j}^{q_{j'}}\left(\frac{\partial}{\partial q}\left(p\left(q, \theta'\right) + q\frac{\partial p\left(q, \theta'\right)}{\partial q}\right)\right)dq + \int_{\theta_j}^{\theta_{j'}}\left(\frac{\partial}{\partial\theta}\left(p\left(q, \theta\right) + q\frac{\partial p\left(q, \theta\right)}{\partial q}\right)\right)d\theta
\end{aligned}$$

Since we are considering a change to a higher fixed cost source, marginal cost must fall, i.e. $\Delta c < 0$. Furthermore, our assumption that preferences are such that equilibria are stable means that for all q and θ,

$$0 \geq \frac{\partial}{\partial \theta}\left(p + q\frac{\partial p\left(q,\theta\right)}{\partial q}\right)$$

and thus as $\theta_{j'} \leq \theta_j$

$$0 \geq \int\limits_{\theta_j}^{\theta_{j'}} \left(\frac{\partial}{\partial \theta}\left(p\left(q,\theta\right) + q\frac{\partial p\left(q,\theta\right)}{\partial q}\right)\right) d\theta$$

i.e., a (weakly) lower θ in the new equilibrium will lead to a (weakly) higher MR, quantity held equal. Consequently, it must be that

$$0 < \int\limits_{q_j}^{q_{j'}} \left(\frac{\partial}{\partial q}\left(p\left(q,\theta'\right) + q\frac{\partial p\left(q,\theta'\right)}{\partial q}\right)\right) dq$$

The second order condition for the choice of quantity implies that for any q and z,

$$0 \geq \frac{\partial}{\partial q}\left(p + q\frac{\partial p\left(q,\theta\right)}{\partial q}\right)$$

which implies $q_{j'} > q_j$.

And finally, we establish that $\Delta N < 0$. By definition

$$\theta = N_{-v} f\left(q\right)$$

so that

$$\Delta\theta = \left(N_{-v}\right)_{j'} \int\limits_{q_j}^{q_{j'}} f'\left(q\right) dq + f\left(q_j\right)\Delta N$$

And by assumption, $f'\left(q\right) \geq 0$ for all q, and we have showed that $\Delta q > 0$, so that $\left(N_{-v}\right)_{j'} \int_{q_j}^{q_{j'}} f'\left(q\right) dq > 0$. But $\Delta\theta \leq 0$ so that it must be $f\left(q_j\right)\Delta N < 0$. By assumption, $f\left(q\right) \geq 0$ for all q, so that $\Delta N < 0$.

B.7 Equation 9

Equation (9) follows from expanding the difference in welfare. We start from the expression for welfare

$$W(N,q) = U(N,q) - Npq$$

We then take differences under the two levels of F (using j' and j subscripts to denote equilibrium levels of variables under the respective sources)

$$\begin{aligned}\Delta W = & \left[U\left(N_{j'},q_{j'}\right) - U\left(N_{j'},q_j\right)\right] + \left[U\left(N_{j'},q_j\right) - U\left(N_j,q_j\right)\right] - \ldots \\ & \ldots N_{j'}\Delta(pq) - (\Delta N)\,p_j q_j\end{aligned}$$

and from the zero-profit condition

$$\begin{aligned} pq &= cq + F \\ \Delta(pq) &= c_{j'}\Delta q + q_j \Delta c + \Delta F \end{aligned}$$

so that

$$\begin{aligned}\Delta W = & \left[U\left(N_{j'},q_{j'}\right) - U\left(N_{j'},q_j\right)\right] + \left[U\left(N_{j'},q_j\right) - U\left(N_j,q_j\right)\right] - \ldots \\ & \ldots N_{j'}\left(c_{j'}\Delta q + q_j\Delta c_j + \Delta F\right) - (\Delta N)\,p_j q_j \\ & = \left[U\left(N_{j'},q_j\right) - U\left(N_j,q_j\right) - (\Delta N)\,p_j q_j\right] + \ldots \\ & \ldots \left[U\left(N_{j'},q_{j'}\right) - U\left(N_{j'},q_j\right) - N_{j'}c_{j'}\Delta q\right] - N_{j'}\left(q_j\Delta c + \Delta F\right) \\ & = \int_{N_j}^{N_{j'}} \left[\frac{\partial}{\partial N}U(\nu,q_j) - p_j q_j\right] d\nu + N_{j'}\int_{q_j}^{q_{j'}} \left[p\left(N_{j'},\vartheta\right) - c_{j'}\right] d\vartheta - N_{j'}\left(q_j\Delta c + \Delta F\right)\end{aligned}$$

which is the expression provided in the text.

B.8 Proposition 2

Proposition 2 If the assumptions on the form of demand from Section 3.1.1 hold and the cost frontier is discrete, a discrete change to a higher fixed cost technology starting from the market equilibrium may increase or decrease welfare.

Proof In this setting, ambiguity can come from two sources. First, depending on the underlying functions G, f, c and the market structure, it may be bad for the planner to make an infinitesimally small increase in F as we proved in Proposition 1. If even an infinitesimally small increase in F is bad for welfare, then a discrete increase of any size will always lower welfare. But conversely, if an infinitesimally small increase in F is good for welfare, there is always a small enough discrete change which is beneficial. This is obvious, and so we omit a proof here.

However, as we discuss in the text, even when an infinitesimal increase in F would be good for welfare, a large enough change in F can be detrimental; we provide an example here. We return to the cournot setting from the proof of Proposition 1 where firms are large, $G(Q) = A\left(Q - \frac{B}{2}Q^2\right)$, and $f(q) = q$. Suppose there are two possible sourcing strategies: $c_1 = \frac{A}{5}$, $F_1 = \frac{1}{25}\frac{A}{B}$ and $c_2 = \frac{A}{10}$, $F_2 = \frac{81}{400}\frac{A}{B}$.[65]

It is well-known that the optimal cournot strategy with N firms and marginal cost c is

$$q = \frac{1}{N+1} \cdot \frac{A-c}{AB}$$

and given equilibrium prices and the free entry condition, this implies (ignoring integer constraints) that the optimal number of firms is

$$N = \frac{A-c}{\sqrt{ABF}} - 1$$

It can be shown that if firms are in an initial symmetric equilibrium from the first source, no firm will wish to deviate to sourcing from location 2.[66] Under the first source, 3 firms will operate and they will generate

$$Q = \frac{12}{20B}$$

while if the planner mandates sourcing from location 2, there will be a monopolist which produces quantity

$$q = Q = \frac{9}{20B}$$

Effectively, the fall in competition has outweighed the gains from lower marginal cost, even though welfare would be increased by a small increase in F.

[65]Note that as there are only two sources, we could easily fit both points with a cost function of the form $c(F) = a\exp(-\mu F)$ which satisfies the assumptions on $c(F)$ for the purposes of the example in Proposition 1.

[66]The easiest way to see this is that sourcing strategy two permits a monopolist to operate earning zero profits while location one permits three firms. Thus no firm will wish to deviate, as if it assumes the quantity by remaining firms is unchanged, it must earn negative profits.

C Material from Section 4

C.1 Microfoundation of consumer utility

In the main text, the utility is

$$U = q_0 + AV^{\alpha}$$
$$V = N_{ri}^{\gamma - \frac{1}{1-\sigma_r}} \left(\int\limits_{\omega \in N_{ri}} q(\omega)^{\frac{\sigma_r - 1}{\sigma_r}} d\omega \right)^{\frac{\sigma_r}{\sigma_r - 1}}$$

In this part of the appendix, we show that this utility can be microfounded as a discrete choice problem in the style of Anderson, de Palma, and Thisse (1987).

Suppose that utility for a consumer who spends X_0 on the outside good, E_1 on the traded good, and purchases the traded good from seller ω is:

$$U = X_0 + A \left(\max_{\omega} V(\omega) \right)^{\alpha}$$
$$V(\omega) = \frac{E_1}{p(\omega)} \varepsilon(\omega)$$

where $\varepsilon(\omega)$ is an idiosyncratic iid Frechet draw of match value with seller ω, and $p(\omega)$ is the price charged by that seller. The Frechet match value has shape parameter $\frac{1}{\mu_c}$ where $\mu_c < 1$[67] and scale parameter $\frac{M_c^{\gamma - \mu_c}}{\Gamma(1-\mu_c)}$, where $\Gamma(\cdot)$ denotes the gamma function. We assume consumers make two sequential decisions: first, they observe the measure of traders and distribution of prices, and choose expenditure on the outside good to maximize expected utility, and second, they observe their idiosyncratic match values and pick the seller who maximizes $V(\omega)$ given their choice of E_1.

The payout for agent with expenditure E_1 from buying variety ω at price $p(\omega)$ with idiosyncratic match value is

$$\frac{E_1}{p(\omega)} \varepsilon(\omega)$$

where $\varepsilon(\omega)$ is distributed Frechet with shape parameter $\frac{1}{\mu_c}$ and scale parameter $\frac{N_r^{\gamma - \mu_r}}{\Gamma(1-\mu_r)}$ where N_r is the number (or measure) of retailers ω and γ is a parameter.

This gives rise to CES demands across varieties ω. Maximizing sub-utility is the same as maximizing the monotone transformation

$$\ln E_1 - \ln p(\omega) + \ln \varepsilon(\omega)$$

[67] This condition delivers consumer elasticity of demand greater than 1.

where $\ln \varepsilon(\omega)$ is distributed Gumbel with location parameter $(\gamma - \mu_r) \ln N_r - \ln \Gamma(1-\mu_r)$ and scale parameter μ_r. As shown in Anderson, de Palma, and Thisse (1987) maximizing this monotone transformation of utility gives rise to CES demands with elasticity of substitution

$$\sigma_r = 1 + \frac{1}{\mu_r}$$

Next, determining the expenditure will depend on the expected value of the maximum draw. Note that because $\left(\max_\omega \frac{E_1}{p(\omega)} \varepsilon(\omega)\right)^\alpha$ is a monotone transformation of $\max_\omega \frac{E_1}{p(\omega)} \varepsilon(\omega)$, the parameter α will not affect the choice of the utility maximizing seller. And, due to the properties of the Frechet distribution, $\max_\omega \frac{E_1}{p(\omega)} \varepsilon(\omega)$ will be distributed Frechet with shape parameter $\frac{1}{\mu_r}$ and scale parameter $\left(\frac{E_1 N_r^{\gamma-\mu_r}}{\Gamma\left(1-\frac{1}{\mu_r}\right)}\right)\left(\sum_\omega p(\omega)^{-\frac{1}{\mu_r}}\right)^{\mu_r}$. Thus, $\left(\max_\omega \frac{E_1}{p(\omega)} \varepsilon(\omega)\right)^\alpha$ will be distributed Frechet with scale $\left(\frac{E_1 N_r^{\gamma-\mu_r}}{\Gamma\left(1-\frac{1}{\mu_r}\right)}\right)^\alpha \left(\sum_\omega p(\omega)^{-\frac{1}{\mu_r}}\right)^{\alpha\mu_r}$ and shape parameter $\frac{1}{\alpha\mu_r}$, so that the expected value for the maximum draw will be

$$\mathbb{E}\left[\left(\max_\omega V(\omega)\right)^\alpha\right] = \frac{E_1^\alpha}{\left(\sum_\omega \left(\frac{1}{(p(\omega))}\right)^{-\frac{1}{\mu_r}}\right)^{-\alpha\mu_r}} N_r^{\alpha(\gamma-\mu_r)}$$

If we define the price index in the usual way (and where following Anderson, de Palma, and Thisse (1987) we define $1 - \sigma_r = -\frac{1}{\mu_r}$), i.e. that

$$P_r \equiv \left(\sum_\omega (p(\omega))^{1-\sigma_r}\right)^{\frac{1}{1-\sigma_r}}$$

then

$$\mathbb{E}\left[\max_\omega V(\omega)^\alpha\right] = \left(\frac{E_1}{P_r N_r^{-\frac{1}{1-\sigma_r}-\gamma}}\right)^\alpha$$

Note that this suggests that $\max_\omega V(\omega)$ is identical to standard CES utility in the event that $\gamma = \frac{1}{\sigma_r - 1}$. If not, following Benassy (1996), γ is a parameter governing gains from variety.

Thus, given a distribution of prices and measure of sellers which lead to the price index $\Omega_r^{-\frac{1}{1-\sigma_r}-\gamma} P_r$, following immediately from the FOC for E_1, the consumer will choose

$$E_1 = \left[\alpha A \left(N_r^{-\frac{1}{1-\sigma_r}-\gamma} P_r\right)^{-\alpha}\right]^{\frac{1}{1-\alpha}}$$

Thus consumers will have expected utility from the differentiated sector of

$$\mathbb{E}\left[\left(\max_{\omega} V\left(\omega\right)\right)^{\alpha}\right] = \left(\frac{\alpha A}{N_r^{-\frac{1}{1-\sigma_r}-\gamma} P_r}\right)^{\frac{\alpha}{1-\alpha}}$$

C.2 Generalization of results from Section 3 to encompass utility like Benassy (1996)

In this section, we consider CES utility with a divergence between the social and private value of variety as in Benassy (1996) which does not fit in the utility form described in Equation (1) and instead matches the form of Equation (11). We show that our results from Section 3 still hold in this framework.

In particular, we extend our results from Section 3 to utility of the form

$$U = \frac{1}{h\left(N\right)} G\left(\sum_{v=1}^{N} h\left(N\right) f\left(q_v\right)\right)$$

In addition to the assumptions we made before, we additionally assume $h > 0$, $-\frac{h'}{h^2}\frac{G}{G'} + f + N\frac{h'}{h} f \geq 0$ (i.e. more variety never makes consumers worse off), and $N\frac{h'}{h} > -1$ (which implies that an increase in the measure or number of varieties lowers consumer prices). Note that Equation (11) fits this form: define $G\left(x\right) = AX^{\alpha\frac{\sigma}{\sigma-1}}$, $f\left(x\right) = x^{\frac{\sigma_r-1}{\sigma_r}}$, and $h\left(x\right) = x^{\gamma\left(\frac{\alpha\sigma}{(\alpha-1)\sigma+1}\right)}$ to obtain the direct utility form of Equation (11).

.

First, we adopt a slightly different definition of the level of competition; now we define (when firms are large) the level of competition faced by firm v, θ_v by

$$\theta_v \equiv h\left(N\right) \sum_{v'=1}^{N} f\left(q_{v'}\right) - f\left(q_v\right)$$

while if firms are small

$$\theta_v \equiv h\left(N\right) \int_{0}^{N} f\left(q_{v'}\right) dv'$$

We also change our notion of stability to account for the changed definition of θ so that $0 \geq \frac{\partial}{\partial\theta}\left(p + q\frac{\partial p(q,\theta)}{\partial q}\right)$ as before, and place the necessary assumptions on the utility function and its components..

Next, we show that, under this definition of competition, both Lemma 1 and Lemma 2 hold. We start with Lemma 1: with this new definition of competition, we can express profits as a function of per-firm quantity, level of competition, and level of fixed cost.[68] $\frac{d\pi}{dF} = 0$ due to the free entry condition, which implies $\frac{d\theta}{dF} = 0$. This implies that $\frac{dq}{dF} > 0$ for the same reason as under the initial utility level. Furthermore, under the new definition of competition (and assumptions of the utility function and its constituent functions), it is easy to verify that $\frac{d\theta}{dq} > 0$ and $\frac{d\theta}{dN} < 0$. Thus $\frac{d\theta}{dF} < 0$. Similarly, for Lemma 2, one can make the same revealed preference argument to establish that $\Delta\theta \leq 0$. Similarly, we can establish $\Delta q > 0$ through the stability condition (the steps are unchanged). And finally, using the new definition of θ, we derive

$$\Delta\theta = \left(N_{j'} h\left(N_{j'}\right) - 1\right) \int_{q_j}^{q_{j'}} f'(q)\, dq + \int_{N_j}^{N_{j'}} \left[N_{j'} f(q_j) h'(N) + f(q_j) h(N_j)\right] dN$$

Note that by assumption, $N_{j'} h\left(N_{j'}\right) > 1$ and $f' > 0$ so the first term is positive. Since $\Delta\theta \leq 0$, the second term must be negative, and by assumption $N_{j'} f(q_j) h'(N) + f(q_j) h(N_j) > 0$. Consequently, it must be that $\Delta N < 0$.

Note that both welfare decompositions presented in the text do not depend on the functional forms of utility. Consequently, both expressions still hold, and as before there is an tradeoff between per-firm quantity gains and variety losses from movements to higher levels of fixed cost. To show that this tradeoff is ambiguous, it suffices to set $h(N) = 1$, in which case the existing examples in Propositions 1 and 2 demonstrate the ambiguity. However, it can also be shown that, for any f and G, as $-\frac{h'}{h^2}\frac{G}{G'} + f + N\frac{h'}{h}f \to 0$ (so that consumers are indifferent to additional variety), the per-firm quantity gains dominate and the planner will always choose to increase the level of fixed cost. Similarly, as $-\frac{h'}{h^2}\frac{G}{G'} + f + N\frac{h'}{h}f \to \infty$, variety gains will dominate any per-firm quantity effects and the planner will always reduce the fixed cost.

C.3 Intermediary payouts

In this section, we provide details about the distribution of intermediaries' shocks.

We assume that (for each intermediation activity u), $\xi_u(j)$ is distributed iid Gumbel

[68] The key is that under this utility and definition of competition

$$\begin{aligned} p_v &= \frac{\partial U}{\partial q_v} \\ &= G'\left(\sum_{v'=1}^{N} h(N) f(q_{v'})\right) f'(q_v) \\ &= G'\left(\theta_v + f(q_{v'})\right) f'(q_v) \end{aligned}$$

i.e. the price for a given firm is simply a function of own quantity and the level of competition as before.

with scale parameter s and location parameter $-s\Gamma'(1)$ where $\Gamma'(1)$ is the derivative of the gamma function evaluated at 1 (and is equal to the negative of the Euler-Mascheroni constant).

We assume that (for each intermediation activity u), $\zeta_u(z)$ is distributed iid Frechet with shape parameter $\frac{1}{\beta}$ and scale parameter $\frac{1}{\Gamma(1-\beta_u)}\left[\frac{\sum_{z\in z_{iju}}\pi_z\pi_z^{\frac{1}{\beta}}}{\sum_{z'\in z_{iju}}\pi_{z'}^{\frac{1}{\beta}}}-\left(\sum_{z\in z_{iju}}\pi_z^{\frac{1}{\beta}}\right)^{\beta}\right]$ where $\Gamma(\cdot)$ is the gamma function, π_z is the expected profits conditional on choosing chain z, and z_{iju} is the set of chains going from j to i for intermediation activity u.

Finally, we assume that $\varepsilon_z(\omega)$ is distributed iid Frechet with shape parameter $\frac{1}{\mu_t}$ and scale parameter $\frac{N_z^{\frac{1}{1-\sigma_z}}}{\Gamma(1-\mu_t)}$ where $\Gamma(\cdot)$ is the gamma function, N_z is the measure of sellers on chain z, and σ_z is the elasticity of demand at the relevant stage of chain z (note that this elasticity is solved for in the body of the paper). Note that this formulation means that the number of sellers serving a chain does not affect downstream payouts.

C.4 Consumer price index

In this section, we provide details for the derivation of the expression of the consumer price index presented in the paper, P_{ij}. This is equal to the probability a given retailer carries any particular chain times that chain's contribution to the consumer price index. Note that the probability of carrying chain z for a retailer is (as provided in Section 4.2.2 of the paper)

$$\Pr(z)=\frac{p_z^{\frac{1}{\beta}(1-\sigma_r)}}{\sum_{z'}p_{z'}^{\frac{1}{\beta}(1-\sigma_r)}}$$

where we can ignore the role of the number of sellers due to the shape parameter of seller match distribution – all that matters for trader payouts are prices. For convenience, we will define ϕ_{ij} by $\left((m_r\tau_{ij})^{\frac{1}{\beta_r}}\phi_{ij}\right)^{1-\sigma_r}=\sum_{z'}p_{z'}^{\frac{1}{\beta_r}(1-\sigma_r)}$ – i.e. it is a (transformation of the) profit index. Using this definition, we can write for location $j\neq o$

$$P_{ij}^{1-\sigma_r}=(M_2\tau_{ij}\tau_{jo}p_o)^{1-\sigma_r}\frac{(M_2\tau_{ij}\tau_{jo}p_o)^{\frac{1}{\beta}(1-\sigma_r)}}{\left((m_r\tau_{ij})^{\frac{1}{\beta}}\phi_{ij}\right)^{1-\sigma_r}}+\sum_{j'}\left(M_3\tau_{ij}\tau_{jj'}\tau_{j'o}p_o\right)^{1-\sigma_r}\frac{\left(M_3\tau_{ij}\tau_{jj'}\tau_{j'o}p_o\right)^{\frac{1}{\beta}(1-\sigma_r)}}{\left((m_r\tau_{ij})^{\frac{1}{\beta}}\phi_{ij}\right)^{1-\sigma_r}}+\sum_{j''}\sum_{j'}\left(M_4\tau_{ij}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o\right)^{1-\sigma_r}\,(M_4$$

$$=\frac{(m_r\tau_{ij})^{1-\sigma_r}}{\phi_{ij}^{1-\sigma_r}}\left[\left(\frac{M_2}{m_r}\tau_{jo}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)}+\sum_{j'}\left(\frac{M_3}{m_r}\tau_{jj'}\tau_{j'o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)}+\sum_{j''}\sum_{j'}\left(\frac{M_4}{m_r}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)}+\ldots\right]$$

and by close analogy

$$P_{io}^{1-\sigma_r}=\frac{(m_c\tau_{io})^{1-\sigma_r}}{\phi_{io}^{1-\sigma_r}}\left[p_o^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)}+\left(\frac{M_2}{m_r}\tau_{oo}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)}+\sum_{j'}\left(\frac{M_3}{m_r}\tau_{oj'}\tau_{j'o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)}+\ldots\right]$$

We next turn to the profit indexes, and we show (for $j \neq i$, we will neglect the origin as the definition is provided in the text and by this point it is clear by analogy)

$$\left((m_c\tau_{ij})^{\frac{1}{\beta}}\,\phi_{ij}\right)^{1-\sigma_r} = (M_2\tau_{ij}\tau_{jo}p_o)^{\frac{1}{\beta}(1-\sigma_r)} + \sum_{j'}(M_3\tau_{ij}\tau_{jj'}\tau_{j'o}p_o)^{\frac{1}{\beta}(1-\sigma_r)} + \sum_{j''}\sum_{j'}(M_4\tau_{ij}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o)^{\frac{1}{\beta}(1-\sigma_r)} + \ldots$$

$$\phi_{ij}^{1-\sigma_r} = \left(\frac{M_2}{m_r}\tau_{jo}p_o\right)^{\frac{1}{\beta}(1-\sigma_r)} + \sum_{j'}\left(\frac{M_3}{m_r}\tau_{jj'}\tau_{j'o}p_o\right)^{\frac{1}{\beta}(1-\sigma_r)} + \sum_{j''}\sum_{j'}\left(\frac{M_4}{m_r}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o\right)^{\frac{1}{\beta}(1-\sigma_r)} + \ldots$$

where since it is independent of i, we define $\phi_{ij}^{1-\sigma_r} \equiv \phi_j^{1-\sigma_r}$ shared across all i. This then lets us define the $\mathcal{P}_j$ as in the text, so that

$$P_{ij}^{1-\sigma_r} = (m_r\tau_{ij})^{1-\sigma_r}\,P_j^{1-\sigma_r}$$

yielding the expression in the text that

$$P_{ij} = m_r\tau_{ij}P_j$$

C.5 Chain length

The expenditure-weighted length of chains going from location j to location i will reflect the share of expenditure on the route allocated to each chain times the length of that chain. Formally, for a location $j \neq o$,

$$\Lambda_{ij} = \frac{1}{\left((m_r\tau_{ij})^{\frac{1}{\beta}}\,\phi_{ij}P_{ij}\right)^{1-\sigma_r}}\left[2\,(M_2\tau_{ij}\tau_{jo}p_o)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 3\sum_{j'}(M_3\tau_{ij}\tau_{jj'}\tau_{j'o}p_o)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 4\sum_{j''}\sum_{j'}(M_4\tau_{ij}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o)\right.$$

As with the price indeces, it is possible to factor out the i-specific components

$$L_{ij} = \frac{1}{(\phi_j\mathcal{P}_j)^{1-\sigma_c}}\left[2\left(\frac{M_2}{m_r}\tau_{jo}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 3\sum_{j'}\left(\frac{M_3}{m_r}\tau_{jj'}\tau_{j'o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 4\sum_{j''}\sum_{j'}\left(\frac{M_4}{m_r}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} +\right.$$

so that we can define the shared part of chain length as

$$\Lambda_j \equiv 2\left(\frac{M_2}{m_r}\tau_{jo}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 3\sum_{j'}\left(\frac{M_3}{m_r}\tau_{jj'}\tau_{j'o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 4\sum_{j''}\sum_{j'}\left(\frac{M_4}{m_r}\tau_{jj'}\tau_{j'j''}\tau_{j''o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + \ldots$$

And for chains from the origin

$$\Lambda_{io} = \frac{1}{\left((m_r\tau_{io})^{\frac{1}{\beta}}\,\phi_{io}P_{io}\right)^{1-\sigma_c}}\left[(M_1\tau_{io}p_o)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 2\,(M_2\tau_{oo}p_o)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 3\sum_{j'}(M_3\tau_{io}\tau_{oj'}\tau_{j'o}p_o)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + \ldots\right]$$

so that we similarly simplify and define

$$\Lambda_o = (p_o)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 2\left(\frac{M_2}{m_r}\tau_{oo}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + 3\sum_{j'}\left(\frac{M_3}{m_r}\tau_{oj'}\tau_{j'o}p_o\right)^{\left(1+\frac{1}{\beta}\right)(1-\sigma_r)} + \ldots$$

Using these definitions, for all ij pairs we can write

$$\Lambda_{ij} = \frac{\Lambda_j}{\left(\phi_j \mathcal{P}_j\right)^{1-\sigma_r}}$$

C.6 Proposition 3

Proposition 3 If the assumptions on the setting and consumer and trader payouts of Section 4.1 hold then the equilibrium is unique.

Proof As established in the prior subsection of this appendix, the P_{ij} are only a function of model parameters and not the equilibrium choices of intermediaries. Thus uniqueness boils down the sourcing choices of retailers and consumer price indeces, and these choices are separable across locations.

First, we show that, holding the measure of traders in a final location fixed, this implies a unique sourcing pattern. In particular, we show that when we define the function $e\left(P_{r,i}\right)$ (for notational simplicity)

$$e\left(P_{r,i}\right) = P_{r,i}^{1-\sigma_r} - N_{r,i}\frac{\sum_j \exp\left(\frac{\pi_{ij}^r}{s}\right)}{\sum_{j'} \exp\left(\frac{\pi_{ij'}^r}{s}\right)}P_{ij}^{1-\sigma_r}$$

this function is strictly decreasing in P_{ij} ; this montonicity implies a unique value of P_{ij} (which in turn implies a unique sourcing pattern). The intuition for this result is that a rise in the consumer price index will make sourcing from all locations more profitable, but disproportionately so for locations which have the lowest price indeces. Formally, we start from the expression for expenditure in terms of the price index

$$\frac{E_i}{P_{r,i}^{1-\sigma_r}} = CN_{r,i}^{\delta}P_{r,i}^{\psi}$$

where have adopted for notational convenience $C \equiv \frac{Z_i(\alpha A)^{\frac{1}{1-\alpha}}}{s\sigma_r}$,$\delta \equiv \frac{-\alpha}{1-\alpha}\left(-\frac{1}{1-\sigma_r} - \gamma\right)$ $\psi \equiv \sigma_r - 1 - \frac{\alpha}{1-\alpha}$. Then holding $N_{r,i}$ fixed

$$\frac{\partial}{\partial P_{r,i}}\exp\left(\frac{\pi_{ij}^r}{s}\right) = \delta CN_{r,i}^{\delta}P_{ij}^{1-\sigma_r}P_{r,i}^{\psi-1}\exp\left(\frac{\pi_{ij}^r}{s}\right)$$

Taking the derivative with respect to $P_{r,i}$ again holding $N_{r,i}$ fixed:

$$e'\left(P_{r,i}\right) = \left(1-\sigma_r\right) P_{r,i}^{-\sigma_c} - \psi N_{r,i}^{\delta} P_{r,i}^{\psi-1} N_{r,i} \cdot \sum_j P_{ij}^{1-\sigma_r} \chi_{ij} \left(P_{ij}^{1-\sigma_r} - \frac{P_{r,i}^{1-\sigma_r}}{N_{r,i}} \right)$$

where $\chi_{ij} \equiv \frac{\exp\left(\frac{\pi_{ij}^r}{s}\right)}{\sum_{j'} \exp\left(\frac{\pi_{ij'}^r}{s}\right)}$ is the sourcing share from j. Note that $N_{r,i} \sum_j \chi_{ij} P_{ij}^{1-\sigma_r} = P_{r,i}^{1-\sigma_r}$ by definition, so that

$$\sum_j P_{ij}^{1-\sigma_r} \chi_{ij} \left(P_{ij}^{1-\sigma_r} - \frac{P_{r,i}^{1-\sigma_r}}{N_{r,i}} \right) = \sum_j \chi_{ij} \left(P_{ij}^{1-\sigma_r} - \frac{P_{r,i}^{1-\sigma_r}}{N_{r,i}} \right) \left(P_{ij}^{1-\sigma_r} - \frac{P_{r,i}^{1-\sigma_r}}{N_{r,i}} \right)$$

which is simply the variance in the price index across locations and by definition weakly positive. Since restrictions on α imply $\delta > 0$ and we assumed $\sigma_r > 1$, it follows that $e'\left(P_{r,i}\right) < 0$ for all $P_{r,i}$. Thus, the solution to $e\left(P_{r,i}\right) = 0$ is unique holding the measure of firms fixed.

Second, we show that there is a unique measure of traders in equilibrium. Per-trader profits are strictly decreasing in the number of traders, such that there is only one measure of traders such that the expected profits are equal to the fixed cost of entry.

Before addressing this point directly, we develop a few expressions which will simplify future steps. We return to our expression for $\frac{E_i}{P_{r,i}^{1-\sigma_r}}$ from earlier, and now express it as

$$\frac{E_i}{P_{r,i}^{1-\sigma_r}} = C N_{r,i}^{\delta'} \left(\sum_j \chi_{ij} P_{ij}^{1-\sigma_r} \right)^{\frac{\psi}{1-\sigma_r}}$$

where $\delta' = \delta + \frac{\psi}{1-\sigma_r}$. Furthermore, it can be shown that

$$\frac{\partial \chi_{ij}}{\partial N_{r,i}} = \frac{1}{s\sigma_r} \chi_{ij} \left(P_{ij}^{1-\sigma_r} - \frac{P_{r,i}^{1-\sigma_c}}{N_{r.,i}} \right) \cdot \frac{\partial}{\partial N_{r,i}} \left(\frac{E_i}{P_{r,i}^{1-\sigma_r}} \right)$$

Using both of these expressions, we find

$$\frac{\partial}{\partial N_{r,i}} \left(\frac{E_i}{P_{r,i}^{1-\sigma_r}} \right) = \frac{\delta'}{N_{r,i}} \frac{E_i}{P_{r,i}^{1-\sigma_r}} + \frac{\frac{\psi}{1-\sigma_r} N_{r,i}}{P_{r,i}^{1-\sigma_r}} \frac{E_i}{P_{r,i}^{1-\sigma_r}} \sum_j P_{ij}^{1-\sigma_r} \frac{\partial \chi_{ij}}{\partial N_{r,i}}$$

$$\frac{\partial}{\partial N_{r,i}} \left(\frac{E_i}{P_{r,i}^{1-\sigma_r}} \right) = \frac{\frac{\delta'}{N_{r,i}} \frac{E_i}{P_{r,i}^{1-\sigma_r}}}{\left(\frac{1}{s\sigma_r} \frac{\frac{\psi}{\sigma_r - 1} N_{r,i}}{P_{r,i}^{1-\sigma_r}} \frac{E_i}{P_{r,i}^{1-\sigma_r}} V + 1 \right)}$$

where we define $V \equiv \sum_j \chi_{ij} \left(P_{ij}^{1-\sigma_r} - \frac{P_{r,i}^{1-\sigma_r}}{N_{r.,i}} \right) P_{ij}^{1-\sigma_r}$ as the variance in the P_{ij}.Note that $V > 0$. Following our assumption that $\gamma < \frac{1-\alpha}{\alpha}$, so that $\delta < -\frac{\alpha}{1-\alpha} \left(\frac{1}{\sigma_r - 1} \right) + 1$ and $\delta' < 0$. Furthermore, our assumptions on α imply $\psi > 0$. Consequently, $\frac{\partial}{\partial N_{r,i}} \left(\frac{E_i}{P_{r,i}^{1-\sigma_r}} \right) < 0$.

We now turn to the main result. Following the distribution of trader shocks, retail firm profits conditional on entry are

$$\mathbb{E}[\pi] = s \ln \left(\sum_j \exp \left(\frac{\pi_{ij}^r}{s} \right) \right)$$

Thus,

$$\frac{\partial \mathbb{E}[\pi]}{\partial N_{r,i}} = C N_{r,i}^{\delta'} \sum_j \chi_{ij} \frac{\partial}{\partial N_{r,i}} \left(\frac{E_i}{P_{r,i}^{1-\sigma_r}} \right)$$

This will be strictly negative under the conditions provided, and there is a unique measure of traders which will enter retail.

C.7 Solving for wholesale equilibrium

This section provides explains how we solve for wholesaling decisions and calculate average chain lengths (as weighted by the number of traders on each chain.[69]. Everything in this Appendix relies on the markup approximation described in Section 5 of the text.[70]

This is organized into five subsections. The first four subsections build toward optimal entry and sourcing decisions of wholesalers, and the fifth develops expressions for average chain lengths across all traders selling in a location. First, we derive an expression for the expected payouts for a trader based in i to source from each source j (partly in terms of equilibrium outcomes we solve for in later steps). Second, we derive expressions for value of goods traded wholesale on each chain. Doing requires two components: the ultimate consumer expenditure at the end of the chain (which depends on the final price on the chain – which depends only on fundamentals of that chain – and characteristics of the final destination, including equilibrium expenditure, price index, and the sourcing decisions of its retailers), as well as the downstream markups and trade costs, which are not reflected in the value of goods upstream. Third, we find the total wholesale expenditure to a given

[69] Note that this also relies on optimal retailing behavior as described in the text and elsewhere in the Appendix

[70] This assumption is not necessary – we could alternatively drop long chains (which play little role since there are sch low flows and low markups that they do not provide many wholesale profits and hence are not important for wholesaling behavior). However, the assumption simplifies the exposition substantially.

location from each source j. Fourth, we establish that this leads to unique sourcing and entry decisions. There is a final (fifth) section which relies on the equilibrium entry and sourcing decisions of the fourth part, and develops expressions for average chain lengths across all traders selling in a location.

C.7.1 Expected payout for each source

In this section, we develop an expression for the expected payouts for a trader based in i to source from each source j (partly in terms of equilibrium outcomes we solve for in later steps).

The first step is an expression for the share of traders serving an ij link on each step of each chain with an ij connection. The probability of choosing step n on chain z among the set of such objects with a link from j to i, denoted $\Pr(z,l)$, follows from the Frechet shocks across chains and is given by

$$\Pr(z,l) = \frac{\left(\pi_{z,l}^{var}\right)^{\frac{1}{\beta+1}}}{\sum_{z',l''}\left(\pi_{z',l''}^{var}\right)^{\frac{1}{\beta+1}}}$$

where $\pi_{z,l}^{var}$ is the variable profits from serving step l on chain z. in turn, the profits from serving step l on chain are given by

$$\pi_{z,l}^{var} = \frac{1}{\sigma_{l,z}}\frac{E_{z,l}}{N_{z,l}}$$

where on step l of chain z, $\sigma_{l,z}$ is the elasticity of demand across traders,[71] $E_{z,l}$ is the total expenditure, and $N_{z,l}$ is the measure of sellers. This expression follows from the standard CES profits and our assumption of symmetric sellers. And under the markups approximation,

$$\frac{1}{\sigma_{l,z}} \approx \frac{1}{\sigma_w}$$

where σ_w is shared for all wholesale steps. This, combined with the previous two expressions, gives us

$$\Pr(z,l) = \frac{E_{z,l}^{\frac{1}{\beta+1}}}{\sum_{z',l''}\left(E_{z',l''}^{\frac{1}{\beta+1}}\right)}$$

Next, we use the prior expression to solve for the profits of sourcing from each source

[71] This could be expressed in terms of step l and $L(z)$ following the formula in Section 4 of the text.

j. We start by denoting by N_{ij}^{w} the measure of wholesalers sourcing from j for i, so that the total number of traders

$$N_{z,l} = \Pr(z,l)\, N_{ij}^{w}$$

Thus, the expected variable payout of sourcing wholesale from j to sell in i (i.e. neglecting fixed costs and idiosyncratic draws), which we denote by $\pi_{ij}^{w,var}$, is given by

$$\begin{aligned}
\pi_{ijw}^{var} &= \sum_{z,l} \Pr(z,l)\, \pi_{z,n} \\
&\approx \frac{1}{\sigma_w} \sum_{z,n} \frac{E_{z,n}}{N_{ij}^{w}} \\
&= \frac{1}{\sigma_w} \cdot \frac{E_{ij}^{w}}{N_{ij}^{w}}
\end{aligned}$$

where E_{ijw} is the total wholesale expenditure going from j to i across all routes.

C.7.2 The value of goods on a given wholesale link

As we showed in the prior subsection, the expected variable profit for sourcing from j to sell in i as a wholesaler is a function of the total wholesale expenditure across all chains along that link. The expenditure values for all chains are simply a function of fundamentals and the choice of retail traders (which are themselves a unique function of fundamentals, as we have shown elsewhere). Thus, in this section, we develop an expression for expenditure on step l of chain z, which we denote by $E_{z,l}$ (and in the next section we will sum across all such chains). Throughout this section, we assume $l > L(z)$, i.e. this is a wholesale step on the chain.

Wholesale expenditure on step l of chain z reflects expenditure by consumers at the end of the chain, plus expenditure on markups and trade costs downstream (as these costs will not be passed up the chain). Expenditure at the end of the chain will be

$$E_{z,L(z)} = N_{r,L(z)} \chi_{rL(z),(L-1)(z)} \left(\frac{p_{z,N(z)}^{\frac{\beta+1}{\beta}}}{\phi_{L(z),(L-1)(z)} P_{r,L(z)}} \right)^{1-\sigma_r} E_{L(z)}$$

where in an abuse of notation, we use $\chi_{L(z),(L-1)(z)}^{r}$ to denote the share of retailers in the destination sourcing from the penultimate location on chain z, $P_{r,L(z)}$ for the consumer price index in the destination, and $N_{r,L(z)}$ to denote the number of retail traders in the destination.

And (in a repetition of results elsewhere in this appendix), the price at the end of the chain will be

$$p_{z,N(z)} = p_o \prod_{s=1}^{L(z)} \tau_{(s-1)(z)s(z)} m_s$$

and the downstream markups and trade costs (denoted by $C_{z,l}$) will be

$$C_{z,l} = \prod_{s=l+1}^{L(z)} \tau_{(s-1)(z)s(z)} m_s$$

where $\tau_{(l-1)(z)l(z)}$ is the trade cost incurred at step s and m_s i the markup at step s.

Taking these three expressions together, the value of expenditure flowing along chain z at step l is

$$\begin{aligned} E_{z,L(z)} &= \frac{E_{z,L(z)}}{C_{z,l}} \\ &= \frac{N_{r,L(z)} \chi_{rL(z),(L-1)(z)}}{C_{z,l}} \left(\frac{p_{z,N(z)}^{\frac{\beta+1}{\beta}}}{\phi_{L(z),(L-1)(z)} P_{r,L(z)}} \right)^{1-\sigma_r} E_{L(z)} \\ &= p_o^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_c)} \left(\prod_{s=1}^{l} \tau_{(s-1)(z)s(z)} m_s \right)^{\frac{\beta+1}{\beta}(1-\sigma_c)} \left(\prod_{s=l+1}^{L(z)} \tau_{(s-1)(z)s(z)} m_s \right)^{\frac{1-\sigma_c(\beta_r+1)}{\beta_r}} \dots \\ &\quad \dots \cdot \left(N_{r,L(z)} \chi_{rL(z),(L-1)(z)} E_{L(z)} \right) \left(\frac{1}{\phi_{L(z),(L-1)(z)} P_{r,L(z)}} \right)^{1-\sigma_r} E_{L(z)} \end{aligned}$$

C.7.3 Total expenditure from each source

As the result in the first subsection shows, a critical object is the aggregate expenditure on wholesale chains passing through the j to i link, E^w_{ij}. We solve for that object by summing across the wholesale expenditure on each chain with an ij link following the second subsection.

We start with a definition of aggregate expenditure on wholesale chains going from j to i, which is simply the sum of expenditure on chains from j to i which feature wholesaler-to-wholesaler sales, E_{ijww}, and those which feature wholesaler-to-retailer sales, E_{ijwr}, i.e. that

$$E_{ijw} = E_{ijww} + E_{ijwr}$$

We start with wholesale-to-wholesale chains. We start with an expression for the profits from sourcing a chain from j going through i, ending up in ι from the penultimate destination φ. Furthermore, we will consider a chain z_{ww} with w_u upstream steps and w_d downstream steps until location k (so that there are w_u+w_d+2 total steps: w_u steps from the origin to j, 1 step from j to i, w_d from i to φ and 1 step from φ to ι). Following the first subsection, the price of this chain will be

$$p_{z_{ww}} = \left(\tau_{o,z(2)} \prod_{n=2}^{w_u-1} \tau_{z(n)z(n+1)}\right) \tau_{ij} \left(\prod_{n=w_u+1}^{w_u+w_d} \tau_{z(n)z(n+1)}\right) \tau_{\iota\varphi} \cdot M_{w_u+w_d+2} \cdot p_o$$

and using the approximation to markups

$$p_{z_{ww}} \approx \left[(Bm_r)^{w_u} \tau_{o,z(2)} \prod_{n=2}^{w_u-1} \tau_{z(n)z(n+1)}\right] \cdot [Bm_r \tau_{ij}] \cdot \left[(Bm_r)^{w_d} \left(\prod_{n=w_u+1}^{w_u+w_d} \tau_{z(n)z(n+1)}\right) \tau_{\iota\varphi} m_r\right] p_o$$

where the term in the first set of brackets reflects trade costs and markups which occur upstream, the term in the middle brackets reflects markups and trade costs from j to i, and the term in the right brackets reflects trade costs and markups which occur downstream of the link from j to i.

Using this expression, we follow the expression in the first subsection to capture the expenditure a wholesale-to-wholesale chain for the link from j to i, denoted $E_{z_{ww}}$.

$$E_{z_{ww}} = \frac{p_{z_{ww}}^{\frac{\beta+1}{\beta}(1-\sigma_r)}}{m_r\,(Bm_r)^{w_d} \prod_{n=w_u+1}^{w_u+w_d} \tau_{z(n)z(n+1)}} \cdot N_{r\iota} \chi_{r\iota\varphi} \left(\frac{1}{\phi_{\iota\varphi} P_{r,\iota}}\right)^{1-\sigma_r} E_\iota$$

Second, we consider wholesale-to-retail chains. We will think of such a chain from j going through i, ending up in l. We will denote this chain z_{wr} with w_u upstream steps and of course only 1 downstream step to l (so that there are w_u+2 total steps: w_u steps from the origin to j, 1 step from j to i, 1 step from i to ι). Then, the price of this chain will be

$$p_{z_{wr}} = \left(\tau_{o,z(2)} \prod_{n=2}^{w_u-1} \tau_{z(n)z(n+1)}\right) \tau_{ij} \tau_{\iota i} \cdot M_{w_u+2} \cdot p_o$$

and using the approximation to markups

$$p_{z_{wr}} \approx \left[(Bm_r)^{w_u} \tau_{o,z(2)} \prod_{n=2}^{w_u-1} \tau_{z(n)z(n+1)}\right] \cdot [Bm_r \tau_{ij}] \cdot [\tau_{\iota i} m_r]\, p_o$$

where the term in the first set of brackets reflects trade costs and markups which occur upstream, the term in the middle brackets reflects markups and trade costs from j to i, and the term in the right brackets reflects trade costs and markups which occur downstream of the link from j to i.

Using this expression, we follow the expression in the first subsection to capture the expenditure a wholesale-to-retail chain for the link from j to i, denoted $E_{z_{wr}}$.

$$E_{z_{wr}} = \frac{p_{z_{wr}}^{\frac{\beta+1}{\beta}(1-\sigma_r)}}{m_r \tau_{\iota i}} N_{\iota i}^r \left(\frac{1}{\phi_{\iota i} P_{r,\iota}} \right)^{1-\sigma_r} E_\iota$$

With these expressions in hand, we first turn to wholesale-to-wholesale expenditure for $j \neq o$[72]

$$\begin{aligned} E_{ijww} &= \sum_{\varphi} \sum_{\iota} \sum_{w_u=1}^{\infty} \sum_{w_d=2}^{\infty} \sum_{z_{ww} \in \{ij\varphi\iota, w_u, w_d\}} E_{z_{ww}} \\ &= (B m_r \tau_{ij} p_o)^{\frac{\beta+1}{\beta}(1-\sigma_r)} \left((B m_r \tau_{jo})^{\frac{\beta+1}{\beta}(1-\sigma_r)} + \tilde{\mathbf{\Upsilon}}_{j'o} \left(\mathbf{I} - \tilde{\mathbf{\Upsilon}} \right)^{-1} \tilde{\mathbf{\Upsilon}}_{jj'}^T \right) \dots \\ &\dots \cdot \sum_{\varphi} \sum_{\iota} \left((B m_r \tau_{\varphi i})^{\frac{\beta+1}{\beta}(1-\sigma_r)} + \tilde{\delta}_{j'i} \left(\mathbf{I} - \tilde{\delta} \right)^{-1} \tilde{\delta}_{\iota j'}^T \right) (m_r \tau_{\varphi \iota})^{\frac{1-\sigma_r(\beta+1)}{\beta}} N_{\iota\varphi}^r \left(\frac{1}{\phi_{\iota\varphi} P_{r,\iota}} \right)^{1-\sigma_r} E_\iota \end{aligned}$$

[72] If $j = o$, it is necessary to add a 1 to the upstream expression for buying from the manufacturer, i.e. the first parenthetical would instead be $1 + (B m_r \tau_{jo})^{\frac{\beta+1}{\beta}(1-\sigma_r)} + \tilde{\mathbf{\Upsilon}}_{j'o} \left(\mathbf{I} - \tilde{\mathbf{\Upsilon}} \right)^{-1} \tilde{\mathbf{\Upsilon}}_{jj'}^T$; we have omitted this expression for brevity and because it is analogous to our expressions for P_{ij} in Section 5.

where

$$\tilde{\mathbf{\Upsilon}}_{j'o} \equiv \begin{pmatrix} (Bm_r\tau_{1o})^{\frac{\beta+1}{\beta}(1-\sigma_r)} & \dots & (Bm_r\tau_{Lo})^{\frac{\beta+1}{\beta}(1-\sigma_r)} \end{pmatrix}$$

$$\tilde{\mathbf{\Upsilon}}^T_{jj'} \equiv \begin{pmatrix} (Bm_r\tau_{j1})^{\frac{\beta+1}{\beta}(1-\sigma_r)} \\ \vdots \\ (Bm_r\tau_{jL})^{\frac{\beta+1}{\beta}(1-\sigma_r)} \end{pmatrix}$$

$$\tilde{\mathbf{\Upsilon}} \equiv \begin{bmatrix} (Bm_r\tau_{11})^{\frac{\beta+1}{\beta}(1-\sigma_r)} & \dots & (Bm_r\tau_{L1})^{\frac{\beta+1}{\beta}(1-\sigma_r)} \\ \vdots & \ddots & \vdots \\ (Bm_r\tau_{1L})^{\frac{\beta+1}{\beta}(1-\sigma_r)} & \dots & (Bm_r\tau_{LL})^{\frac{\beta+1}{\beta}(1-\sigma_r)} \end{bmatrix}$$

$$\tilde{\delta}_{j'i} \equiv \begin{pmatrix} (Bm_r\tau_{1i})^{\frac{1-\sigma_r(\beta+1)}{\beta}} & \dots & (Bm_r\tau_{Li})^{\frac{1-\sigma_r(\beta+1)}{\beta}} \end{pmatrix}$$

$$\tilde{\delta}^T_{\iota j'} \equiv \begin{pmatrix} (Bm_r\tau_{\iota 1})^{\frac{1-\sigma_r(\beta+1)}{\beta}} \\ \vdots \\ (Bm_r\tau_{\iota L})^{\frac{1-\sigma_r(\beta+1)}{\beta}} \end{pmatrix}$$

$$\tilde{\delta} \equiv \begin{bmatrix} (Bm_r\tau_{11})^{\frac{1-\sigma_r(\beta+1)}{\beta}} & \dots & (Bm_r\tau_{L1})^{\frac{1-\sigma_r(\beta+1)}{\beta}} \\ \vdots & \ddots & \vdots \\ (Bm_r\tau_{1L})^{\frac{1-\sigma_r(\beta+1)}{\beta}} & \dots & (Bm_r\tau_{LL})^{\frac{1-\sigma_r(\beta+1)}{\beta}} \end{bmatrix}$$

Similarly, for sales to retailers for $j \neq o$[73]

$$\begin{aligned} E_{ijwr} &= \sum_{\iota} \sum_{w_u=1}^{\infty} \sum_{z_{ww} \in \{ij\varphi\iota, w_u, w_d\}} E_{z_{wr}} \\ &= (Bm_r\tau_{ij}p_o)^{\frac{\beta+1}{\beta}(1-\sigma_r)} \left((Bm_r\tau_{jo})^{\frac{\beta+1}{\beta}(1-\sigma_r)} + \tilde{\mathbf{\Upsilon}}_{j'o} \left(\mathbf{I} - \tilde{\Upsilon} \right)^{-1} \tilde{\mathbf{\Upsilon}}^T_{jj'} \right) \dots \\ &\dots \cdot \sum_{\iota} (m_r\tau_{i\iota})^{\frac{1-\sigma_r(\beta+1)}{\beta}} \left(\frac{N_{rij}}{m_r} \left(\frac{1}{\phi_{\iota i} P_{r,\iota}} \right)^{1-\sigma_r} \right) E_\iota \end{aligned}$$

where $\tilde{\Upsilon}$,$\tilde{\mathbf{\Upsilon}}_{j'o}$, and $\tilde{\mathbf{\Upsilon}}^T_{jj'}$ are as defined above.

Thus, using the prior expressions, we can solve for the share of traders on any given route, and hence expected profits from sourcing from j for wholesalers in i as a function of fundamentals and the equilibrium number of wholesalers serving the j to i route.

[73] If $j = o$, it is necessary to add a 1 to the upstream expression for buying from the manufacturer as in the wholesale-wholesale expression, i.e. the first parenthetical would instead be $1 + (Bm_r\tau_{jo})^{\frac{\beta+1}{\beta}(1-\sigma_r)} + \tilde{\mathbf{\Upsilon}}_{j'o} \left(\mathbf{I} - \tilde{\mathbf{\Upsilon}} \right)^{-1} \tilde{\mathbf{\Upsilon}}^T_{jj'}$.

C.7.4 Equilibrium sourcing

We now return to find expected sourcing shares and entry. In this subsection, we show that both of these objects are unique. Thus we can in practice construct the necessary matrices following subsection 3 to find wholesale expenditure flows between every location pair, and then solve numerically to find the unique measure of wholesalers and the sourcing pattern. Throughout this subsection, we will refer to E_{ij}^{w} without using the complicated expressions derived in the prior subsection. Note that the aggregate wholesale expenditure on routes from j to i is exogenous with respect to wholesale sourcing locations and entry – it depends only on parameters and retail sourcing and entry decisions.

The expected wholesale profits (inclusive of fixed costs) of sourcing from j for i, denoted π_{ijw}, are given by

$$\pi_{ijw} = \pi_{ijw}^{var} - F_{ij}$$

(where we have an expression for π_{ijw}^{var} from the first subsection) and given the Gumbel distribution of fixed costs and the assumptions on the location parameters, expected profits (pre-entry) will be

$$\pi_{i}^{w} = s \ln \left(\sum_{j} \exp\left(\frac{\pi_{ijw}}{s}\right) \right) - f_{e,i}$$

By combining this expression with results from prior subsections, we can use the expression for variable profits to write

$$\chi_{wij} = \exp\left(\frac{\frac{1}{\sigma_w} \cdot \frac{E_{ijw}}{\chi_{wij} N_{wi}} - F_{ij} - f_{i,e}}{s} \right)$$

and

$$\pi_{iw} = s \ln \left(\sum_{j} \exp\left(\frac{\frac{1}{\sigma_w} \cdot \frac{E_{ijw}}{\chi_{wij} N_{wi}} - F_{ij}}{s} \right) \right) - f_{i,e}$$

where N_{wi} is the total measure of wholesalers in i.

Proposition A1: The measure of wholesalers and wholesale sourcing shares (for each location) are unique functions of fundamentals.

Proof: First, we show that, conditional on N_{wi}, the share of wholesalers which source from every location j is unique. And second, we show that this implies a unique measure of wholesalers which enter in location i.

We start with the expression for wholesale sourcing shares above

$$\chi_{wij} = \exp\left(\frac{\frac{1}{\sigma_w} \cdot \frac{E_{ijw}}{\chi_{wij} N_{wi}} - F_{ij} - f_{i,e}}{s}\right)$$

and we show that $\frac{\partial \chi_{wij}}{\partial N_{wi}} < 0$ for every value of N_{wi}. This implies that χ_{wij} is a unique function of N_{wi}:

$$\begin{aligned}
\frac{\partial \chi_{wij}}{\partial N_{wi}} &= \chi_{wij} \left(\frac{-\frac{1}{\sigma_w} \cdot \frac{E_{ijw}}{\chi_{wij} N_{wi}}}{s}\right) \left(\frac{1}{\chi_{ijw}} \frac{\partial \chi_{wij}}{\partial N_{wi}} + \frac{1}{N_{wi}}\right) \\
&= \frac{-\frac{E_{wij}}{N_{wi}}}{\sigma_w s N_{wi} + \frac{E_{wij}}{\chi_{wij}}} \\
&< 0
\end{aligned}$$

And second,

$$\begin{aligned}
\frac{\partial \pi_{wi}}{\partial N_{wi}} &= -s \frac{E_{wij}}{N_{wi}} \sum_j \left(\frac{\chi_{wij}}{\sigma_w s N_{wi} \chi_{wij}^2 + E_{ijw} \chi_{wij}}\right) \\
&< 0
\end{aligned}$$

This implies that there is only one value of N_{wi} under which the free entry condition holds. Thus equilibrium is unique.

C.7.5 Wholesaling chain length

The last part of this subsection is to describe wholesaling chain lengths. Once we describe the expenditure flowing on chains of a particular length, then the remaining steps are trivial.

The result will have the form of a 4-touple: wholesale flows through j to i with ν upstream steps and υ downstream steps, denoted $E_{ij\nu\upsilon w}$. For the purposes of calculating this expression, we develop two scalars: $M_{uj\nu}$ which captures upstream costs, and $M_{di\varphi\upsilon}$.

If $\nu = 1$, then $M_{uo1} = 1$ (note that the only j with 1 upstream step is the origin, as manufacturer sourcing is not possible from other places).

If $\nu = 2$, then $M_{uj2} = (B m_r \tau_{jo})^{\frac{\beta+1}{\beta}(1-\sigma_r)}$.

If $\nu \geq 3$, then $M_{uj\nu} = \tilde{\boldsymbol{\Upsilon}}_{j'o}\left(\tilde{\Upsilon}\right)^{\nu-3}\tilde{\boldsymbol{\Upsilon}}^{T}_{jj'}$ (where $\tilde{\boldsymbol{\Upsilon}}_{j'o}$, $\tilde{\Upsilon}$, and $\tilde{\boldsymbol{\Upsilon}}^{T}_{jj'}$ are defined as in the prior subsection).

If $\upsilon = 2$, then $M_{di\varphi 2} = (Bm_r\tau_{\varphi i})^{\frac{1-\sigma_r(\beta+1)}{\beta}}$ (since this is a wholesale transaction, it is impossible for $\upsilon = 1$).

If $\upsilon \geq 3$, then $M_{di\varphi 3} = \tilde{\delta}_{j'i}\tilde{\delta}^{\nu-3}\tilde{\delta}^{T}_{\varphi j'}$. (again where the vectors and matrices are defined as in the prior subsection).

With these definitions in hand, we can then describe

$$E_{ij\nu\upsilon w} = (Bm_r\tau_{ij}p_o)^{\frac{\beta+1}{\beta}(1-\sigma_r)}M_{uj\nu}\sum_{\varphi}\sum_{\iota}M_{di\varphi\upsilon}\left(m_r\tau_{\varphi\iota}\right)^{\frac{1-\sigma_r(\beta+1)}{\beta}}N^{r}_{\iota\varphi}\left(\frac{1}{\phi_{\iota\varphi}P_{r,\iota}}\right)^{1-\sigma_r}E_{\iota}$$

D Material from Section 5

D.1 Markups approximation

We start with the expression for the markup at step l and find

$$\frac{\sigma_l}{\sigma_l - 1} = m_r\left[1 - \frac{1}{\sigma_r}\left(1 - \left(\frac{\mu_w + 1}{\mu_w}\right)^{1-n}\right)\right]$$

and then substitute into the expression for the log of the aggregate markup

$$\ln M_L = L\ln m_r + \sum_{l=1}^{L}\ln\left(1 - \frac{1}{\sigma_r}\left(1 - \left(\frac{\mu_w + 1}{\mu_w}\right)^{1-l}\right)\right)$$

We next take a second order approximation to $\ln M_L$ around $L = 2$ and $\frac{\mu_w+1}{\mu_w} = 1$ (i.e. as μ_w gets very large) to obtain

$$\ln M_L \approx L\ln m_r - \frac{1}{\sigma_r}\left(\frac{1}{\mu_w}\right)(L-1) + \frac{1}{2}\left(\frac{1}{\mu_w}\right)^2\left[\frac{m_r + 1}{\sigma_r m_r}\right]$$

By exponentiating both sides we obtain

$$M_L = \exp\left(\frac{1}{2}\left(\frac{1}{\mu_w}\right)^2\left[\frac{m_r+1}{\sigma_r m_r}\right]\right)m_r\left(m_r e^{-\frac{1}{\mu_w\sigma_r}}\right)^{L-1}$$

which implies a wholesale markup between any two steps L and $L - 1$ (for $L > 0$)

$$m_w = \frac{M_L}{M_{L-1}} = m_r e^{-\frac{1}{\mu_w\sigma_r}}$$

In general in the text we will treat $M_L \approx m_r \left(m_r e^{-\frac{1}{\mu_w \sigma_r}}\right)^{L-1}$. We do this for two reasons. First, it avoids needlessly distorting the retail markup (which would otherwise take place as the approximation implies a markup at $L = 1$ of $\exp\left(\frac{1}{2}\left(\frac{1}{\mu_w}\right)^2 \left[\frac{m_r+1}{\sigma_r m_r}\right]\right) m_r$). And second, in practice $\exp\left(\frac{1}{2}\left(\frac{1}{\mu_w}\right)^2 \left[\frac{m_r+1}{\sigma_r m_r}\right]\right)$ is quite close to 1. Note that omitting this factor does not affect our estimation of the wholesale markup, and under our parameter values,

$$\exp\left(\frac{1}{2}\left(\frac{1}{\mu_w}\right)^2 \left[\frac{m_r+1}{\sigma_r m_r}\right]\right) \approx 1.037$$

so omitting this factor is of little practical importance.

D.2 Price index approximation

Using our markup approximation (and defining $B \equiv m_r e^{-\frac{1}{\mu_w \sigma_r}}$), we are able to obtain closed-form solutions our expressions for the $\mathcal{P}_j$. Starting from the expression for P_j in the text for a non-origin source

$$\mathcal{P}_j^{1-\sigma_r} \approx \frac{1}{\phi_j^{1-\sigma_r}} \left[(B m_r \tau_{jo} p_o)^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)} + \sum_{j'} \left((B m_r)^2 \tau_{j'o} \tau_{jj'} p_o \right)^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)} + \sum_{j''} \sum_{j'} \left((B m_r)^3 \tau_{j''o} \tau_{j'j''} \tau_{j\ldots} \right.\right.$$

Then,

$$\begin{aligned}
\mathcal{P}_{j \neq o}^{1-\sigma_c} &\approx \frac{p_o^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)}}{\phi_j^{1-\sigma_r}} \left[(B m_r \tau_{jo})^{\frac{1}{\beta_r}(1-\sigma_r)} + \mathbf{T}_{jo} \left(\sum_{n=0}^{\infty} \mathbf{T}^n \right) \mathbf{T}_{jj'}^T \right] \\
&= \frac{p_o^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)}}{\phi_o^{1-\sigma_r}} \left[(B m_c \tau_{jo})^{\frac{1}{\beta_r}(1-\sigma_r)} + \mathbf{T}_{jo} \left(\mathbf{I} - \mathbf{T} \right) \mathbf{T}_{jj'}^T \right]
\end{aligned}$$

where we define

$$\mathbf{T} \equiv \begin{pmatrix} (B m_r \tau_{11})^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)} & \cdots & (B m_r \tau_{L1})^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)} \\ \vdots & \ddots & \vdots \\ (B m_r \tau_{1L})^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)} & \cdots & (B m_r \tau_{LL})^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)} \end{pmatrix}$$

and we use $\mathbf{T}_{jo}$ to denote the first row of $\mathbf{T}$ and $\mathbf{T}_{jj'}^T$ to denote the transpose of the jth row of $\mathbf{T}$.

Essentially identical steps can be taken to obtain

$$\mathcal{P}_o = \frac{p_o^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)}}{\phi_o^{1-\sigma_r}} \left[1 + (Bm_c\tau_{oo})^{\frac{1}{\beta_r}(1-\sigma_r)} + \mathbf{T}_{jo}\left(\mathbf{I}-\mathbf{T}\right)^{-1}\mathbf{T}^T_{oj'}\right]$$

$$(j \neq o)\ \phi_j^{1-\sigma_r} = p_o^{\frac{1}{\beta_r}(1-\sigma_r)}\left[(Bm_r\tau_{jo})^{\frac{1}{\beta_r}(1-\sigma_r)} + \hat{\mathbf{T}}_{j'o}\left(\mathbf{I}-\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}^T_{jj'}\right]$$

$$\phi_o^{1-\sigma_r} = p_o^{\frac{1}{\beta_r}(1-\sigma_r)}\left[1 + (Bm_r\tau_{oo})^{\frac{1}{\beta_r}(1-\sigma_r)} + \hat{\mathbf{T}}_{j'o}\left(\mathbf{I}-\hat{\mathbf{T}}\right)^{-1}\hat{\mathbf{T}}^T_{oj'}\right]$$

where we define

$$\hat{\mathbf{T}} \equiv \begin{pmatrix} (Bm_r\tau_{11})^{\frac{1}{\beta_r}(1-\sigma_r)} & \cdots & (Bm_r\tau_{L1})^{\frac{1}{\beta_r}(1-\sigma_r)} \\ \vdots & \ddots & \vdots \\ (Bm_r\tau_{1L})^{\frac{1}{\beta_r}(1-\sigma_r)} & \cdots & (Bm_r\tau_{LL})^{\frac{1}{\beta_r}(1-\sigma_r)} \end{pmatrix}$$

and we use $\hat{\mathbf{T}}_{jo}$ to denote the first row of $\hat{\mathbf{T}}$ and $\hat{\mathbf{T}}^T_{jj'}$ to denote the transpose of the jth row of $\mathbf{T}$.

Similarly, the expressions for expenditure-weighted chain length in Section 4.3.4 can be approximated as

$$(j \neq o)\ \Lambda_j = \frac{p_o^{\frac{1}{\beta_r}(1-\sigma_r)}}{\left(\tilde{P}_j\phi_j\right)^{1-\sigma_r}}\left[2\left(Bm_r\tau_{jo}\right)^{\left(1+\frac{1}{\beta_r}\right)(1-\sigma_r)} + \mathbf{T}_{jo}\left(\sum_{n=0}^{\infty}(n+3)\,\mathbf{T}^n\right)\mathbf{T}^T_{jj'}\right]$$

$$\Lambda_o = \frac{p_o^{\frac{1}{\beta_r}(1-\sigma_r)}}{\left(\tilde{P}_o\phi_o\right)^{1-\sigma_r}}\left[1 + \mathbf{T}_{jo}\left(\sum_{n=0}^{\infty}(n+2)\,\mathbf{T}^n\right)\mathbf{T}^T_{oj'}\right]$$

where $\mathbf{T}$, $\mathbf{T}_{jo}$, $\mathbf{T}^T_{jj'}$, $\mathcal{P}_j$ and ϕ_j are as previously defined.