Spatial Inequality and Informality in Kenya's Firm Network *

Verena Wiedemann† Benard Kipyegon Kirui‡ Vatsal Khandelwal§
Peter Wankuru Chacha¶

February 2025

Click here for the most recent version.

Abstract

The spatial configuration of domestic supply chains plays a crucial role in the transmission of shocks. This paper investigates the representativeness of formal firm-to-firm trade data in capturing domestic trade patterns in Kenya — a context with a high prevalence of informality. We first document a series of stylized facts to show that formal sector data is not representative of overall economic activity. We then link granular transaction-level data on formal firms with data on informal economic activity to estimate a structural model and predict a revised network that accounts for informal firms. We find that formal sector data overstates the spatial concentration of aggregate trade flows and underaccounts for trade within regions and across regions with stronger social ties. Additionally, the higher the incidence of informality in a sector and region, the more we underestimate its vulnerability to domestic output shocks and overestimate its vulnerability to import shocks.

Keywords: informality, supply chains, spatial inequality, firm networks.
JEL classification: D22, D85, E26, O11, O17, R12.

*We thank the Kenya Revenue Authority (KRA) for the outstanding collaboration. Romeo Ekirapa, Simon Mwangi, and Benard Sang provided excellent technical support and advice. We thank Elizabeth Gatwiri and Daniela Villacreces Villacis for excellent research assistance. We thank Andrea Bacilieri, Raphael Bradenbrink, Banu Demir Pakel, Kevin Donovan, Douglas Gollin, Justice Tei Mensah, Luke Heath Milsom, Sanghamitra Warrier Mukherjee, Solomon Owusu, Piyush Panigrahi, Nina Pavcnik, Simon Quinn, Gabriel Ulyssea, Alexander Teytelboym, Christopher Woodruff, Hannah Zillessen and participants of seminars and conferences at Oxford, CSAE, STEG, the World Bank's DaTax group, IFC, KIPPRA, Montreal, the Bank of Italy/OECD, and Jeune Street for comments and feedback. We gratefully acknowledge financial support from the Private Enterprise Development in Low Income Countries (PEDL) and the Structural Transformation and Economic Growth (STEG) programmes, both of which are joint initiatives by the Centre for Economic Policy Research (CEPR) and the UK Foreign, Commonwealth & Development Office (FCDO). Verena Wiedemann further acknowledges funding from the Oxford Economic Papers Research Fund (OEP) and the German National Academic Scholarship Foundation. This study has been approved by the Department of Economics Research Ethics Committee at Oxford (protocol no.ECONCIA20-21-23), and the Kenyan National Commission for Science, Technology and Innovation (protocol no.NACOSTI/P/20/5923). The views in this paper are those of the authors, and do not necessarily represent those of the KRA or any other institution the authors are affiliated with.

†Economic Research Unit, International Finance Corporation (World Bank Group)
‡Privatization Commission Kenya
§Department of Economics, University of Exeter
¶International Monetary Fund

1 Introduction

Limited opportunities for export-led growth and concerns over the unequal allocation of gains from trade have led policymakers and researchers to focus on domestic supply chains and market integration to enhance economic development (Topalova, 2010; Atkin and Donaldson, 2015; Bustos et al., 2020; Grant and Startz, 2022; Goldberg and Reed, 2023). Progress in understanding the structure of domestic supply chains is facilitated by the increasing availability of granular transaction-level firm network data, which are often sourced from tax records (see e.g. Panigrahi, 2022; Fujiy et al., 2022; Adão et al., 2022; Alfaro-Ureña et al., 2022; Boken et al., 2023).[1] These advancements reflect a broader trend in the literature on development and structural transformation, where novel micro-data allow researchers to generate new insights to answer classic economic questions (Lagakos and Shu, 2023).

While some of the most exciting insights stem from non-traditional data sources, like credit registries (Bustos et al., 2020), smartphone data (Blanchard et al., 2021; Kreindler and Miyauchi, 2023), matched employer-employee data (Dix-Carneiro et al., 2024), and transaction-level firm data, the underlying data-generating process of such data is often skewed towards particular segments of the economy, e.g. taxpayers. This can leave us with a limited view due to the size and significance of the informal sector in many economies.[2]

In this paper, we ask how observing only a selected segment of the economy, in our case formal firms, might bias what we can learn about the patterns of firm-to-firm trade from tax records.[3] Given the scarcity of data on the informal sector, empirical evidence on the degree of bias caused by ignoring informality is typically rather limited. We address this problem by combining transaction-level administrative tax records of over 76,000 formal firms in Kenya with data on informal sector activity obtained from population census data and national accounts.

We first establish a series of stylized facts on both the formal and the informal sector to show that formal sector data is not representative of overall economic activity. We further employ a structural model to predict a revised firm network that accounts for informality. We find that extrapolating from formal sector data leads researchers to mismeasure the firm network and consequently mis-predict key economic indicators such as the degree of spatial inequality in trade flows, the importance of urban hubs, and the regional impact of economic shocks. Finally,

[1] Transaction-level survey data (see e.g. Startz, 2021) or administrative industry-specific data (see e.g. Hansman et al., 2020), which often cover a wider range of firm and buyer-supplier characteristics, are a popular, albeit sometimes costly, complement to data relying on tax records.

[2] For example, the informal sector has been documented to play a crucial role in adjustment to trade shocks (McCaig and Pavcnik, 2018; Dix-Carneiro and Kovak, 2019; Dix-Carneiro et al., 2024).

[3] One of the key advantages of transaction-level firm-to-firm trade data over traditional sources like input-output tables is the ability to explore the rich regional heterogeneity in economic activity, rather than being limited to national aggregates.

we implement bounding exercises to test the sensitivity of our results to alternative assumptions about how informal firms interact with the rest of the economy. Our results indicate that incorporating data on informality alongside datasets on formal firm-to-firm trade can improve the accuracy of economic predictions, leading to better-informed policy decisions.

The Kenyan context is particularly well-suited to answering this question. With VAT-paying formal firms contributing only 36% of Kenya's GDP, the informal sector constitutes a sizable segment of the economy. Moreover, as East Africa's largest economy, Kenya boasts a domestic market with substantial geographic and socio-economic regional heterogeneity. While questions about who benefits from globalization and the relevance of domestic and regional markets for future growth are particularly crucial for emerging economies in Africa (Atkin and Donaldson, 2015; Goldberg and Reed, 2023), research on these topics using data with national coverage remains sparse for the region. A series of unusually granular data sets allow us to observe both formal and informal economic activity at the sectoral and regional levels—an advantage that is often difficult to achieve in contexts of similar income levels, where statistical bureaus tend to face resource constraints and are frequently limited to focusing on national aggregates.[4]

We begin by documenting three stylized facts on formal and informal economic activity to assess the representativeness of the VAT data. We ask whether the observed patterns of domestic trade in the formal sector arise due to the systematic selection of firms into the administrative data or if they reflect the underlying structure of the economy. First, we show that the VAT-paying sector only contributes to 36% of Kenya's GDP, accounts for 2.5% of licensed businesses in the country, and employs just 5% of the total workforce. Next, we document that trade among formal firms is substantially more concentrated around Kenya's metropolitan areas than both population and overall economic activity. This concentration is driven by inequality along the extensive margins of the firm network, i.e. the location of firms and trading relationships, rather than transaction volumes. Finally, we document that informality is not evenly distributed across space but varies systematically across sectors, regions, and firm position in the supply chain. For instance, informal firms are more likely to be located downstream of large formal firms, and informality negatively correlates with regional economic size and income. As a result, we expect that accounting for the informal sector can systematically alter the structure of the observed firm network.

Next, we introduce and estimate a network formation model with heterogeneous node types following Bramoullé et al. (2012) to show how accounting for informal firms could affect the observed network and predict a revised network. In our adaptation of the model, we classify firms based on their sector, location, and size, due to the substantial heterogeneity along these

[4] https://blogs.worldbank.org/africacan/for-the-first-time-the-relative-economic-size-of-kenyas-counties-is-clear

three dimensions documented in the stylized facts. As a result, the model provides predictions for the number of links between firms of different sector-location-size types. The network formation process is as follows: A newborn firm first chooses a specific type of firm to link to in accordance with its own "bias". This bias can be reflective of the firm's underlying production technology or geographic location. Then, it forms a specific proportion of its links with firms of this type via undirected search and the remainder via preferential search. In other words, the new firm chooses a certain proportion of suppliers independent of its network environment (undirected), but the remainder from the pool of the suppliers of these suppliers (preferential).[5] We first estimate this network formation model to predict the Kenyan firm network as it is.[6] We find that new firms choose 45% of their suppliers through undirected search, conditional on their bias, and the remaining 55% of suppliers are found via existing suppliers.[7]

We then predict a revised network that accounts for informal firms by combining the model with granular real-world data on the sectoral and regional composition of the informal sector. To incorporate informal firms, we use updated information on the sectoral and spatial dispersion of informal economic activity from the population census and a survey of small firms by the Kenya National Bureau of Statistics. Using this information, we update the probability of firms being born in a given sector, location, and of a certain size. We rely on the assumption that informal firms, conditional on sector and geography, are similar in terms of their linking patterns to the smallest quartile of formal firms observed in our data. By using small formal firms' behaviour as an initial proxy for informal firm linking patterns, we address concerns that informal firms might have different linking patterns compared to larger formal firms operating in the same sector and location (e.g. due to internal economies of scale (Grant and Startz, 2022)). We provide evidence showing that the linking patterns of small firms are similar to what we would expect from informal firms – they link more locally and buy more from intermediaries relative to their larger peers. However, informal firms might encounter additional obstacles specific to informality.[8] Hence, we implement additional sensitivity checks that account for alternative scenarios with lower linking probabilities between the formal and informal sector.

We use the revised network to answer the question of interest: How do spatial patterns of trade

[5]Our modeling choices, such as focusing on the extensive margin of trade (i.e. trading relationships rather than volumes) and the number of buyers rather than suppliers, are motivated by our stylized facts and other regularities commonly observed in firm network data (Bacilieri et al., 2023).

[6]The model predicted network not only fits the empirical distribution of outlinks well, but further performs well with respect to untargeted moments, such as capturing the prominence of Nairobi and Mombasa in the production network.

[7]In comparison, Chaney (2014) finds that only 40% of all relationships of French exporters with international trade partners are formed via preferential attachment. Our estimate of 55% of links being formed as a result of preferential attachment could suggest that information frictions are potentially even more binding for firms in Kenya's domestic firm network.

[8]These can include wedges introduced by the VAT system itself (De Paula and Scheinkman, 2010; Gadenne et al., 2022).

change when informal firms are accounted for? First, we find that sectors and regions with the highest levels of informality have more outlinks in the revised network relative to the baseline network. The spatial inequality in outlinks declines by 7% and the prominence of urban hubs like Nairobi and Mombasa declines. We show that while this decrease in inequality of outdegrees is driven by an increase in both inter-county and intra-county trade, intra-county trade rises by a larger margin. Moreover, once informal firms are accounted for, the number of trade relationships between counties is more sensitive to the strength of social ties between them. In line with both the enhanced prominence of intra-county trade and trade among counties with stronger social ties, we find that the predicted network is more partitioned in that it has more clusters with links among them rather than links across.

Next, we simulate the pass-through of domestic and import shocks to sector-regions using both the network estimated by the structural model and the revised network obtained after accounting for informal firms. When relying on the revised network that includes informal firms, we find a larger adverse impact of domestic output shocks on sector-regions with a higher level of informality relative to the case where we rely on formal sector data only. Our results suggest that a 1 percentage point decrease in the formal sector share results, on average, in an underestimation of the reduction in output due to a domestic shock by 5 percentage points. Conversely, when considering the pass-through of an import shock, we find that relying solely on the formal network to study its impact on aggregate output introduces a bias in the opposite direction. The economy is less exposed to import shocks than predicted if informal firms are accounted for. This discrepancy arises because import shocks primarily affect larger formal firms, which carry less weight in the overall firm network once we incorporate informality.

Finally, we consider the sensitivity of our results with respect to alternative assumptions about the linking patterns of informal firms. We conduct bounding exercises where we further restrict the degree of integration of informal firms with the formal sector. Drawing on survey evidence and the existing literature (Böhme and Thiele, 2014; Gadenne et al., 2022), we assume informal firms sell no output to the formal sector and source a smaller proportion of their inputs from formal firms compared to small formal firms. We find that this further reduces spatial inequality in outlinks and diminishes the prominence of urban hubs relative to the formal network. Moreover, we continue to significantly underestimate the impact of domestic shocks while overestimating the effects of trade shocks.

Our paper contributes to the literature on macroeconomic development, informality, firm networks, and spatial inequality. First, we contribute to a growing body of research at the intersection of trade and macroeconomic development that integrates granular administrative data such as employer-employee records and data from credit registries, with broader data sources

like population censuses to achieve a more accurate assessment of aggregate economic outcomes. To date, this literature has primarily focused on employment outcomes, sector shares (see e.g. Albert et al., 2021), and consumption (see e.g. Fan et al., 2023), where informal activity is somewhat more observable. However, informal activity along supply chains remains particularly elusive (Böhme and Thiele, 2014; Atkin and Khandelwal, 2020). Our results highlight the implications of the non-random selection of firms into administrative records. This is particularly important, given the growing reliance on such data in the literature.

Our approach to employ a structural model to bridge gaps in our understanding of informal firm dynamics also aligns with the recent literature in this field (see e.g. Ulyssea, 2018; Dix-Carneiro et al., 2024). Unlike related studies that focus on firm and worker-level dynamics, we do not model the endogenous response of firms and workers to simulated shocks. Crucially, however, our research design allows us to examine the role of informality for Kenya's region-level input-output matrix.[9] This is particularly relevant for research that seeks to complement predictions about aggregate national welfare with welfare estimates at the regional level to study geographic heterogeneity in the impact of international trade (Topalova, 2010), infrastructure investments (Arkolakis et al., 2023; Demir et al., 2024) or climate and weather shocks (Albert et al., 2021; Castro-Vincenzi et al., 2024).

Second, we contribute to the literature on spatial production networks (Bernard et al., 2019; Panigrahi, 2022; Miyauchi, 2024; Arkolakis et al., 2023), shock propagation in firm networks (see e.g. Baqaee, 2018; Huneeus, 2018; Carvalho et al., 2021), and urban primacy (as published in Jefferson (1989), Jefferson, 1939; Memon, 1976; Ades and Glaeser, 1995; Soo, 2005). We analyze the spatial distribution of formal firms in an economy with a large informal sector and demonstrate that ignoring informality can lead to overestimating spatial inequality in firm-to-firm trade and the extent of urban primacy. This oversight may cause researchers to underestimate the economic connectedness and vulnerability of smaller regions.

Third, we contribute to a sizeable literature on estimating the size of the informal sector (Schneider and Enste, 2000; La Porta and Shleifer, 2014; Elgin et al., 2021). This literature uses cross-country regressions to show that the relative size of the formal economy increases with income levels (Brandt, 2011; La Porta and Shleifer, 2014; Ulyssea, 2018). We confirm that this pattern extends to Kenya's domestic economy, demonstrating that the formal sector share correlates with income levels across regions within the country. Our finding that formal sector activity is concentrated in Kenya's metropolitan areas mirrors Zárate (2022)'s finding from spatial patterns within Mexico City, which exhibits a similar formal-core, informal-periphery

[9]Input-output channels and the links between a more formalised manufacturing sector and a more informal service economy are, for example, an important channel for how trade shocks feed through to informal firms in Brazil (Dix-Carneiro et al., 2024).

structure. Our findings also align with the literature on the link between the size of markets and the firm size distribution (Kumar et al., 1999; Laeven and Woodruff, 2007; Gollin, 2008; McCaig and Pavcnik, 2015). Additionally, we complement a literature in public finance that studies reasons for why informality arises along supply chains, and how tax policy can alter the incidence of informality (De Paula and Scheinkman, 2010; Zhou, 2022; Gadenne et al., 2022; Almunia et al., 2023). Relative to this literature, we focus on reconstructing a more complete network that includes informal firms rather than studying the marginal firm's decision to formalize.

Finally, we contribute to the growing literature on estimating network statistics and reconstructing networks in the presence of missing data. Our approach is similar in spirit to the graphical reconstruction technique proposed in Chandrasekhar (2016) who provide two methods to correct biases in network statistics from missing data. However, the data are missing in a non-random manner in our context i.e. the preferences for network formation of the missing nodes are likely to be systematically different from the nodes that we observe. We first account for this by estimating a structural model that allows for heterogeneous preferences for network formation among firms of different sizes, locations, and sectors. We then make systematic, data-driven, assumptions about the linking probabilities of the missing informal firms to construct a revised network. In complementary work, Bacilieri et al. (2023) examine how varying reporting thresholds for firm-to-firm transactions affect the comparability of aggregate network statistics. While they demonstrate that missing data can bias network statistics, our focus differs in two key ways. First, transaction reporting thresholds are not a concern in our setting; we instead address other sources of informality and hence network incompleteness in tax data. Second, while they analyze aggregate statistics, we examine distributional implications at the subnational level – a margin where we find informality plays a substantial role.

The paper is organized as follows: Section 2 describes our data. Section 3 examines how representative the trade patterns captured in the administrative data are of overall economic activity and discusses the role of the informal sector. We then estimate a network formation model with preferential attachment (Bramoullé et al., 2012) in Section 4. Section 5 describes the patterns we observe in our revised network that now incorporates informal firms, while Section 6 analyzes the sensitivity of these results to alternative assumptions. Section 7 concludes.

2 Data Description

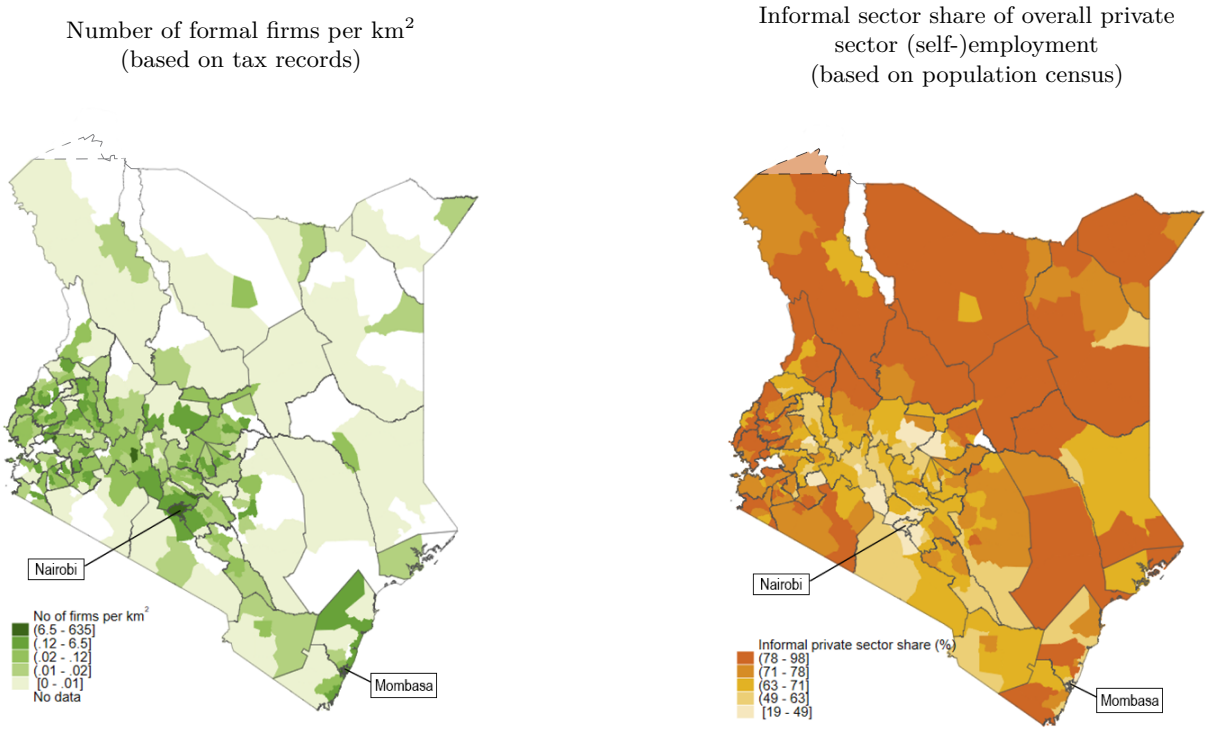
2.1 Administrative Data

Our analysis of the formal sector draws on micro data from value-added (VAT) and pay-as-you-earn (PAYE) tax returns collected by the Kenya Revenue Authority. The VAT returns include details on firm-to-firm transactions between VAT-registered firms. Sales and purchases

with non-registered parties (e.g. tax exempt parties, non-VAT-registered businesses, and final consumers) are recorded as an aggregate monthly figure.

VAT applies to firms with an annual turnover of KShs five million and above ($38,400 as of May 2024). Once a firm is VAT-registered and has crossed the threshold of KShs five million, they are required to continue filing VAT returns in years with lower turnover. We only include entities that identify as private companies or partnerships in their tax-registration form and exclude all government-owned firms, government agencies, international organisations, NGOs, trusts, and clubs. We also restrict our analysis to firms with annual purchases greater than zero and annual sales of KShs five million or more in at least one year between 2015 and 2022.[10] In addition to this, we utilize the tax registration forms to obtain basic self-reported information on each firm's 4-digit sector classification and headquarter location.

Figure 1: Location of formal firms and informal sector employment shares



The left map shows the density of firm headquarter locations at the sub-county level, i.e. the number of firms per km^2. The right map shows the share of informally employed people as a share of the local labour force, at the subcounty level. Sub-counties represent the second administrative layer. The borders of Kenya's 47 counties are outlined in grey.

Figure A1 plots the sector composition and the respective sales and purchase channels of firms covered in these administrative records. Manufacturing and wholesale and retail firms together account for almost half of the sales we observe in the tax records. The graph on the left in Figure 1 shows the geographical dispersion of formal firm headquarters.

[10]This ensures that firms that registered for VAT to bid for tender but were never operational are not included for analysis.

2.2 Data on Informal Activity

2.2.1 Margins of Informality

We first discuss how informality arises in firm networks constructed from VAT data. The most obvious reason is that firms are simply not registered for VAT, i.e. they are informal on the extensive margin. A second reason why VAT data falls short of covering overall economic activity is that even VAT-registered firms might not fully report all their transactions in VAT records. This is analogous to the intensive margin of informality introduced by the literature on labour markets (Ulyssea, 2018). For example, a wholesaler we interviewed in Nairobi's Central Business District explained how the notion of an extensive and intensive margin of informality extends to firm-to-firm transactions:

> *"All firms purchase from manufacturers and importers paying input VAT. They even have an interest in getting purchases that have VAT on it to inflate the input VAT. What they do to mitigate the VAT levy, [is that] they downplay their output VAT (i.e. sales). Some customers will purchase with receipt and output VAT on it. Some customers will purchase without a receipt."*

Table 1 summarises four different margins of informality that can occur in firm networks: an extensive margin at the firm level and an intensive margin at the transaction level. Within each category, informality can occur due to either non-compliance or simply because a firm/transaction is too small to be taxed.[11]

Table 1: Margins of informality in firm networks

	Extensive	**Intensive**
Below tax threshold	Small firms	Small transactions
Above tax threshold	Non-compliance	Non-compliance

The only margin of informality, which is not a concern in our setting, is the potential neglect of small VAT transactions. This is because the Kenya Revenue Authority requires firms to record transactions of any size, conditional on both parties being VAT-registered. Moreover, we address some issues around non-compliance by relying on information from a firm's trade partner to recover some omitted transactions and under-reported trade volumes. We now turn towards a discussion of the available measures of informality.

2.2.2 Measurement

Table 2 provides an overview of the data sources we draw on to capture informal economic activity, including the most granular level of disaggregation possible.

We use this data to obtain an updated distribution of overall (formal and informal) economic activity that we can compare to the administrative data. We mainly rely on measures that

[11] Depending on the tax code not all of these categories arise in every setting. Moreover, VAT exemptions can be a legal reason why firms or transactions above the VAT threshold are not captured in administrative tax records.

capture the proportion of overall economic activity that can be traced back to the formal sector, i.e. formal sector shares of total employment, value added, and number of firms. Table 3 summarises how we compute the various measures of informality. Importantly, we not only measure informality as the gap between overall economic activity and what is captured in the administrative data, but further rely on measures generated independently of the administrative records. This helps us rule out that our measures of informality capture idiosyncratic patterns that are specific to the VAT system.

Table 2: Overview of data sources

Source	**Year**	**Aggregation**	**Key indicators**
Population & housing census (census)	2019	sector AND county	formal & informal employment
Gross County Product (GCP)	2019	sector AND county	Gross County Product
Census of establishments (CoE)	2017	sector OR county	# of formal sector establishments
Micro, small & medium sized enterprises survey (MSMEs)	2016	firm-level	main input source and buyer
Census of industrial production	2010	sector AND county	sales of multi-establishment firms

All data are collected and published by the Kenya National Bureau of Statistics. **Sources:** 2019 Kenya Population & Housing Census KNBS (2019); Gross County Product KNBS (2022); Census of Establishments KNBS (2017); Small & Medium-Sized Enterprises Survey KNBS (2016); Census of Industrial Production 2010 (KNBS, 2010).

Table 3: Measures of informality

Unit	**Numerator** (formal sector)	**Denominator**	**Source**	**Use admin data**
Employment	No. formal priv. sector employ.	Working population	Census	✗
Employment	No. employ. in licensed firms	No. employ. in all firms	MSMEs	✗
Employment	No. employ. VAT firms	No. employ. in licensed firms	MSMEs	✓
No. firms	No. licensed firms	All firms	MSMEs	✗
No. firms	No. VAT firms	All firms	MSMEs	✓
Value added	Value added VAT firms	Gross County Product	GCP	✓

For details on the data sources by KNBS see Table 2. The term "all firms" refers to both licensed and unlicensed businesses based on KNBS (2016) estimates.

Throughout this paper, we draw on indicators for informal activity, computed from three types of sources: employment figures, the number of firms, and value added (i.e. the difference between sales and purchases). Our preferred measure of informality uses formal sector employment as a share of total employment, drawing on a comprehensive labor force module in the 2019 population census (KNBS, 2019). This measure, which later serves as a key input for predicting the revised network with informal firms, offers two distinct advantages. First, it enables simultaneous disaggregation of informal employment at both sectoral and regional levels. Second, it allows us to distinguish between private and public sector employment – a distinction unavailable in alternative measures. The graph on the right in Figure 1 shows the geographical dispersion of informal activity as per this measure. The measure correlates strongly (ρ= 0.83, Table A2) with another employment-based measure computed using the administrative records (see Figure A8). Measures of informality based on the number of firms rely on estimates of the

universe of businesses in KNBS (2016),[12] while the value-added measure utilizes estimates of the regional economic size captured by the Gross County Product (KNBS, 2022).[13]

Finally, we exclude agricultural and non-market service sectors from our informality measures where possible, as their tax records only cover a small and very specific sub-population of firms and employees. In the case of agricultural firms, the administrative data only capture large-scale commercial agriculture, which is often primarily export-oriented (see Figure A1 and Chacha et al. (2024)). Non-market services are dominated by non-profit organizations and the government with only a few for-profit VAT firms.

In addition to the above, we also use data from a survey with small and medium size enterprises (KNBS, 2016) to derive some insights into the sales and purchase patterns of the informal sector. The survey data only record the main type of buyer and supplier of a firm and hence cannot be directly used to reconstruct a network with informal firms. However, we will use these data to inform the assumptions of our model.

3 Representativeness of the Formal Firm Network

In this section, we establish three stylized facts about the representativeness of VAT data. First, VAT-registered firms account for a limited share of economic activity. Second, trade flows among formal firms show higher spatial concentration around urban centers than economic activity overall. We find that these patterns are not an idiosyncratic feature of the VAT data, but reflect spatial concentration in formal activity more broadly. Third, the incidence of informality varies systematically across space, sectors, and supply chain position.

Fact 1: The VAT-paying sector only accounts for 36% of Kenya's GDP.

We compare value added observed in the administrative records to GDP records from national accounts to comment on the discrepancy between VAT data and overall economic activity. The gross value added generated by VAT-paying firms corresponds to 36% of Kenya's annual GDP, on average for the years between 2015 and 2019 (see Appendix Table A3).[14] The gap between value added in the administrative data and aggregate GDP figures arises for two reasons. The first reason is the differential treatment of sectors in the tax code, in particular the treatment of financial services, non-market services, and agriculture. The second reason is informality.

[12] KNBS (2016) obtains information on the number of licensed businesses from county governments and estimates the number of unlicensed businesses based on household survey data.

[13] To estimate the Gross County Product, KNBS (2022) relies on a series of data sets including the 2016 MSME survey and the 2019 population census. Most data sets which cover informal activity are only collected intermittently.

[14] This is substantially lower than high income economies like Chile, where 80% of the country's GDP can be attributed to VAT-paying firms (Huneeus, 2018).

If we exclude sectors that are to a large extent exempt from VAT (i.e. non-market services and agriculture) or have special reporting rules applied to them (i.e. financial services), the VAT sector accounts for 67% of residual economic activity.[15] This implies an overall informal sector share of 33%. Our estimate suggests a larger informal sector compared to the 26% estimated by Elgin et al. (2021) for 2018 and 29% estimated by Hassan and Schneider (2019) for 2013.[16]

This exercise highlights that data from the formal sector may not fully capture the breadth of overall economic activity. The vast majority of firms in Kenya are in fact too small to be VAT-registered. A KNBS (2016) report estimated 7.4 million businesses to operate in Kenya, only one-fifth of which are licensed, and a mere fraction captured in VAT data. In 2016, the VAT data only captured 2.5% of businesses with a county license. Similarly, formally registered employees in VAT-paying firms only represent 5% of the total workforce (as per population census) in 2019.

As an initial investigation into how informality might affect the firm-to-firm trade patterns we observe, we correlate the number of firm-to-firm links in the administrative data with local levels of informality. Even after controlling for population and travel time to metropolitan areas, counties with higher informality show fewer firm-to-firm links in the administrative data (Table A1).[17] These findings motivate us to have a closer look at what types of firm network links we might miss out on in settings with high informality.

Fact 2: Domestic formal sector trade flows are disproportionately centered around Kenya's metropolitan areas.

Next, we show that Kenya's economic activity captured by the administrative data is strongly concentrated around its metropolitan areas, Nairobi and Mombasa. As much as 68% of the sales volume within the network of formal firms is generated by Nairobi-headquartered firms. However, in 2019, as little as 9% of Kenya's population lived in Nairobi County and the city contributed only 33% of Kenya's GDP outside the agricultural sector (see Table 4). Figure A2 shows that while the spatial distribution of firm headquarters is concentrated around these urban centers, the spatial distribution of population is less unequal. These comparisons suggest that the role of urban centers in the Kenyan firm network is disproportionate relative to their

[15] Table A3 details how the GDP share of the VAT sector changes if each of them is removed sequentially.

[16] Both of these studies utilised model-based approaches to estimate aggregate informality as a share of GDP. Part of this difference might stem from our greater reliance on empirical data, while another part might stem from the fact that we focus on the VAT sector as our definition of the formal economy. By doing so, we apply one of the most stringent possible definitions of formality for firms.

[17] We exclude Nairobi and Mombasa-based firms in this regression, as both are strong outliers both in terms of the number of firm-to-firm links and the local incidence of informality. Including them suggests an even stronger correlation.

population and their contribution to aggregate GDP.[18]

Table 4: Geographic concentration of economic activity by degree of formalisation

	Nairobi	**Mombasa**	**Rank regression**	
	in %		α	SE
Population overall	9	3	1.29	0.18
Population of cities & towns	31	9	0.85	0.01
GDP	25	5	1.00	0.07
GDP w/o agriculture	33	7	0.97	0.05
GDP w/o non-market services	25	5	0.91	0.08
No. MSMEs	14	3	0.86	0.17
Employment in MSMEs	19	3	0.78	0.13
No. licensed MSMEs	18	3	0.73	0.09
Employment in licensed MSMEs	28	3	0.67	0.07
No. SMEs	37	3	0.58	0.06
Employment in SMEs	36	3	0.60	0.05
No. census establishments	36	4	1.10	0.12
No. firms census of industrial production	48	6	0.54	0.02
Sales census of industrial production	61	7	0.32	0.03
No. VAT firms	64	9	0.63	0.03
Employment in VAT firms	62	9	0.36	0.03
Value added of VAT firms	72	10	0.38	0.03
VAT network sales	68	13	0.35	0.02
VAT network outlinks	69	11	0.35	0.02
VAT network purchases	60	9	0.43	0.02
VAT network inlinks	65	10	0.48	0.02

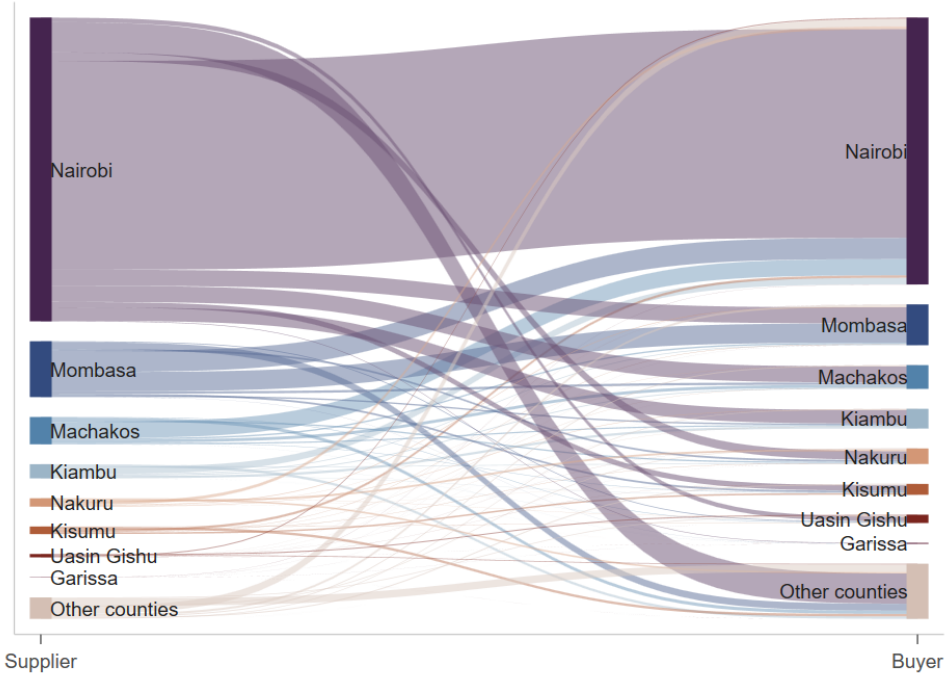
The columns for Nairobi and Mombasa report their share of the respective national aggregate figures (e.g., Nairobi's contribution to Kenya's GDP). The Pareto exponent α is the estimated coefficient from a county-level regression of each county's rank (log) on the respective measure x (log): $\log \text{rank} = \log A - \alpha \log x$. All measures reported in the final section of the table are derived from the administrative data. All other measures are based on data sources summarised in Table 2.

The county-to-county trade flows in Figure 2 underscore the primacy of Nairobi and Mombasa. The size of each segment on the left is proportional to the respective county's sales within the network, while segments on the right are proportional to purchases. The color of the trade flows aligns with the county of origin. Of the over 21.5 million firm-to-firm transactions in 2019, 89% involved at least one firm based in Nairobi or Mombasa.

We use Pareto exponents (α) to measure spatial concentration (Gabaix, 2009; Soo, 2005) beyond simply considering the share of activities attributed to major cities. The third column in Table

[18] Nairobi and Mombasa first emerged as Kenya's primate urban centers during the establishment of a European colonial economic system. Both locations, and Nairobi in particular, were strategically developed as entrepôts along the Kenya-Uganda railroad and the region's communication network (Memon, 1976; Obudho, 1997). The railroad in turn followed existing caravan routes. Mombasa and Nairobi then gradually replaced Zanzibar as the major trading hub of the region (Memon, 1976). In 1960, Nairobi-based firms generated 49% of turnover and employed 46% of the workforce of Kenya's wholesale sector (MoF, 1963, as cited in Memon (1976)). Back then, Nairobi accounted for as little as 3% of Kenya's population. Mombasa accounted for 35% of turnover and 27% of employment in the wholesale sector.

Figure 2: County-level trade flows between formal firms



The figure shows inter-firm trade flows aggregated at the county level. The size of each node (segment) is proportional to the county's share of purchases and sales relative to the aggregate volume of firm-to-firm trade between formal firms in Kenya. The colour of the edges (links between segments) indicates the direction of the trade flow. They take the colour of the supplying county (e.g., goods and services provided by firms in Nakuru to firms in Nairobi take the colour of the segment for Nakuru). The width of each edge (links between segments) is proportional to the share of the trade flow with respect to the aggregate volume of trade flows in the transaction-level administrative data. To improve readability, we only separate the trade flows for eight counties (prioritising those with the largest aggregate amount of transactions and those that act as regional hubs). We bundle the trade flows for the remaining 39 counties.

4 reveals that while county-level GDP and population show relatively even distributions (α close to one), measures of economic activity derived from the administrative data exhibit much higher spatial inequality. The Pareto exponents for employment, value added, and trade flows are 57-76% lower than for overall economic activity. Outlinks with buyers ($\alpha = 0.35$) are more spatially concentrated than inlinks from suppliers ($\alpha = 0.48$), indicating that the supply of inputs to the broader network is concentrated in a few counties.

What margins of trade drive this concentration? In Appendix A.4, we show that the spatial concentration is primarily driven by firm locations and the number of firm-to-firm relationships. The number of transactions and average trade volumes per transaction only play a minor role in explaining spatial variation in trade patterns.

A potential concern is that the observed spatial concentration is driven by the fact that we only observe firm headquarter locations, which in turn are more likely to be based in Nairobi or Mombasa. In Appendix A.5 we use micro-data from the 2010 Census of Industrial Production (KNBS, 2010) to compare the spatial concentration of sales and firm locations with

and without multi-establishments. We find that the excess spatial concentration introduced by multi-establishments cannot explain the aggregate concentration patterns of formal private sector activity.

To assess whether the spatial concentration is particular to the VAT data or captures patterns innate to formal activity, we further report spatial concentration observed in various other measures of economic activity ranging from the least formal to the most formal. This also allows us to show that less formal economic activity is more evenly dispersed than formal economic activity. For instance, the universe of both unlicensed and licensed businesses (KNBS, 2016) exhibits a more even dispersion across space compared to licensed businesses alone. In turn, licensed businesses show a more equal distribution than formal entities engaged in industrial production (KNBS, 2010), many of which were likely VAT-paying firms in 2010. Combined, this evidence suggests that the spatial concentration of economic activity observed in the administrative data is likely a formal sector phenomenon.

Fact 3: Incidence of informality varies by sector, geography, and position along the supply chain.

We will now investigate whether informality is randomly distributed across the economy or systematically varies by sector, geography, and position in the supply chain. If informality is not randomly distributed, we expect that accounting for the informal sector will systematically alter the structure of the observed network. Such corrections could have implications for patterns of sectoral connectivity and county-level linkages, thereby refining our predictions of how individual segments of the economy respond to shocks.

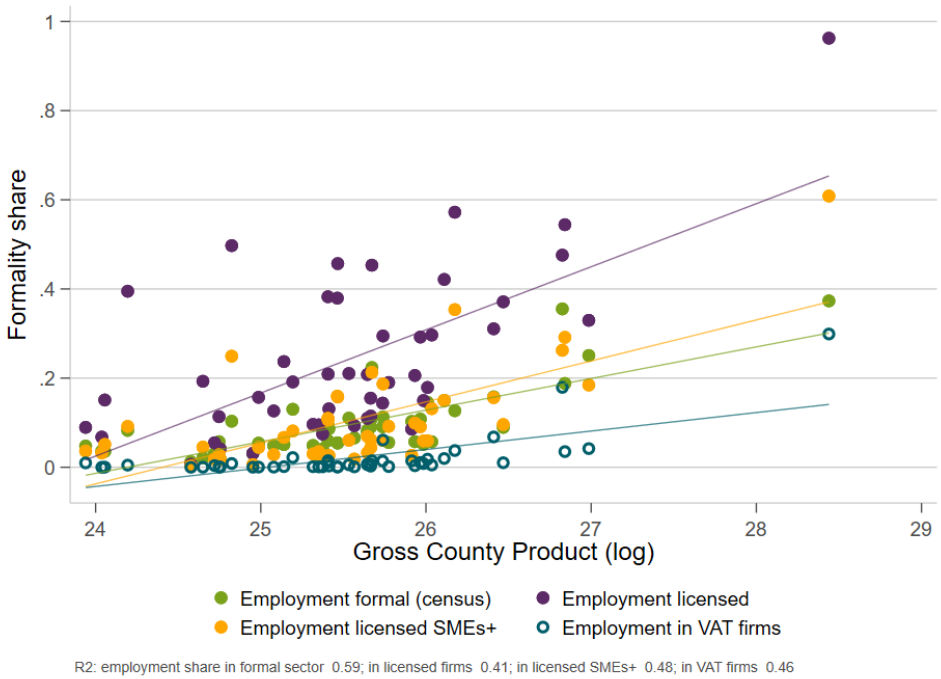
Informal-sector shares correlate negatively with regional economic size and income.

First, we explore the spatial distribution of informal firms, which we find, predominantly reside in smaller markets. Figure A3 plots the distribution of formal sector shares across counties, measured using both value-added and employment metrics. The graph shows that in most counties, the formal sector accounts for less than 20% of economic activity.

We find a strong correlation between a county's formal sector share and both its economic size (measured by Gross County Product) and income level (measured by Gross County Product per capita). As shown in Figure A4, economic size alone explains between 35% and 52% of the variation in formal sector shares across counties. This pattern is consistent across all three measures of economic activity: employment, value added and the number of firms. To validate that this positive correlation between market size and formal sector share is not merely an artifact of the administrative data, Figure 3 presents correlations between Gross County Product and

three additional employment-based formality measures that do not rely on the administrative data. Notably, while more stringent definitions of informality yield flatter slopes, the R^2 remains stable. This consistency suggests that economic size explains similar proportions of county-level informality variation regardless of the measurement approach.

Figure 3: Share of formal sector employment and regional market size



The first measures uses the formal sector employment share according to the 2019 population census, the second measure considers the number of employees in licensed businesses, the third uses the same measure but disregards micro-enterprises, and the fourth measure considers employment in the tax records. Each measure represents a share, i.e. captures the proportion of economic activity that can be attributed to the formal sector. For an exact definition of each measure see Table 3.

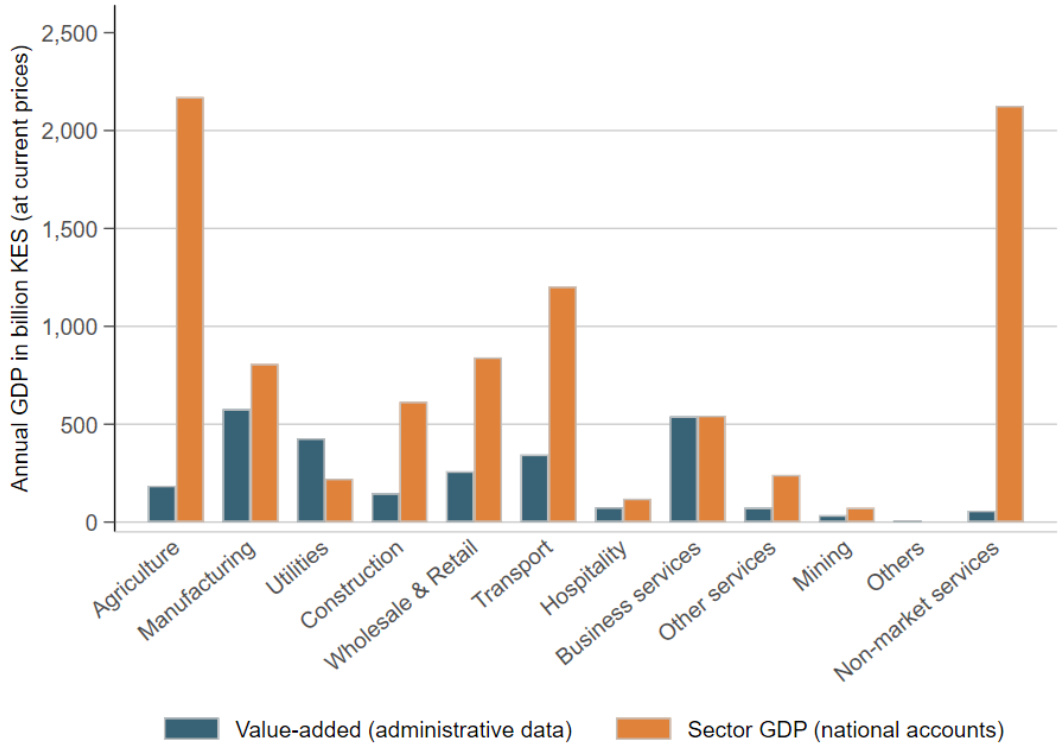
The incidence of informality systematically varies across sectors.

Beyond geographic patterns, informality also varies systematically across sectors. Figure 4 compares a sector's value added (from administrative data) with its contribution to Kenya's GDP (from national accounts). Manufacturing and business services show the closest alignment between these measures, which suggests that the bulk of economic activity in those two sectors takes place in the formal economy. This pattern is consistent with the fact that both sectors rely predominantly on inputs from other firms and tend to sell to other businesses (Figure A1).

In contrast, downstream sectors closer to final consumers exhibit larger disparities between value added and GDP contributions (Figure A7). This pattern aligns with weaker self-enforcement in consumer-facing sectors inherent to most VAT systems (Pomeranz, 2015; Naritomi, 2019). We observe similar patterns in both the extensive margin of informality (comparing firm counts across data sources, Figure A5) and the intensive margin (comparing formal and informal em-

ployment from the 2019 population census, Figure A6). Both measures indicate higher informality in downstream sectors such as wholesale, retail, and other services.

Figure 4: Value added by VAT firms vs GDP



This graph compares the sector-level contribution to national GDP to the value added (sales - purchases) of firms covered in the administrative tax records for 2019.

Informal firms are located downstream of larger firms.

While the previous section indicates that informal firms predominantly operate in downstream sectors, we now utilize survey data to further document their location along the supply chain. We find that informal firms predominantly operate downstream of large formal firms and in consumer-facing roles. If interactions between large firms and smaller, often informal firms occur, they often follow the following pattern: large firms serve as input providers to informal businesses, whilst informal firms primarily act as distributors, serving end consumers.[19]

We draw on survey data on trading partners of micro, small and mid-sized enterprises (MSMEs) by KNBS (2016), which asks about a firm's main source of input and main type of customer. Only 2.3% of all MSMEs state that a large firm is their main customer, while 14.5% rely on large firms as their main source of inputs.[20] Figure 5 shows that the pattern holds across sectors.[21]

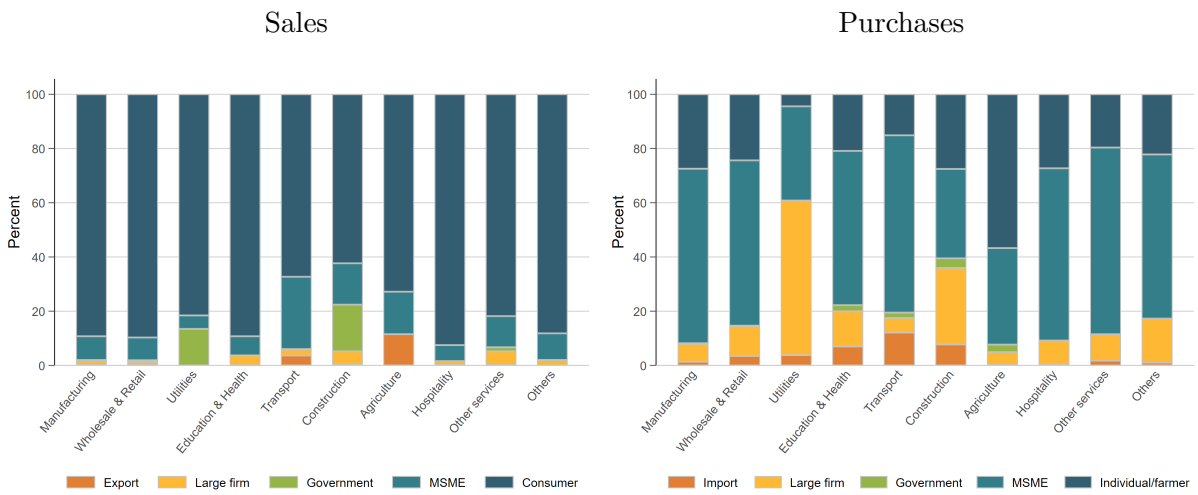
[19] Cordaro et al. (2022) document this pattern in Kenya, showing how microenterprises distribute fast-moving consumer goods for multinationals.

[20] KNBS (2016) defines large firms as those with more than 99 employees, which is larger than the average VAT-paying firm.

[21] The survey responses can be interpreted as a lower bound estimate of the interaction between the VAT-

These results align with findings by Böhme and Thiele (2014); Zhou (2022) who document similar linking patterns between formal and informal firms in Benin, Burkina Faso, Côte d'Ivoire, Mali, Sénégal, Togo (Böhme and Thiele, 2014) and India (Zhou, 2022) respectively.[22] Gadenne et al. (2022) use granular data covering a wide range of sectors to document that non-VAT paying firms that participate in a simplified tax scheme in West Bengal, India, also sell little to VAT-paying firms, but purchase between 50-75% of their inputs from VAT-paying businesses. The higher

Figure 5: Links of small and medium sized enterprises to large firms



The figure draws on data from the 2016 Small and Medium Enterprises (MSME) Survey by the Kenya National Bureau of Statistics (KNBS, 2016). We restrict the sample to participating firms with an annual revenue below the VAT registration cut-off. The survey asks each firm for their main input sources and their main customer type. Note that the customer and supplier category "MSME" also contains medium sized firms which can include formal tax-registered firms. The percentage captured in the "Large firm" category thus represents a lower bound on linkages between small non-VAT registered businesses and large VAT-registered private sector firms. KNBS (2016) defines non-MSMEs/large firms as entities with more than 99 employees.

incidence of informality in downstream sectors, as well as informal firms being more likely to purchase from larger firms rather than vice versa, are consistent with the underlying enforcement structure of VAT systems. The VAT system incentivises downstream firms to request receipts from suppliers to claim input VAT deductions against their output VAT obligations. However, end consumers and VAT-exempt entities lack incentives to demand receipts as they cannot claim VAT refunds (Naritomi, 2019). Consequently, we indeed expect downstream sectors to exhibit higher shares of economic activity outside the VAT system.

registered and non-VAT-registered sector. While MSMEs primarily trade with other MSMEs, the survey does not differentiate between micro, small, and medium enterprises. Under Kenya's Micro and Small Enterprises Act No.55 of 2012, small enterprises include those with up to 50 employees and KShs five million annual turnover (KNBS, 2016), thus medium-sized enterprises often exceed the VAT threshold.

[22]Note that Böhme and Thiele (2014) covers multiple sectors, while Zhou (2022) focuses on manufacturing firms only. Meagher (2013) provides a more detailed literature review on links between the formal and informal economy.

4 A Network Formation Model with Heterogeneity in Sectors, Regions, and Firm Size

We now present and estimate a network formation model to (i) predict the formal firm network as observed in the data and (ii) estimate a revised network that accounts for informal firms. We will use the revised network to measure the extent to which ignoring informality has implications for the spatial patterns of domestic trade and the geographic variation in the pass-through of domestic and international trade shocks.

4.1 Model Motivation

We rely on the network formation model outlined in Bramoullé et al. (2012). This model is particularly well-suited for our purposes for three reasons.

First, it focuses on the entry of nodes into the network and the formation of links among them. In other words, the model captures the extensive margin of trade, i.e. firm location and firm-to-firm links. As discussed earlier and document in Appendix A.4, we show that these two components account for 70-90% of the variation in trade flows.

Second, it allows us to easily incorporate three key dimensions of firm heterogeneity that can affect network formation - sectors, geography, and size. The sectoral dimension captures the underlying input-output structure, while the geographic dimension also allows us to study the question of spatial inequality. The size dimension captures both the well-documented positive relationship between firm sales and firm-to-firm connections(Bernard et al., 2022; Bacilieri et al., 2023) as well as potential differences in how small firms organize their supply chains across space and sectors.[23] Table B1 shows that small firms within the same sector and county are less likely to directly source from manufacturing firms (upstream), but instead are more likely to source from retailers or wholesalers (downstream). Further, they are less likely to source from Nairobi-based suppliers and more likely to source locally.

Third, the model incorporates a flexible network formation process such that the emergent degree distribution can follow a power law. The underlying dynamic network formation process gives rise to the substantial inequality in outdegrees across firms that has been widely documented in the literature (Bernard and Moxnes, 2018; Panigrahi, 2022; Bernard et al., 2022; Bacilieri et al., 2023). Figure B1 plots the outdegree and indegree distribution of the formal firms in our VAT data, revealing a very unequal degree distribution that resembles a power law. Our framework is flexible and allows us to estimate the share of firm-to-firm links formed

[23] Economies of scale in trade cost at the firm level give rise to supply chain structures with several intermediaries (Grant and Startz, 2022). Hence, we expect the linking patterns of small firms to diverge from large firms. Relevant for our case, economies of scale can result in firms of different sizes, but operating within the same geography and sector exhibiting different sourcing patterns.

via preferential attachment versus undirected search (often referred to as *random search* in the networks literature (Jackson and Rogers, 2007)).

4.2 Model Setup

Consider an economy with a set of firms denoted by N. Each firm $i \in N$ is of a given type $\theta_i \in \Theta$ where Θ is the set of all possible types. In our application, we specify firm types as unique sector-county-size combinations, i.e. all firms in the same sector, county, and size group are classified as the same type. Our aim is to predict a matrix that captures the number of links that exist between every possible pair of firm types.

The network formation process is as follows. In every period t, a new buyer firm of type θ enters with probability $p(\theta)$. Hence, the number of firms in the network in any given time period is equal to the number of time periods that have passed since $t = 0$. In order for its operations to be viable, the new firm needs to source inputs from a fixed number of suppliers m. To do this, it first chooses a sector-county-size pair (i.e. a type) with probability $p(\theta, \theta')$ for all θ' in Θ. The probabilities $p(\theta, \theta')$ represent the firm's bias in terms of sectors, regions, and firm size types it wants to link with. In other words, the probability that a buyer of type θ finds a supplier of type θ' may not necessarily be equal to the probability of θ' in the firm population. These biases can reflect production technologies or homophilous preferences arising out of search costs and information frictions. Firms in a location θ might find it easier to link to firms in location θ' that is close to them as opposed to firms in location θ'' that is far. Likewise, firms in sectors that supply services like electricity or telecommunication, which almost every firm requires as inputs, might find themselves with linking probabilities $p(\theta, \theta')$ that exceed their entry probability $p(\theta)$.

Having chosen the sector-region-size type it wants to link with, the firm now relies on two different search technologies to form its m links: first, undirected search (a.k.a. random search). Here, the new firm 'randomly' links to other firms of the chosen type. It forms a fraction r of its total m links in this manner. Second, preferential attachment. The new firm forms the remaining fraction $1 - r$ of its m links to suppliers by searching among the existing suppliers it acquired via undirected search. In other words, once the buyer firm forms links to the first set of suppliers, it then 'randomly' links with the suppliers of its suppliers. The second step of this process is preferential in that suppliers that are more connected are more likely to be chosen. This process continues for several time periods and the network evolves accordingly.[24]

[24]The model takes the distribution of firm types as given, abstracting from firms' endogenous entry decisions across sectors, regions, and size categories (and subsequently, formality status). Instead, we capture these entry patterns through exogenous probabilities $p(\theta)$ for each firm type, which correspond to the observed spatial, sectoral, and size distributions in our data. While one could extend the model to microfound these entry choices — for instance, to explain the concentration of economic activity in Nairobi —we prioritise matching the observed proportions of different firm types rather than explaining their underlying determinants.

Note that while a firm's number of buyers (outdegree) evolves over time, the number of suppliers (indegree) that a firm has is fixed to m and does not change as new firms are born. While this is a strong assumption that we will maintain, we can also imagine this to reflect a fixed production technology that the firm needs to operate. Our focus on endogenizing the outdegree distribution (i.e. the number of buyers of a firm) is motivated by two stylized facts. First, the outdegree distribution in firm networks has been widely documented to exhibit a substantially higher degree of heterogeneity than the indegree distribution (Bacilieri et al., 2023), a fact that also replicates when we consider the spatial distribution of firm links in Figure B2 and Figure B3. Second, given that informal firms are more likely to operate downstream in the supply chain, not observing them is more likely to affect the outdegree rather than the indegree of formal firms.

Ultimately, we are interested in the number of links between each sector-county-size type and their outdegrees. To this end, consider a matrix B where each row and column represents a type $\theta \in \Theta$. Its $\theta\theta'$'th entry is then equal to $p(\theta)\frac{p(\theta,\theta')}{p(\theta')}$. Bramoullé et al. (2012) rely on $\mathbf{B}$ to derive the matrix π whose ij'th entry shows the number of directed links at time t between buyers of type i and suppliers of type j which are born in t_0:

$$\pi^t_{t_0} = m\frac{r}{1-r}(f(t,\mathbf{B}) - \mathbf{I}) \tag{1}$$

Here, t refers to the time period, $\mathbf{I}$ is the identity matrix, and f is a scaled geometric series of the matrix $\mathbf{B}$ defined as follows:

$$f(t,\mathbf{B}) = \sum_{\mu=0}^{\mu=\infty} \frac{((1-r)\log(t)\mathbf{B})^\mu}{\mu!}$$

Newly entered buyers form m inlinks in every period. As a result the outdegree of existing firms, i.e. the suppliers of the newly entered firms, evolves over time. Thus, the matrix $\pi^{\mathbf{t}}_{\mathbf{t_0}}$ gives the expected outdegree (i.e. number of buyers) of each column node born in time t_0 to a row, computed at time t. The purpose of the dynamic network formation process is to rationalise the heterogeneity in outdegree.

4.3 Estimation Strategy

Given the granular data on the empirical formal sector firm-to-firm network, we are able to obtain the majority of the model parameters directly from the data (see Table 5 for an overview). These include all entry probabilities $p(\theta)\ \forall\ \theta \in \Theta$ that a firm enters in a given sector, county, and size group as well as all interaction probabilities $p(\theta,\theta')$ between all possible sector-county-size types. We use the cross-section from 2019, the last pre-COVID year of our panel, to obtain the $p(\theta)$s

and $p(\theta, \theta')$s.[25]

The parameter we need to estimate is r, the fraction of input links a firm obtains via undirected search independent of the network environment.

Table 5: Model parameters

Parameters	Description	Source	Proxy	Value
r	Share of suppliers via random search	Estimated	-	0.45
$p(\theta)$	Entry probability of type θ	Data	Share of firms observed as θ	(0,0.12]
$p(\theta, \theta')$	Linking probability of θ and θ'	Data	Share of links between θ and θ'	(0,1]
m	Indegree	Data	Avg. number of suppliers	30
t	Number of entry periods	Data	Number of firms in admin data	56822

First, we classify firms into types defined by unique sector-location-size combinations. Sectors refer to 13 aggregate sectors, namely, agriculture, mining, manufacturing, utilities, construction, transportation and logistics, wholesale, hospitality, retail, business services, non-market services, other services, and miscellaneous (incl. international organisations and non-classified firms). Locations are given by the county in which the firm is located. Within each sector and county we further group firms into large and small firms. We define small firms as firms in the bottom sales quartile within a sector-county group. By restricting ourselves to two size bins only, we avoid having too few observations in each firm-type bin and the matrix of linking probabilities becoming too sparse. For example, all firms in the top three sales quartiles of Nairobi's manufacturing sector are classified as the same type.

Next, we compute the probability that a type exists for all types in Θ. We do so by dividing the number of formal firms of a sector-county-size type by the total number of formal firms in the economy. The interaction probabilities $p(\theta, \theta')$ then represent the fraction of a sector-county-size type θ's supplier relationships that it forms with type θ'. We compute the above probabilities for all possible combinations of types and use them to construct the matrix $\mathbf{B}$. Moreover, we follow Jackson and Rogers (2007) and define m as the average indegree (i.e. average number of suppliers) in the network. The variable t, by definition, is equal to the number of firms observed in the data equal to 56,822.

Using the parameters from the empirical data, we are able to predict the matrix of type-to-type network links $\pi(r)$ for different choices of $r \in [0, 1]$. Appendix B.2 discusses the practical steps needed to construct the matrix during the estimation.

In addition to the predicted version of the matrix π, we also observe the actual π in the data where the ij'th entry of π is just the number of links between types i and j. We match the model predicted matrix and the matrix in the data using the method of moments procedure to obtain

[25] We exclude a small proportion of firms firms with zero suppliers in the data, as the model requires all entrants to form m buying links.

r^*. Each moment is weighted by the probability with which we observe a specific sector-region-size type in the data. In doing so, we assign greater weight to more common sector-region-size types whose probabilities tend to be more stable over time. r^* is defined as follows:

$$r^* = \arg\min \sum_{\theta} \sum_{\theta'} p(\theta')(\pi_{model}(\theta, \theta'; r) - \pi_{actual}(\theta, \theta'))^2 \tag{2}$$

r^* is obtained by minimising the distance between the model predicted matrix of type-by-type interactions and the corresponding matrix obtained from the data (method of moments). We estimate r using simulated annealing. With only one parameter to estimate, we can plot the objective function for various values of r to ensure that our estimated value is indeed the global minimum (see Figure B4).

4.4 Estimation Results

Our estimation strategy yields a result of $r^* = 0.45$. It suggests that a newly entered firm chooses 45% of its m suppliers at random, and the remaining 55% among the suppliers of its existing suppliers. A network with 55% of all links being formed via preferential attachment suggests a prominent role for information frictions as firms rely on their suppliers to form new links. It aligns with previous research documenting the importance of relational contracts in Kenya and neighbouring economies (Fafchamps, 2003). In a variant of this model, Chaney (2014) estimates $r = 0.6$ for French exporters forming links with trade partners abroad, which also suggests a substantial, but not quite as prominent role of information asymmetries.

4.5 Model Fit

To assess how well our model does in fitting the targeted outdegree distribution (i.e. distribution of the number of buyers), we plot the degree distribution (i.e. total number of outlinks) of each sector-county-size type as observed in the data and as predicted by the model. Figure B5 shows that the key properties of the outdegree distribution are replicated by the model's predictions. The model and data match particularly well in the right tail of the distribution i.e. the part that is specifically targeted by the preferential attachment framework.

We also estimate the Pareto exponent α, which was not explicitly targeted by the model, for both degree distributions. We obtain an α of 0.36 from the model and 0.37 from the data. In addition, since we have previously shown that the formal firm network is spatially concentrated, we also assess how well the model predicts the share of buyer relationships in the economy that originate from firms in Nairobi, Mombasa, to firms in other counties. We find that the model performs well on this dimension too (Table 7). In the administrative data, 69% (11%) of all outlinks in the economy are captured by Nairobi (Mombasa) based firms and the model predicts

a share of 70% (11%).

5 Predicting a Revised Network

With the estimated model at hand, we are now able to tackle the question of how including informal firms might affect the spatial patterns of domestic trade. Our proposed thought experiment is the following: suppose we were to observe informal firms. What would happen to the outdegree distribution of various types θ? To obtain a 'revised network' that accounts for informal firms, we rely on updated information on the spatial and sectoral dispersion of economic activity in Kenya – now including the informal sector. In model terms, our exercise shifts the probabilities $p(\theta)$ with which we observe nodes of certain sector-region-size types θ to be born. Knowing r^* and our updated $p(\theta)$s, we can then once again predict the type-by-type matrix of firm-to-firm links π, keeping everything else constant. We will also make additional assumptions about the linking probabilities of informal firms.

5.1 Predicting the Sector-County Profile of Non-VAT Firms

To incorporate informal firms into the network, we first update the firm-type probabilities $p(\theta)$ for each sector-county-size cell, this time accounting for the entire firm size distribution. To update $p(\theta)$, we ideally would want to observe the number of firms N_{cs} in each sector s, county c, and size cell – irrespective of their formality status. However, none of the KNBS records available to us feature a breakdown of the firm count along both the sector s and the county c dimension, let alone size dimension. Therefore, instead of the firm count, granular sector and region level information on formal and informal employment in the 2019 census labour force module to compute our alternative entry probabilities $p(\theta_a)$:

$$p(\theta_{a,\,large\ formal}) = \frac{e_{sc}^{\text{formal}}}{\sum_c^{47}\sum_s^{13} o_{sc} + e_{sc}} \times \frac{1}{\overline{x}_{sc}^{\text{formal}}} \tag{3}$$

$$p(\theta_{a,\,informal/small\ formal}) = \frac{o_{sc}^{\text{formal}} + o_{sc}^{\text{informal}} + e_{sc}^{\text{informal}}}{\sum_c^{47}\sum_s^{13} o_{sc} + e_{sc}} \times \frac{1}{\overline{x}_{sc}^{\text{informal}}} \tag{4}$$

where o_{sc} is the number of self-employed people, and e_{sc} the number of (wage) employed people.

The denominator sums total private sector employment (both wage and self-employed, formal and informal) across all 13 sectors and 47 counties. The updated sector-region-size probabilities $p(\theta_a)$ will again sum to one. The updated $p(\theta_a)$s hence capture a relative change in the number of firms rather than an absolute change.

Using simple employment shares to compute $p(\theta_a)$ relies on the assumption that the mapping

of employees to firms is the same across all sectors and regions. However, empirically, manufacturing firms, for example, tend to be larger than businesses in the hospitality sector. Nairobi hosts larger firms than Mandera County in Kenya's north. We therefore re-scale the number of employees by the average firm size in each sector-county-size cell $\overline{x}_{sc}^{formal,informal}$. For small formal and informal firms, we rely on the KNBS (2016) survey to compute the average number of employees, while we use the administrative data for large formal sector firms.[26]

For agriculture and non-market services, we estimate their $p(\theta_a)$ drawing only on formal private sector employment. Formal VAT-paying firms occupy a very specific niche in both cases (e.g. formal firms in these sectors are disproportionately export-oriented or the sector is dominated by non-profit entities, see discussion in Section 2.2.2) and informal employment takes vastly different forms (e.g. mainly reflects subsistence farming).

How does the probability $p(\theta)$ that a formal firm enters in a given sector-county-size cell shift to $p(\theta_a)$ once we account for informal firms? Figure C1 suggest that a 10 percentage point increase in formality corresponds with a 0.5 percentage point increase in $p(\theta)$-$p(\theta_a)$ (0.35 standard deviations). As expected, $p(\theta_a)$ is lower than $p(\theta)$ for sectors and counties with a high degree of formality, indicating their importance for the overall economy is overstated in the administrative data. To recap, our proposed revised network accounts for informal firms being born into the network based on their sector-region profile. Rather than thinking of the exercise as adding new firms, we adjust the weights of each sector-region-size type.

5.2 Assumptions about Linking Probabilities of Informal Firms

Another challenge for integrating informal firms into the network of formal firms arises due to the lack of granular data on the sectoral and geographic composition of how informal firms link with both each other and with the formal sector. An ideal data set would provide details on sourcing and selling patterns by sector, geography, and formality status. In the absence of such data and given the strong correlation of size and formality status, our default approach will be to assume that informal firms exhibit linking preferences $p(\theta, \theta')$ similar to those of small formal firms, conditional on sector and geography.

This approach is particularly attractive given its straightforward implications for the sectoral and geographic composition of linking patterns. It further allows for informal and small formal

[26] If big formal firms employ informal casual workers not captured in the administrative data, we understate their size and hence overestimate the probability of big formal firms in the network. As a result, our revised network becomes biased towards the baseline network that only covers the formal sector. This is also illustrated in Figure 6 where we compare the spatial inequality in county-level outdegrees in the baseline network to several scenarios that account for informal firms. Comparing the two scenarios where in one case we use the simple employment shares to compute $p(\theta_a)$ and in the other case further adjust for differences in firm size across sectors and counties, we find that in line with the intuition on the implications of informal workers in formal firms, the former scenario is closer to the original network with only formal firms.

Table 6: County-level changes in the dispersion of outdegrees

County outdegree	Δ sd/mean (in %)
All counties	-7.5
Without Nairobi & Mombasa	-18.0

The above table reports the difference in outdegrees between the original and the revised network - aggregated at the county level. We look at the coefficient of variation as the key metric. Adjusting for the mean accounts for the fact that the change in the number of outlinks predicted by the model needs to be interpreted in relative rather than absolute terms. We exclude the outdegrees of Nairobi and Mombasa when we compute the coefficient of variation in the second row.

firms to experience different wedges in link formation, provided these wedges generate aggregate linking patterns that are still comparable to those of small formal firms at the sector-county-size level. Consider, for instance, a small formal retailer and an informal retailer both seeking to purchase soap. While neither can source directly from manufacturers in Nairobi, the small formal retailer might purchase from a large formal wholesaler locally, whereas the informal retailer—potentially excluded due to their tax status—might source from an alternative local wholesaler. Despite relying on different suppliers, both retailers exhibit similar sectoral and geographic sourcing patterns. The similarity assumption is also motivated by findings in the administrative data, where we document that small formal firms tend to source more locally and from intermediaries (see Table B1).

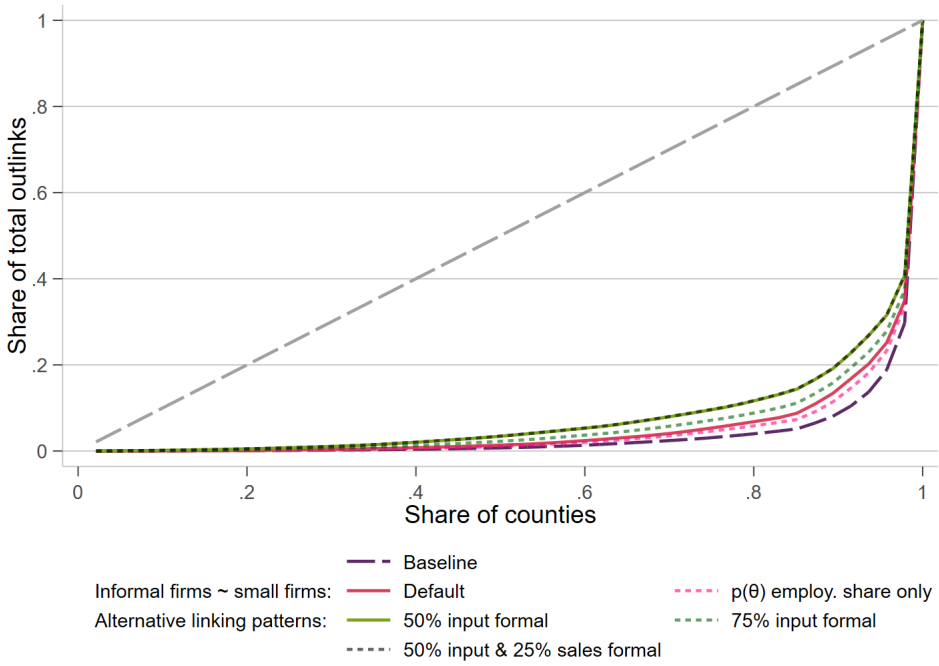
Nevertheless, this assumption many not fully capture the implications of challenges specific to informal firms for their linking patters. Hence, we later introduce modifications of our approach, drawing on stylized facts from the existing literature, to test the sensitivity of our results to alternative assumptions about the linking patterns of informal firms.

5.3 Characteristics of the Revised Network

How does the revised network that accounts for informal firms compare to the baseline network? First, we find that firm types in sectors and counties with a high incidence of informality are predicted to have a relatively larger increase in outdegrees (Figure C2). Second, the share of total outlinks attributed to Nairobi- and Mombasa-based firms declines from 80% to 75% (see Table 7). While these two cities maintain their prominent role in the network, this shift is meaningful for the remaining counties in relative terms as it represents a 25% increase in their outdegrees. Third, accounting for informal firms reduces the variation in outdegrees across counties by 7.5% (Table 6). We visualize this reduction in inequality by plotting the Lorenz curve for county-level outlinks in Figure 6.

What are key drivers of this reduction in inequality in outdegrees? First, Nairobi and Mombasa become less important as a destination for products and services from other counties. In Figure 7 we plot the row-normalised adjacency matrices, before and after accounting for informal firms,

Figure 6: Inequality in county-level outlinks in the baseline and revised network



To visualise the change in inequality between the baseline and the revised network, we plot the Lorenz curve for the number of outlinks at the county level. The default scenario uses the entry probabilities $p(\theta)$ specified in Equations 3 and 4 and assumes similar linking patterns for informal firms and small formal firms conditional on their sector and county of operation.

at the county and sector level respectively. The matrix is normalised such that each row sums to one. A smaller proportion of a county's total outlinks now connects with firms in Nairobi, i.e. the column with the lightest colour in the baseline matrix.

Second, downstream relationships with firms in the same county now become relatively more prominent. The values of the diagonal entries of the adjacency matrix increase between baseline and revised network. This is consistent with the stylized fact discussed earlier that smaller firms are more likely to source from the same county (Table B1). With the exception of five counties, most notably Nairobi and Mombasa, trade within the county gains in importance for all of the remaining 42 counties. This is even more explicit in Figure C4 where we compare the change in both inter- and intra-county links for the baseline and revised network. Once we account for informal firms, both inter- and intra-county outdegree increases for the median county and on average. However, the increase in intra-county outdegrees is higher than the increase in inter-county links for 83% of the counties.[27] If informal firms purchase an even higher share of their inputs locally, the predicted shift towards intra-county trade represents a lower bound. We will discuss this in Section 6.1.

If counties are selling less to Nairobi and Mombasa, where do inter-county trade links shift? We

[27] In fact, while inter-county trade rises for the median county, 18 out of 47 counties have fewer links with other counties in the revised network.

Table 7: The importance of Kenya's primary cities in a revised network

	Nairobi	**Mombasa**	**Other counties**
		in %	
Population	9	3	88
GDP	25	5	70
GDP w/o agriculture	33	7	60
Number of outlinks			
Model fit			
Administrative data	69	11	20
Model predicted network	70	11	19
Revised network: informal firms ≈ small firms			
Default scenario	65	10	25
$p(\theta)$ using employ. share only	66	10	23
Revised network: alternative linking patterns for informal firms			
0% sales, 50% input to/from formal	59	9	32
0% sales, 75% input to/from formal	63	10	28
25% sales, 50% input to/from formal	59	9	32

The above table documents the share of overall firm-to-firm links which have a supplier based in Nairobi, Mombasa or the remaining 45 counties. We the spatial dispersion of outlinks in the data, the predicted (formal sector) network as well as our default scenario for the revised network and four alternative scenarios to assess sensitivity.

find that the number of bilateral trade links now becomes more sensitive to social ties between counties. In Table C2, we regress the number of links between two counties on both travel distance and social connectedness (Bailey et al., 2021), comparing the baseline and the revised network.[28] The results show that inter-county links correlate more strongly with the strength of social ties between counties in the revised network.

The increased importance of within-county trade and trade between socially connected counties gives rise to a network with a more pronounced group structure. We quantify the extent to which the network is partitioned by measuring the network's modularity (Newman, 2006). The modularity of a network is higher when groups of nodes have more links among each other than what we would expect in a random network. We compute the modularity of the weighted adjacency matrix at the sector-county level. We find that modularity in the revised network with informal firms increases by about 60% suggesting that the revised network exhibits a more pronounced group structure. To further characterize this group structure, we apply a community detection algorithm to the trade flows between counties as per the revised network. As illustrated by Figure C3, the group structure now correlates strongly with Kenya's geography, i.e., geographically proximate counties are now more likely to be clustered in the same group.

[28] Social connectedness is measured using friendship network data provided by a popular social media platform, see Bailey et al. (2021).

Next, we explore for which group of counties we tend to underestimate the number of links the most. To do so, we regress the change in county-level outlinks between baseline and revised network on county characteristics like the aggregate level of formality, population, Gross County Product, and market access in Table C3. The results indicate that our baseline network, based solely on formal sector data, particularly underestimates connectivity in smaller counties and counties with high market access. Notably, once we control for these other characteristics, the correlation with a county's aggregate formality share is no longer statistically significant.

Finally, we find that incorporating informal firms also reshapes inter-sector trade patterns (Figure 7). Sectors with substantial informal activity like other services, retail, and wholesale, now gain prominence as buyers in the network. Manufacturing, wholesale, and mining experience the largest relative gains in new links.

5.4 Simulating the Effect of Economic Shocks

As a next step, we ask how the newly predicted network that accounts for informal firms compares to the previous network in terms of its role in propagating domestic and international shocks. How does the predicted impact of the shock depend on whether we account for informality? Are sectors and regions with more informality more or less vulnerable to shocks than the administrative data would suggest? To answer these questions, we first simulate a series of domestic output shocks that reduce each firm type's output and then analyse how it affects the output of all other types, both directly and indirectly, by propagating through the network over multiple time periods. Then, we simulate international supply shocks that affect firm types depending on their exposure to international markets. We discuss the results for both domestic and international shocks below.

5.4.1 Domestic Shocks

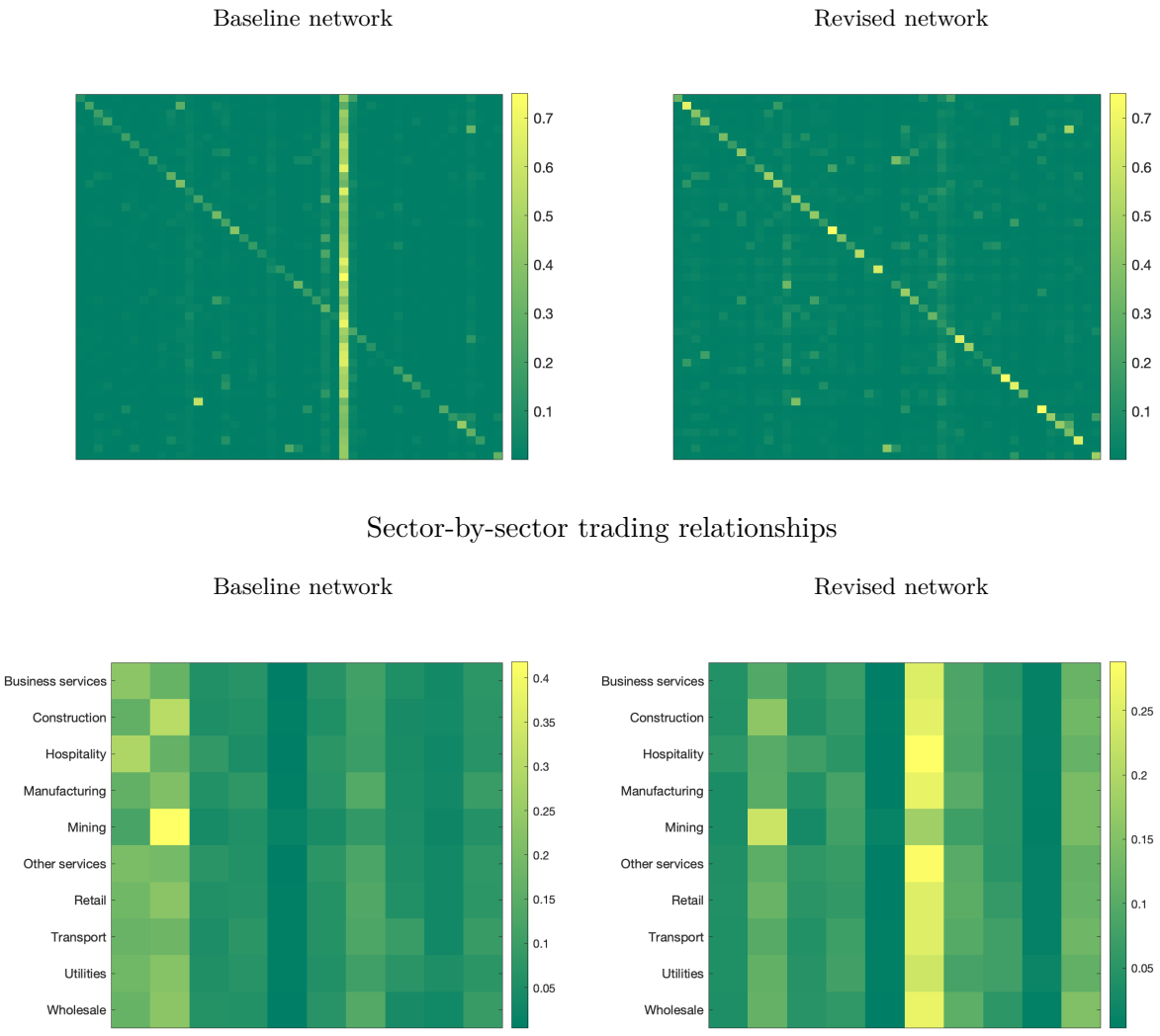
Following the supply-side version of classic input-output models (Sargent and Stachurski, 2022), we define firm type j's output y_{jt} in period t as the sum of inputs it purchases from other types i plus payments to other factors of production (value added) v_{jt}:[29]

$$y_{jt} = \sum g_{ij} y_{it-1} + v_{jt} \tag{5}$$

The intermediate inputs purchased from other firm types are the product of each supplier i's total output in the previous period y_{it-1} and the fraction it sells to type j, g_{ij}. The g_{ij}s represent the normalized cells of our type-by-type matrix π that captures the total number of links between all types. We normalise the rows of π by dividing each entry in a row by the sum of that row. We

[29] Alternatively, v_{jt} can also be interpreted as a type-specific and period-specific shock to output.

Figure 7: Baseline versus revised network



The above figures show heatmaps of the predicted row-normalised adjacency matrix of the network (where row sells to column) as per the baseline $p(\theta)$ on the left and augmented $p(\theta_a)$ on the right at the county level (top) and sector level (bottom).

abstract from any endogenous network adjustments (see e.g. Panigrahi, 2022; Eaton et al., 2022; Arkolakis et al., 2023). We assume that v_{jt} is an independent draw from a uniform distribution $U[-10, 10]$ for every type j in every time period t. Each type starts with a randomly chosen output drawn from the distribution $U[0, 100]$ in $t = 0$.

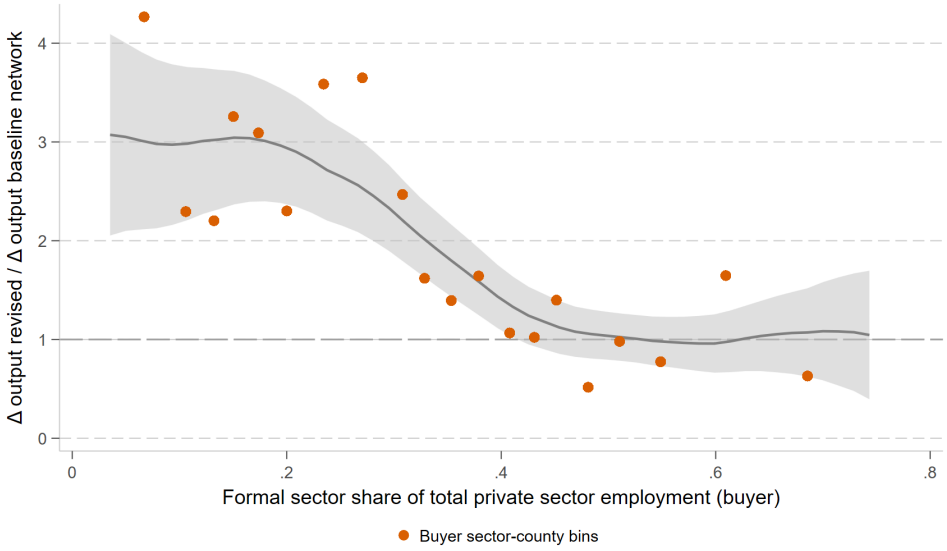
Using this set-up, we first simulate the output process without any shock. Then, we simulate the output process following a negative output shock to sector-region-size type j's value added vjt in the first time period.[30] We repeat this exercise for all types j.

To study the relevance of unobserved informal firms, we conduct our simulation exercise twice.

[30] We compute the impact of the shock on each type's output over 100 periods of time by comparing the two output processes. All of the outputs reported below are averages across the 100 time periods.

In the first scenario, we use the matrix π derived from administrative records. In the second scenario, we use an alternative version of π using our revised network that accounts for the presence of informal firms.[31] Our primary question is: how do domestic shocks impact each type when informal firms are considered versus when they are not? For each simulated shock, we compute: (i) the absolute reduction in output of each type using the original adjacency matrix (excluding informal firms) and (ii) the absolute reduction in output of each type using the new adjacency matrix (including informal firms), both averaged across all time periods. This yields a matrix of shock impacts where each row corresponds to a supplier who is shocked and each column corresponds to a buyer who faces the impact of the shock. We then aggregate across rows to compute the average impact of the shocks on buyers.

Figure 8: How do output shocks pass-through in a revised network that takes into account informal firms? - % change in output drops and the level of formality



The above graph plots the percentage change of the output reduction in response to domestic output shocks for two scenarios: the baseline network using only administrative data and the revised network including informal firms. We aggregate the output reduction across buyer types at the sector and county level, weighted by the entry probability $p(\theta)$ for each size and formality type. The x-axis shows the formal sector share for each sector-region pair.

We find that the higher the incidence of informality in a sector and region, the more we underestimate the adverse impact of a domestic output shock. Figure 8 shows that our established employment-based measure of informality negatively correlates with the impact of the shock under the two scenarios. To align with our most granular measure of informality, we aggregate the response to shocks at the sector-region level.[32] As shown in Table C4, a one percentage point decrease in the formal sector share corresponds with a 4.9 percentage point larger output

[31] We ensure that the random component of output, υ_{jt}, is identical across these two scenarios for each type j in every time period t to ensure it does not affect our results.

[32] Put differently, we average the impact across each buyer's suppliers and then compute a weighted sum for each sector-region cell. The weights are determined by a type's entry probabilities.

drop following a domestic shock.

Figure C5 presents the distribution of the ratio of the shock impact at the sector-county level (revised network/baseline network). This ratio exceeds one for 48% of the sector-county pairs and 42% of the sector-county-size types, indicating the baseline network on average underestimates the impact of domestic shocks for these types. Among the types for which we underestimate the impact, 73% are types with small firms, an indicator that the omission of informal firms is the primary driver of this result.

It is important to note that by considering the changes in output reduction between the two scenarios, we focus on relative shifts. These shifts are more notable for sector-county pairs that face a relatively smaller impact initially due to their peripheral role in the formal sector network. However, even after accounting for informal firms, the aggregate impact of the shock remains largest for Nairobi- and Mombasa-based sectors. This mirrors the finding that the two cities still account for a sizable share of outlinks, as discussed in the previous section.

5.4.2 Import Shocks

In addition to a domestic shock, we consider the impact of a reduction in output in response to an adverse shock to international suppliers whom Kenyan firms source from. As before, firm j's output can be written as follows:

$$y_{jt} = \sum g_{ij} y_{it-1} + m_{jw} y_w + \upsilon_{it} \tag{6}$$

Firm j's output now additionally depends on world output y_w in line with its import share m_{iw}, which we obtain from the administrative data. We re-normalise the rows of the adjacency matrix such that $\sum_j g_{ij} + m_{jw} = 1$. Next, we simulate a series of negative shocks to y_w and analyse how it affects total output in the economy and the heterogeneous effects on various firm types.

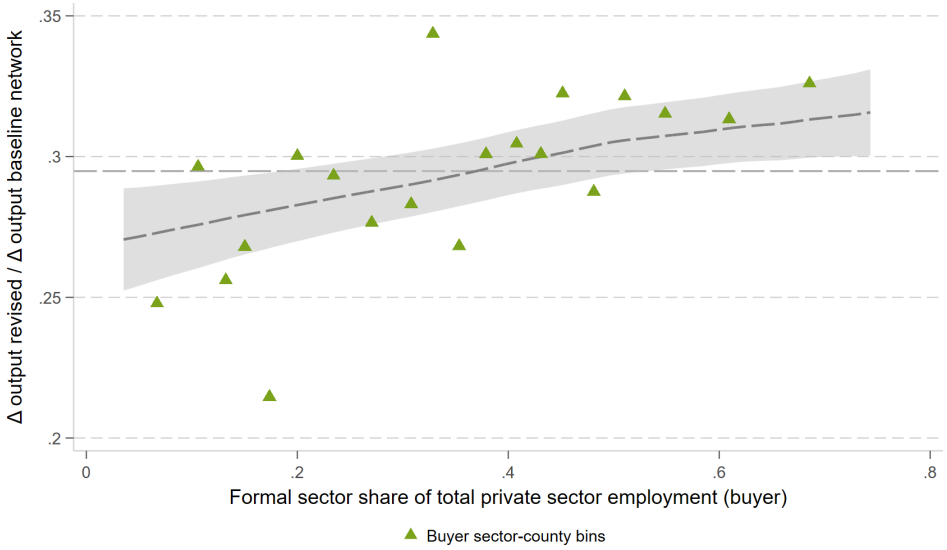
The bottom graph of Figure C5 plots the impact of the shock relying on the revised network as a proportion of the impact based on the baseline network. Unlike domestic shocks, our findings for import shocks indicate that extrapolating from data on the formal sector network to the overall economy leads to an overestimation of the reduction in output. The impact is consistently less negative when using the revised network.

This effect is particularly pronounced in sectors and regions with a higher incidence of informal activity (see Figure 9). Specifically, Table C4 shows that a 10 percentage point increase in the informal sector share corresponds to a 1 percentage point overestimation of the reduction in output. This pattern emerges because formal firms in predominantly informal markets have more

unobserved connections than captured in administrative data, reducing their effective exposure to import shocks.

Why do the predictions differ for domestic and import shocks? When accounting for informality, sectors and counties with a high share of informal activity become more prominent in the network, making them more susceptible to economic shocks. Conversely, this adjustment reduces the relative importance of formal-dominated sectors, which typically have higher import shares and international exposure. By adjusting their prominence (i.e., modifying their entry probabilities and considering informality), we find that the economy seems more resilient to trade shocks but more vulnerable to domestic shocks. The intuition behind our result aligns with the mechanism discussed in Di Giovanni and Levchenko (2012), who show that smaller economies tend to experience more volatility due to having fewer firms and less diversification. Applied to our setting, focusing only on formal sector firms leads to overstating the importance of internationally-linked formal firms and underestimating the diversification of the regional economy.[33]

Figure 9: How does a shock to import markets pass-through in a revised network that takes into account informal firms? - Output drops and the level of formality



The above graph plots the percentage change of the output reduction in response to international trade shocks for two scenarios: the baseline network using only administrative data and the revised network including informal firms. We aggregate the output reduction across buyer types at the sector and county level, weighted by the entry probability $p(\theta)$ for each size and formality type. The x-axis shows the formal sector share for each sector-region pair.

[33] Manelici et al. (2024) show that investments by foreign multinationals in Mexico largely affect domestic formal sector growth with muted effects on the informal sector. Note that their setting accounts for an endogenous response of firms, while we do not consider endogenous network adjustments.

6 How Sensitive are Results to Alternative Linking Patterns?

What happens if we relax the assumption that informal firms have similar linking patterns to small formal firms? Given our lack of disaggregated data on how informal firms link with the formal sector, we stress-test our results relying on alternative assumptions motivated by evidence provided in the literature and consistent with the stylized facts presented in Section 3. First, informal firms almost exclusively sell to final consumers. Second, informal firms purchase from formal firms, but not as much as formal firms purchase from each other (Böhme and Thiele, 2014; Gadenne et al., 2022). Finally, informal firms predominately source locally (Amodio et al., 2024). These assumptions allow us to construct extreme bounds for linking patterns that we might observe if we were to have granular network data with informal firms.

6.1 Alternative Assumptions about Linking Patterns of Informal Firms

Consider now the three groups of firms; small formal firm firms $s_i \in F_s$, large formal firms $l_i \in F_l$, and informal firms $n_i \in N$ with their types (i.e. sector and country composition) denoted by θ_{l_i}, θ_{s_i}, and θ_{n_i} respectively. We will drop the index i in what follows as the conditions do not vary across firms, once we take their sector, county, and size (informal, small, or formal) into account.

We start by considering the sales patterns of informal firms. We consider two alternative linking probabilities that capture the differential sales patterns of informal firms. In one scenario, we set the probability that an informal firm sells to formal firms to zero. In a second scenario, we assume that informal firms form one-fourth of their outlinks with formal firms, where the sector and geographic composition of these links follows those of small formal firms. These assumptions are motivated by stylized facts documented earlier which show that micro, small, and medium enterprises with sales below the VAT threshold rarely sell to large firms in Kenya. This is also in line with findings in Böhme and Thiele (2014) for informal firms in six urban centers in West Africa, and firms that participate in simplified tax scheme and are similar to informal firms in West Bengal, India in Gadenne et al. (2022). Finally, it is consistent with the set up of VAT systems requiring firms to ask for a receipt from their supplier in order to claim input VAT (Pomeranz, 2015). In summary, for all formal and informal types we make the following extreme assumptions regarding the sales patterns of informal firms:

$$p(\theta_n, \theta_l) = 0 \qquad \text{sensitivity: } p(\theta_n, \theta_l) = 0.25 \times p(\theta_s, \theta_l) \tag{7}$$

$$p(\theta_n, \theta_s) = 0 \qquad \text{sensitivity: } p(\theta_n, \theta_s) = 0.25 \times p(\theta_s, \theta_s)$$

We now turn to informal firms and their sourcing patterns. In this case, we assume that informal

firms source only 50% of their inputs from formal firms (large or small) in one scenario and 75% in an additional sensitivity check. Conditional on sourcing from the formal sector, the sectoral and geographic preferences of informal firms will again follow those of small formal firms. As documented in our empirical section, MSMEs do source some of their inputs from large firms. Böhme and Thiele (2014) find that informal firms buy about half as much from formal firms than the amount formal firms source from each other. Gadenne et al. (2022) find that the smallest group of firms under the simplified tax scheme in West Bengal source a similar share from VAT firms, while the largest non-VAT firms source as much as three-quarters of their inputs from VAT firms. We rely on these point estimates to make the following assumption:

$$p(\theta_l, \theta_n) = 0.5 \times p(\theta_l, \theta_s) \qquad \text{sensitivity: } p(\theta_l, \theta_n) = 0.75 \times p(\theta_l, \theta_s) \tag{8}$$

$$p(\theta_s, \theta_n) = 0.5 \times p(\theta_s, \theta_s) \qquad \text{sensitivity: } p(\theta_s, \theta_n) = 0.75 \times p(\theta_s, \theta_s)$$

Informal firms will then source the remainder of their inputs from other informal firms. This requires assumptions about their preferences on which sectors and counties to source from. Motivated by evidence from the literature where Amodio et al. (2024) find that the vast majority of small, largely informal firms in Ethiopia obtain their inputs from local sources, we allow inputs obtained from other informal firms to only be sourced locally.

Consider any counties a and b in the set of counties C and let $\theta_{f,a}$ be the type of small or large firm f in $F_s \cup F_l$ in county a. The assumption implies the following:

$$p(\theta_{n,a}, \theta_{n,b}) = [1 - \sum_{f \in F_s \cup F_l} p(\theta_f, \theta_n)] \times p(\theta_{n,b}) \qquad if \quad a = b \tag{9}$$

$$p(\theta_{n,a}, \theta_{n,b}) = 0 \qquad if \quad a \neq b$$

Finally, we now compute separate $p(\theta)$s for informal and small formal firms, splitting the main term of Equation 4. For additional details see Section C.2.

6.2 Implications for the Revised Network

Table C1 and Figure 6 summarize how alternative assumptions about informal firms affect the distribution of outlinks across counties. If we assume informal firms source half of their inputs from the formal sector but have no formal buyers, inequality in county-level outlinks declines by 16% compared to our model-predicted network. This represents an 8.5 percentage point larger reduction than our default scenario where small formal and informal firms share similar linking patterns. The share of total outlinks accounted for by Nairobi- and Mombasa-based firms now drops down to 69%, down from 80% in the baseline network and 75% in the first scenario with informal firms (see Table 7).

Allowing informal firms to sell 25% of their output to formal firms, instead of none, has little implication for both the overall decline in inequality and Nairobi's share of outlinks. Allowing for a larger share of informal firms' inputs to be sourced from the formal sector results in a smaller reduction in inequality (11%) and a higher share of Nairobi- and Mombasa-based links (73%), more closely matching the initial network with informal firms. This pattern is driven by the greater reliance of informal firms on formal suppliers in this scenario, which in turn tend to locate in counties with larger formal economies.

Our simulations of domestic and international trade shocks under these alternative assumptions reinforce our earlier findings, showing even stronger relative effects in sectors and counties with high informality (Figure C6). This more pronounced impact stems from increased intra-county linkages among informal firms. While more prominent local linkages reduce spatial inequality and urban concentration, they amplify the impact of domestic shocks through stronger within-county multiplier effects.

7 Conclusion

How representative are the patterns observed in formal firm-to-firm trade data of overall domestic trade patterns in contexts with high levels of informality? In this paper, we show that formal firm-to-firm trade data lead us to overestimate the spatial concentration of overall domestic trade flows. We first explore this by presenting a series of stylized facts that exploit both VAT data on firm-to-firm trade as well as alternative sources of data on the informal sector. We then complement these results with findings from a structural model that allows us to predict a revised network that accounts for the dispersion of informal activity across space and sectors.

We find that formal sector data underrepresents trade within counties and trade between counties that have strong social ties. This has implications, for example, for predictions about the propagation of shocks. We simulate output shocks using the revised network to show that formal sector data leads us to underestimate the impact of domestic output shocks in regions with high informality. Conversely, we may overstate the local output effects of international trade shocks in sectors and regions with a high incidence of informality. This is because formal sector data places more weight on formal firms with stronger links to international markets, when in fact the overall economy has much weaker ties to import markets.

An important question for future research, beyond the scope of this paper, is whether the observed spatial concentration of formal sector firm networks is a result of market frictions or a feature of structural transformation (Gollin, 2008). Understanding its drivers can inform policy recommendations about the optimal distribution of formal economic activity across space.

References

Adão, R., Carrillo, P., Costinot, A., Donaldson, D. and Pomeranz, D. (2022), 'Imports, exports, and earnings inequality: Measures of exposure and estimates of incidence', *The Quarterly Journal of Economics* **137**(3), 1553–1614.

Ades, A. F. and Glaeser, E. L. (1995), 'Trade and circuses: explaining urban giants', *The Quarterly Journal of Economics* **110**(1), 195–227.

Albert, C., Bustos, P. and Ponticelli, J. (2021), The effects of climate change on labor and capital reallocation, Technical report, National Bureau of Economic Research.

Alfaro-Ureña, A., Manelici, I. and Vasquez, J. P. (2022), 'The effects of joining multinational supply chains: New evidence from firm-to-firm linkages', *The Quarterly Journal of Economics* **137**(3), 1495–1552.

Almunia, M., Henning, D. J., Knebelmann, J., Nakyambadde, D. and Tian, L. (2023), *Leveraging Trading Networks to Improve Tax Compliance: Experimental Evidence from Uganda*, Centre for Economic Policy Research.

Amodio, F., Benveniste, E., Pham, H. and Sanfilippo, M. (2024), 'The local (informal) multiplier of industrial jobs'. Mimeo.

Antràs, P., Chor, D., Fally, T. and Hillberry, R. (2012), 'Measuring the upstreamness of production and trade flows', *American Economic Review* **102**(3), 412–16.

Arkolakis, C., Huneeus, F. and Miyauchi, Y. (2023), Spatial production networks, Technical report, National Bureau of Economic Research.

Atkin, D. and Donaldson, D. (2015), Who's getting globalized? the size and implications of intra-national trade costs, Working Paper 21439, National Bureau of Economic Research.

Atkin, D. and Khandelwal, A. K. (2020), 'How distortions alter the impacts of international trade in developing countries', *Annual Review of Economics* **12**(1), null.

Bacilieri, A., Borsos, A., Astudillo-Estevez, P. and Lafond, F. (2023), 'Firm-level production networks: what do we (really) know', *INET Oxford Working Paper* **2023**.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R. and Stroebel, J. (2021), 'International trade and social connectedness', *Journal of International Economics* **129**, 103418.

Baqaee, D. R. (2018), 'Cascading failures in production networks', *Econometrica* **86**(5), 1819–1838.

Bernard, A. B., Dhyne, E., Magerman, G., Manova, K. and Moxnes, A. (2022), 'The origins of firm heterogeneity: A production network approach', *Journal of Political Economy* **130**(7), 1765–1804.

Bernard, A. B. and Moxnes, A. (2018), 'Networks and trade', *Annual Review of Economics* **10**, 65–85.

Bernard, A. B., Moxnes, A. and Saito, Y. U. (2019), 'Production networks, geography, and firm performance', *Journal of Political Economy* **127**(2), 639–688.

Blanchard, P., Gollin, D. and Kirchberger, M. (2021), 'Perpetual motion: Human mobility and spatial frictions in three african countries', *CEPR Discussion Papers* **No. 16661**.

Böhme, M. H. and Thiele, R. (2014), 'Informal–formal linkages and informal enterprise performance in urban west africa', *The European Journal of Development Research* **26**, 473–489.

Boken, J., Gadenne, L., Nandi, T. and Santamaria, M. (2023), 'Community networks and trade', *CEPR Working Paper* **DP17787**.

Bramoullé, Y., Currarini, S., Jackson, M. O., Pin, P. and Rogers, B. W. (2012), 'Homophily and long-run integration in social networks', *Journal of Economic Theory* **147**(5), 1754–1786.

Brandt, N. (2011), 'Informality in mexico', *OECD Economics Department Working Papers* (896).

Bustos, P., Garber, G. and Ponticelli, J. (2020), 'Capital accumulation and structural transformation', *The Quarterly Journal of Economics* **135**(2), 1037–1094.

Carvalho, V. M., Nirei, M., Saito, Y. U. and Tahbaz-Salehi, A. (2021), 'Supply chain disruptions: Evidence from the great east japan earthquake', *The Quarterly Journal of Economics* **136**(2), 1255–1321.

Castro-Vincenzi, J., Khanna, G., Morales, N. and Pandalai-Nayar, N. (2024), Weathering the storm: Supply chains and climate risk, Technical report, National Bureau of Economic Research.

Chacha, P. W., Kirui, B. K. and Wiedemann, V. (2024), 'Supply chains in times of crisis: Evidence from kenya's production network', *World Development* **173**, 106363.

Chandrasekhar, A. (2016), 'Econometrics of network formation', *The Oxford Handbook of the Economics of Networks* pp. 303–357.

Chaney, T. (2014), 'The network structure of international trade', *American Economic Review* **104**(11), 3600–3634.

Cordaro, F., Fafchamps, M., Mayer, C., Meki, M., Quinn, S. and Roll, K. (2022), Microequity and mutuality: Experimental evidence on credit with performance-contingent repayment, Technical report, National Bureau of Economic Research.

De Paula, A. and Scheinkman, J. A. (2010), 'Value-added taxes, chain effects, and informality', *American Economic Journal: Macroeconomics* **2**(4), 195–221.

Demir, B., Javorcik, B. and Panigrahi, P. (2024), 'Breaking invisible barriers: Does fast internet improve access to input markets?', *CESifo Working Paper* **11567**.

Di Giovanni, J. and Levchenko, A. A. (2012), 'Country size, international trade, and aggregate fluctuations in granular economies', *Journal of Political Economy* **120**(6), 1083–1132.

Dix-Carneiro, R., Goldberg, P. K., Meghir, C. and Ulyssea, G. (2024), Trade and domestic distortions: The case of informality, Technical report.

Dix-Carneiro, R. and Kovak, B. K. (2019), 'Margins of labor market adjustment to trade', *Journal of International Economics* **117**, 125–142.

Eaton, J., Kortum, S. and Kramarz, F. (2011), 'An anatomy of international trade: Evidence from French firms', *Econometrica* **79**(5), 1453–1498.

Eaton, J., Kortum, S. S. and Kramarz, F. (2022), Firm-to-firm trade: Imports, exports, and the labor market, Technical report, National Bureau of Economic Research.

Elgin, C., Kose, M. A., Ohnsorge, F. and Yu, S. (2021), 'Understanding informality', *CERP Discussion Paper* **16497**.

Fafchamps, M. (2003), *Market institutions in sub-Saharan Africa: Theory and evidence*, MIT press.

Fan, T., Peters, M. and Zilibotti, F. (2023), 'Growing like india—the unequal effects of service-led growth', *Econometrica* **91**(4), 1457–1494.

Fujiy, B. C., Ghose, D. and Khanna, G. (2022), 'Production networks and firm-level elasticities of substitution', *STEG Working Paper Series* **WP027**.

Gabaix, X. (2009), 'Power laws in economics and finance', *Annu. Rev. Econ.* **1**(1), 255–294.

Gadenne, L., Nandi, T. K. and Rathelot, R. (2022), 'Taxation and supplier networks: Evidence from india', *Working Paper* .

Goldberg, P. K. and Reed, T. (2023), 'Demand-side constraints in development: The role of market size, trade, and (in)equality', *Econometrica* .

Gollin, D. (2008), 'Nobody's business but my own: Self-employment and small enterprise in economic development', *Journal of Monetary Economics* **55**(2), 219–233.

Grant, M. and Startz, M. (2022), Cutting out the middleman: The structure of chains of intermediation, Technical report, National Bureau of Economic Research.

Hansman, C., Hjort, J., León-Ciliotta, G. and Teachout, M. (2020), 'Vertical integration, supplier behavior, and quality upgrading among exporters', *Journal of Political Economy* **128**(9), 3570–3625.

Hassan, M. and Schneider, F. (2019), 'Size and development of the shadow economies of 157 countries worldwide: Updated and new measures from 1999 to 2013', *IZA Discussion Paper* **No. 10281**.

Herrendorf, B., Rogerson, R. and Valentinyi, A. (2022), New evidence on sectoral labor productivity: Implications for industrialization and development, Technical report, National Bureau of Economic Research.

Huneeus, F. (2018), 'Production network dynamics and the propagation of shocks', *Working Paper* .

Jackson, M. O. and Rogers, B. W. (2007), 'Meeting strangers and friends of friends: How random are social networks?', *American Economic Review* **97**(3), 890–915.

Jefferson, M. (1939), 'The law of the primate city', *Geographical Review* **29**(2), 226–232.

Jefferson, M. (1989), 'Why geography? The law of the primate city', *Geographical Review* **79**(2), 226–232.

Klenow, P. J. and Rodriguez-Clare, A. (1997), 'The neoclassical revival in growth economics: Has it gone too far?', *NBER macroeconomics annual* **12**, 73–103.

KNBS (2010), 'Basic Report on the 2010 Census of Industrial Production'.

KNBS (2016), Micro, Small and Medium Establishment (MSME) Survey: Basic Report2016, Technical report, Kenya National Bureau of Statistics.

KNBS (2017), Report on the 2017 Kenya Census of Establishments (CoE), Technical report, Kenya National Bureau of Statistics.

KNBS (2019), 2019 Kenya Population and Housing Census: Volume I, Technical report, Kenya National Bureau of Statistics.

KNBS (2022), Gross county product 2021 (gcp), Technical report, KNBS.
URL: *https://www.knbs.or.ke/reports/kenya-gross-county-product-2021/*

Kreindler, G. E. and Miyauchi, Y. (2023), 'Measuring commuting and economic activity inside cities with cell phone records', *Review of Economics and Statistics* **105**(4), 899–909.

Kumar, K., Rajan, R. and Zingales, L. (1999), What determines firm size?, Technical report, NBER Working Paper.

La Porta, R. and Shleifer, A. (2014), 'Informality and development', *Journal of Economic Perspectives* **28**(3), 109–26.

Laeven, L. and Woodruff, C. (2007), 'The quality of the legal system, firm ownership, and firm size', *The Review of Economics and Statistics* **89**(4), 601–614.

Lagakos, D. and Shu, M. (2023), 'The role of micro data in understanding structural transformation', *Oxford Development Studies* **51**(4), 436–454.

Manelici, I., Vasquez, J. P. and Zárate, R. D. (2024), 'The gains from foreign multinationals in an economy with distortions', *Mimeo* .

McCaig, B. and Pavcnik, N. (2015), 'Informal employment in a growing and globalizing low-income country', *American Economic Review* **105**(5), 545–550.

McCaig, B. and Pavcnik, N. (2018), 'Export markets and labor allocation in a low-income country', *American Economic Review* **108**(7), 1899–1941.

Meagher, K. (2013), 'Unlocking the informal economy: A literature review on linkages between formal and informal economies in developing countries', *Work. ePap* **27**, 1755–1315.

Memon, P. A. (1976), 'Urban primacy in kenya', *IDS Working Paper Series, University of Nairobi* **282**.

Miyauchi, Y. (2024), 'Matching and agglomeration: Theory and evidence from japanese firm-to-firm trade', *Econometrica* **92**(6), 1869–1905.

MoF (1963), Survey of distribution (1960), Technical report, Republic of Kenya, Ministry of Finance, Republic of Kenya.

Naritomi, J. (2019), 'Consumers as tax auditors', *American Economic Review* **109**(9), 3031–72.

Newman, M. E. (2006), 'Modularity and community structure in networks', *Proceedings of the national academy of sciences* **103**(23), 8577–8582.

Obudho, R. A. (1997), 'Nairobi: National capital and regional hub', *The urban challenge in Africa: Growth and management of its large cities* pp. 292–334.

Panigrahi, P. (2022), 'Endogenous spatial production networks: Quantitative implications for trade and productivity', *Working Paper* .

Pomeranz, D. (2015), 'No taxation without information: Deterrence and self-enforcement in the value added tax', *American Economic Review* **105**(8), 2539–69.

Sargent, T. J. and Stachurski, J. (2022), 'Economic networks: Theory and computation', *arXiv preprint arXiv:2203.11972* .

Schneider, F. and Enste, D. H. (2000), 'Shadow economies: Size, causes, and consequences', *Journal of Economic Literature* **38**(1), 77–114.

Soo, K. T. (2005), 'Zipf's law for cities: a cross-country investigation', *Regional Science and Urban Economics* **35**(3), 239–263.

Startz, M. (2021), 'The value of face-to-face: Search and contracting problems in nigerian trade', *Working Paper* .

Storeygard, A. (2016), 'Farther on down the road: transport costs, trade and urban growth in sub-saharan africa', *The Review of Economic Studies* **83**(3), 1263–1295.

Topalova, P. (2010), 'Factor immobility and regional impacts of trade liberalization: Evidence on poverty from india', *American Economic Journal: Applied Economics* **2**(4), 1–41.

Ulyssea, G. (2018), 'Firms, informality, and development: Theory and evidence from brazil', *American Economic Review* **108**(8), 2015–47.

Zárate, R. D. (2022), *Spatial misallocation, informality, and transit improvements: Evidence from mexico city*, The World Bank.

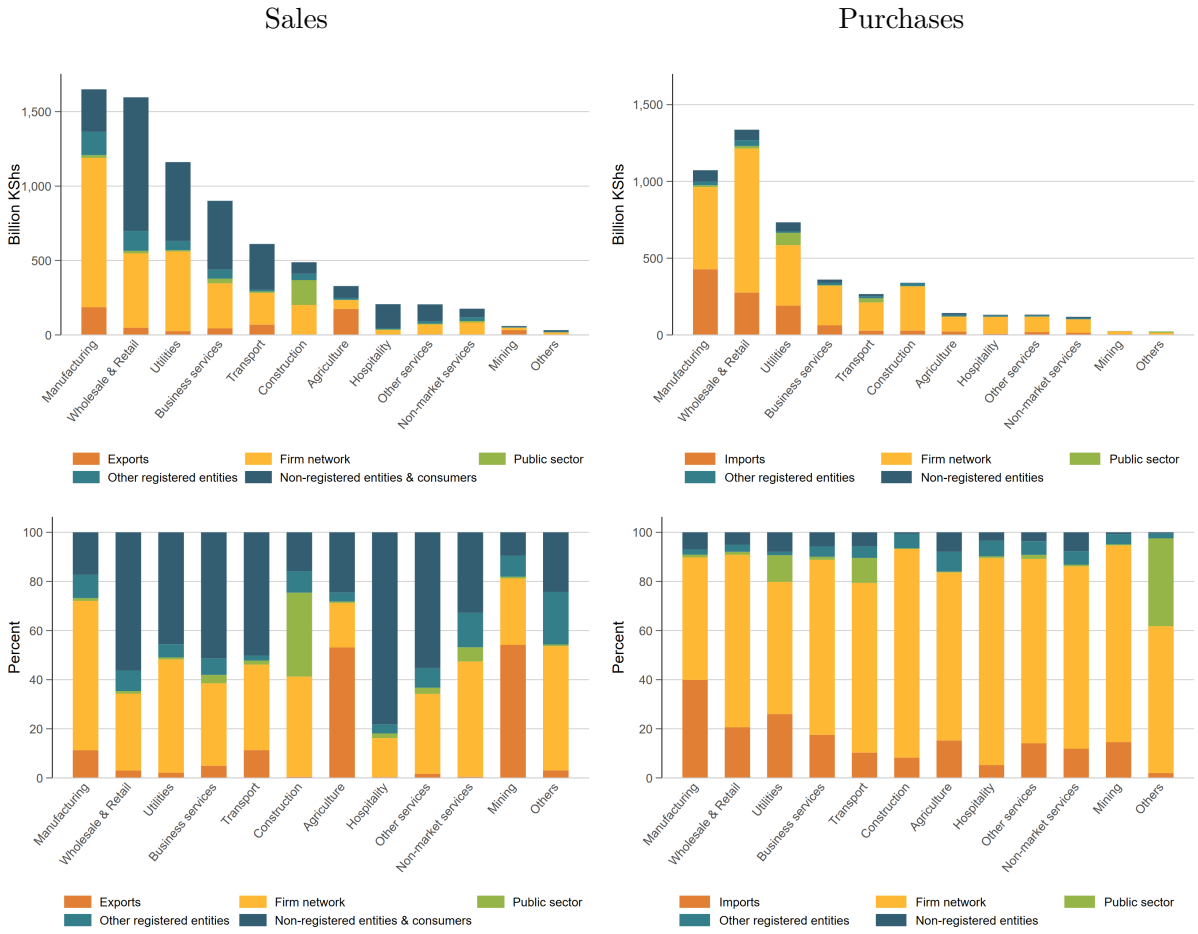
Zhou, Y. (2022), The value added tax, cascading sales tax, and informality, *in* M. Bussolo and S. Sharma, eds, 'Hidden Potential: Rethinking Informality in South Asia', World Bank Publications, chapter The Value Added Tax, Cascading Sales Tax, and Informality, pp. p. 61–90.

Appendix

A Material for Data and Empirical Section

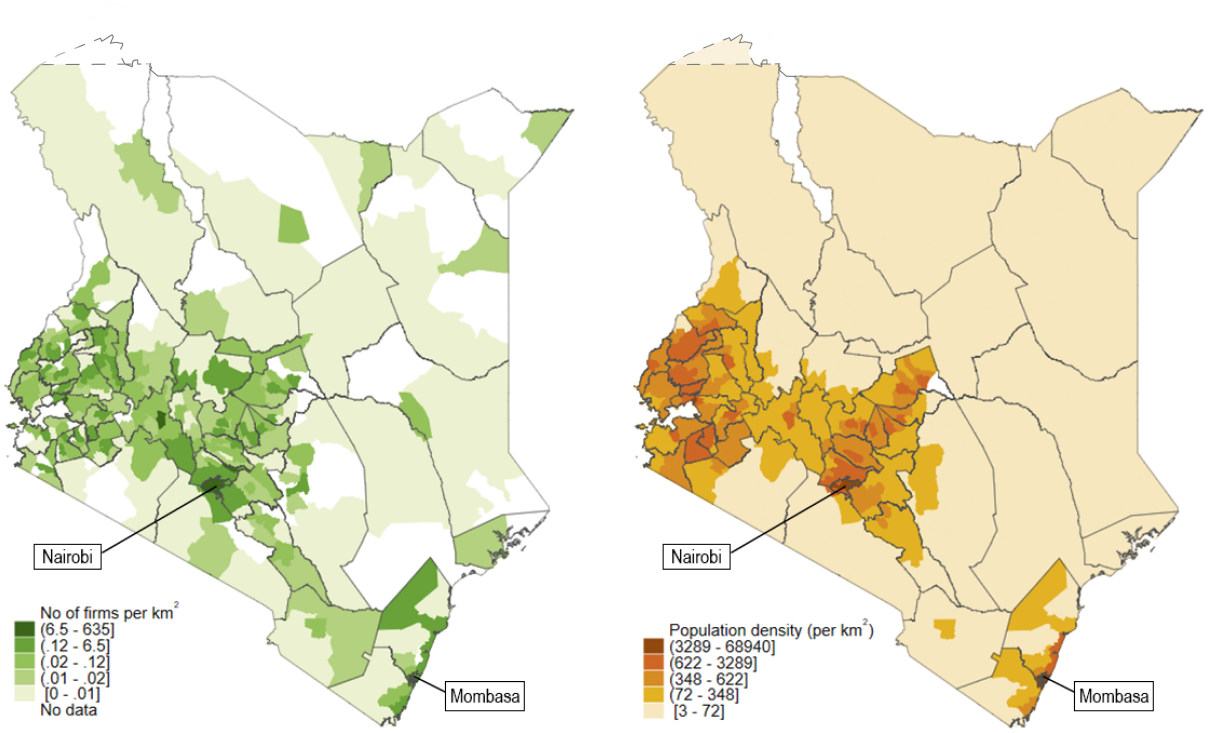
A.1 Supplementary Graphs and Tables

Figure A1: Composition of sales and purchases by sector



The figures in the first row show sector-level aggregate sales (domestic + exports) and purchases (domestic + imports) for 2019. In the second row, we plot the share of each buyer and supplier type as a percentage of total sector-level sales and purchases.

Figure A2: Firm headquarter locations and population density



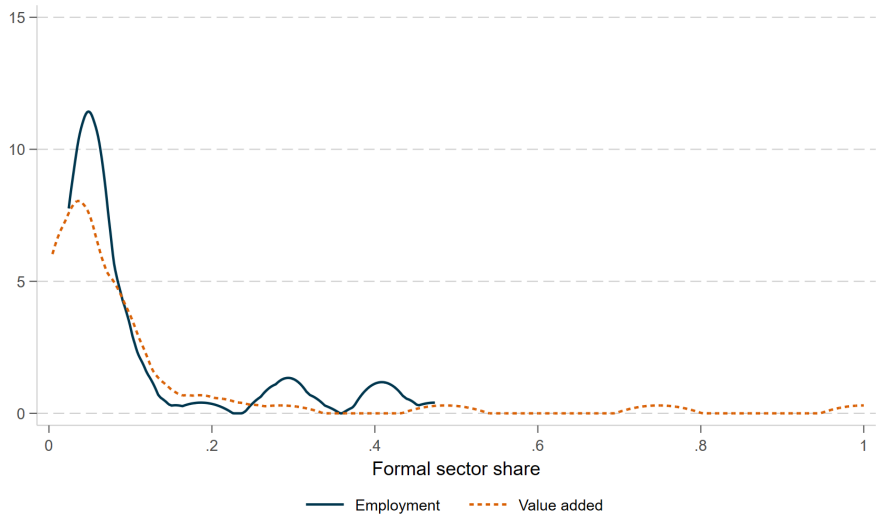
The left map shows the density of firm headquarter locations at the sub-county level, i.e. the number of firms per km^2. The right map shows the population density - also at the sub-county level. The borders of Kenya's 47 counties, the first administrative layer, are outlined in grey.

Table A1: Firms in counties with a higher informal sector share have fewer links in the administrative data

	total		mean		median		90th percentile		final demand
	buyers	suppliers	buyers	suppliers	buyers	suppliers	buyers	suppliers	%
Formal sector share (sector-county, %)	0.043***	0.037***	0.016**	0.009*	0.011**	0.007*	0.019**	0.011*	-0.166
	(0.014)	(0.012)	(0.007)	(0.005)	(0.005)	(0.004)	(0.008)	(0.006)	(0.181)
Population	1.559***	1.415***	0.441***	0.282***	-0.180	0.072	0.543***	0.455***	-5.515*
	(0.294)	(0.266)	(0.138)	(0.090)	(0.132)	(0.088)	(0.140)	(0.117)	(2.763)
Travel time to Nairobi	-0.699***	-0.591***	-0.301**	-0.170**	0.007	-0.139**	-0.318**	-0.225**	0.424
	(0.243)	(0.178)	(0.128)	(0.078)	(0.083)	(0.067)	(0.132)	(0.098)	(2.200)
Travel time to Mombasa	-0.552**	-0.449**	-0.326**	-0.246***	-0.164	-0.282***	-0.378***	-0.262***	6.256***
	(0.244)	(0.176)	(0.132)	(0.083)	(0.127)	(0.084)	(0.137)	(0.095)	(1.998)
No. observations	450	472	450	472	379	471	450	472	470
R2	0.469	0.540	0.400	0.326	0.266	0.315	0.408	0.307	0.242
Sector FE	✓	✓	✓	✓	✓	✓	✓	✓	✓

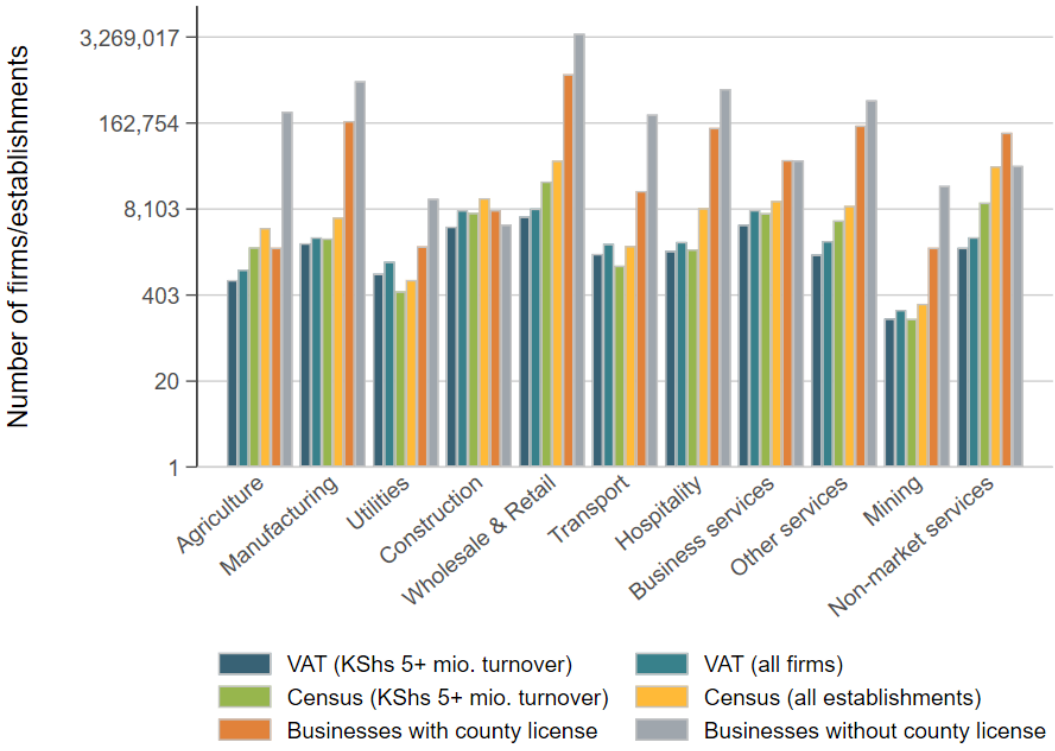
In this table we regress the number of firm-to-firm links aggregated at the sector and county level on informal sector employment shares from the population census, which we observe at the same level of disaggregation. In the last column we regress the share of sales to non-registered entities (consumers or firms outside the VAT system) on the formal sector share. Standard errors are clustered at the county level.

Figure A3: County-level formal sector shares



The above graphs plot the probability density function (pdf) for the dispersion of formal sector shares across Kenya's 47 counties. The value added-based measure relies on the difference between county-sector-level national accounts and the administrative data to obtain formal sector shares. For the employment-based measure we rely on information about formal and informal employment at the sector and county level from the 2019 census.

Figure A5: The extensive margins of informality - in which sectors do informal firms operate?

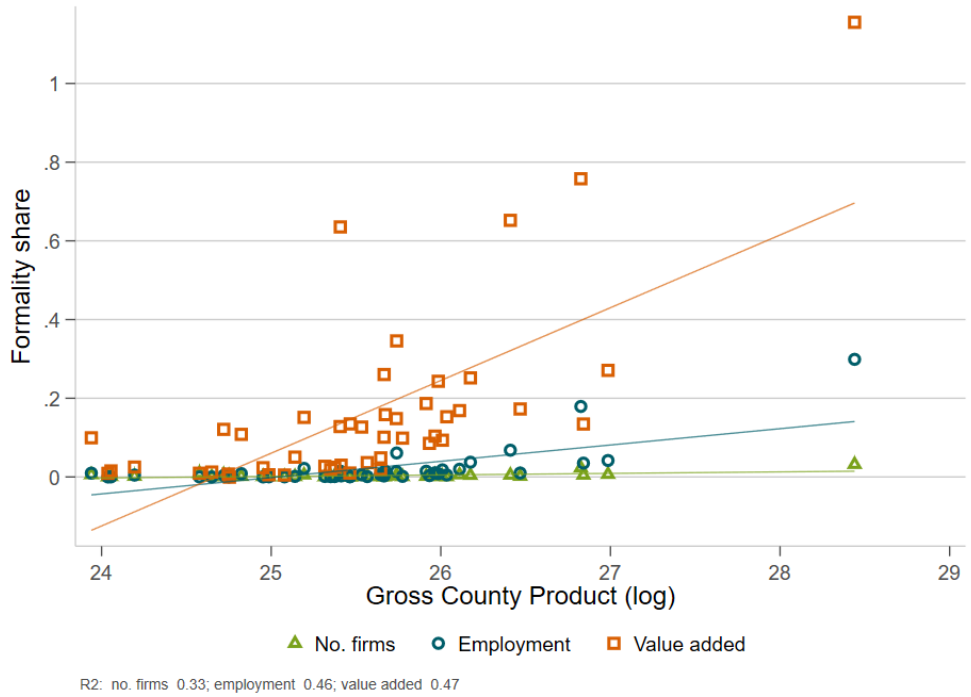


The graph compares the number of firms covered in the administrative data and with an annual revenue of over KShs 5 million in 2016 to the number of firms with annual revenues above KShs 5 million in the 2016 Census of Establishments (CoE) (KNBS, 2017) as well as any firm captured in either data set. Further, it plots the of licensed and unlicensed businesses reported by KNBS in KNBS (2016).

Figure A4: Informality, market size, and income levels

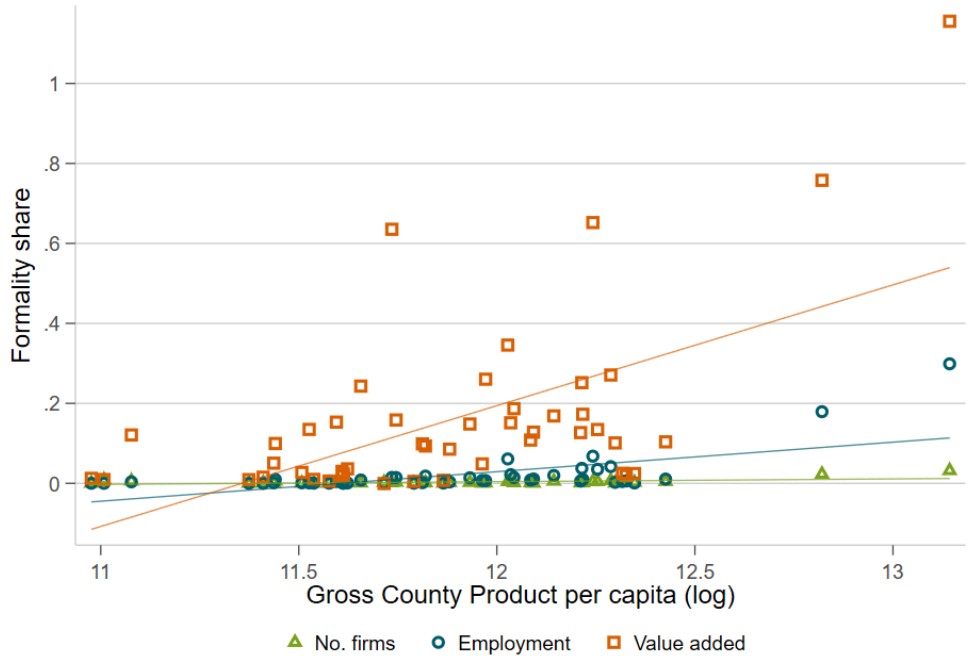
Correlation of the formal sector share and ...

... Gross County Product



R2: no. firms 0.33; employment 0.46; value added 0.47

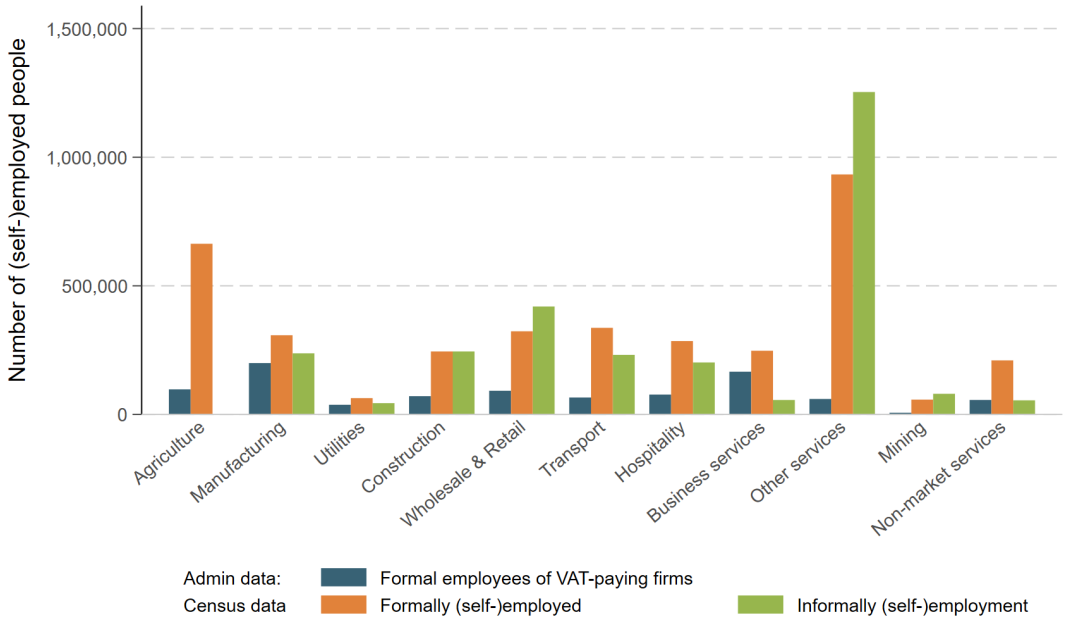
... Gross County Product per capita



R2: no. firms 0.28; employment 0.40; value added 0.34

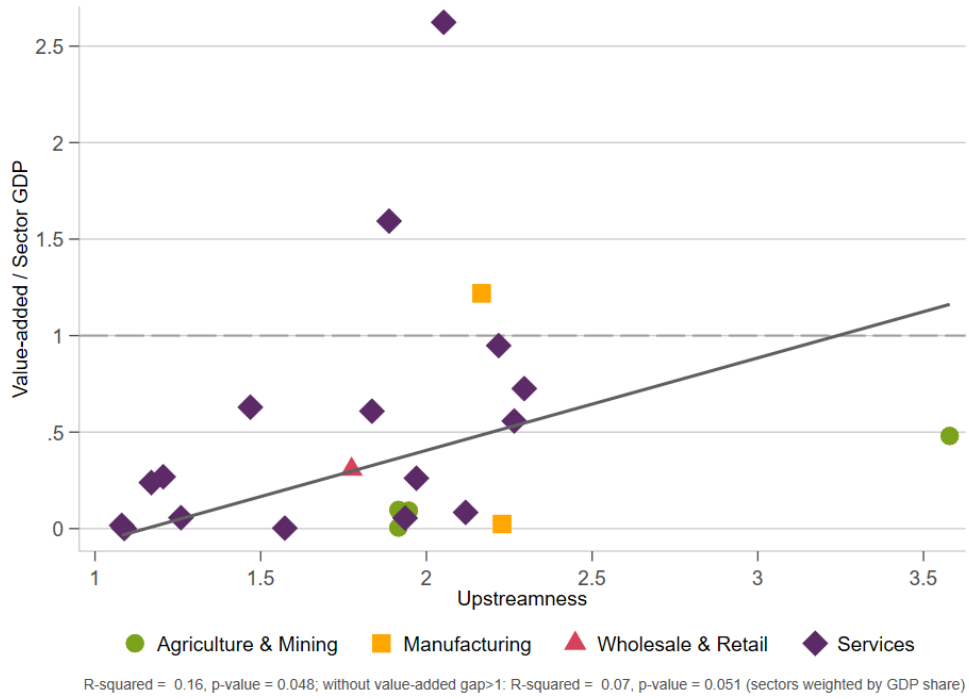
The two graphs plot the correlation of the formal sector share with the Gross County Product in absolute and per capita terms respectively. Each marker represents one of Kenya's 47 counties.

Figure A6: Formal and informal employment in private enterprises



The graph compares the number of formal employees employed in VAT-paying firms to the number people who stated they are formally or informally employed in a private sector entity. To improve readability we omit the bar for informal employment in the agricultural sector. As per 2019 population census over 11 million people are informally employed in the agricultural sector.

Figure A7: The GDP/value-added gap and upstreamness



We plot the gap between value-added in the VAT and national accounts figures at the sub-sector level for the most granular sector classification reported in national accounts. We correlate it with a measure of upstreamness (Antràs et al., 2012), which captures how removed a sector is from final consumers (it takes a value of one if the sector sells everything directly to final consumers).

A.2 Measures of Informality

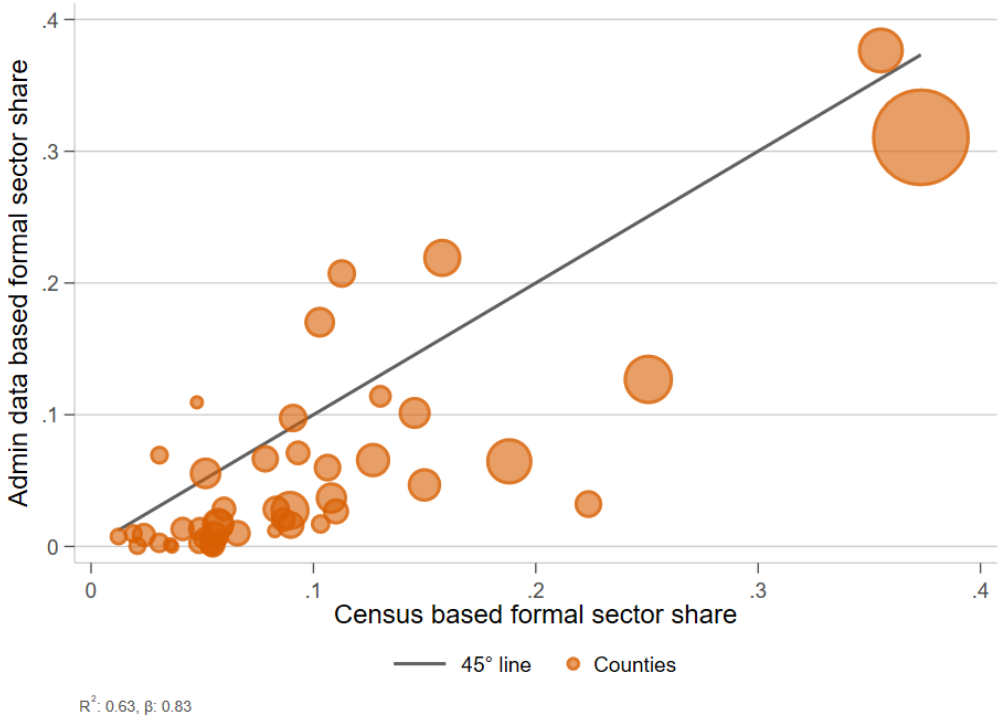
As documented in Table A2, the two employment-based KNBS measures correlate well with all measures based on the administrative data. The measure capturing licensed businesses as a share of the universe of businesses in Kenya (including micro-enterprises) in contrast only correlates weakly with them. This likely reflects the fact that many of the licensed firms are very small themselves and their geographic dispersion does not correlate as strongly with the tax records. Any employment in licensed businesses (second row) is likely concentrated in larger licensed firms, which is why the employment based measure aligns more strongly with the administrative data relative to the simple firm count.

Table A2: Correlation of formality measures

	Formality measures based on admin data		
KNBS measures	No. firms	Employment	Value added
Employment (census)	0.78	0.83	0.78
Employment (licensed MSMEs)	0.58	0.69	0.62
No. firms (licensed)	0.20	0.16	0.11

The above table shows the correlation coefficients of different measures of the formal sector share. Each measure represents a share, i.e. captures the proportion of economic activity that can be attributed to the formal sector. The labels indicate the underlying unit of measurement and the source of the data. All measures are aggregated at the county level.

Figure A8: Comparison of formal sector shares based on census versus administrative records



The above graph correlates the share of the formal sector computed using employment figures from the administrative records with the share of formal private sector employment as per the 2019 population census (KNBS, 2019). Each market represents a county. The size of each marker is proportional to the economic size of the county, i.e. its Gross County Product. To avoid mechanical correlation between the two measures we use total employment in licensed firms as the denominator for the administrative data. The KNBS estimate for employment in licensed firms is based on micro data that is distinct from the population census. Alternatively, one could use total employment in all MSMEs, which, however, includes many self-employed people. The correlation results are very similar for both alternatives.

A.3 The VAT-Paying Sector as a Share of GDP

The most relevant sector that is not well captured in the VAT data is agriculture, which generates 21%-23% of Kenya's GDP. While part of the sector receives special tax treatment due to exemptions of mainly unprocessed agricultural commodities, some of the GDP gap can also be attributed to informality in the classic sense due to the prevalence of small holders in the sector. Non-market services include education, health, public administration, and real estate (Herrendorf et al., 2022). They contribute 22% to Kenya's GDP, but are barely represented in the VAT data as most of the entities operating in these sectors are VAT exempt, not-for-profit, or the underlying sector's size in the national accounts is estimated using non-market prices (see penultimate column of Table A3). Figure 4 highlights another sizeable gap for "others", which includes international organisations, unclassified firms, and financial services.

Table A3 illustrates that the value added generated by the VAT sector has been declining over time as a proportion of GDP. This downward trend in value added can be attributed to two factors. First, the introduction of a fuel tax in September 2018, which was previously VAT

exempt, has led to a reduction in value added. The impact of this tax is particularly relevant for the utilities sector. However, this sector alone cannot fully explain the overall downward trend and kink in the data. Second, sectors that have significantly contributed to Kenya's growth over the years, such as agriculture, real estate, financial services, and public administration, are not well captured in the VAT data.

Table A3: Share of GDP covered in the administrative records

	Share of GDP (%)						
Year	All	ex Fin.	ex NMS+Fin.	ex Agri.	ex NMS+Fin.+Agri.	NMS	Agri.
2015	36	39	50	42	66	22	21
2016	40	43	56	46	73	22	21
2017	37	40	52	45	71	22	22
2018	37	40	52	45	70	22	22
2019	28	30	39	34	53	22	23

The mid-section of the above table reports the share of GDP captured by the VAT data sequentially excluding (*ex*) specific sectors. Fin. refers to financial services. NMS refers to non-market services, i.e. education, health, public administration, and real estate (Herrendorf et al., 2022). Agri. refers to the agricultural sector. The first five data columns report the proportion of GDP captured by value added of the VAT-paying firms. The final two columns report the GDP share of non-market services and agriculture respectively. GDP figures are based on national accounts data (in current prices) published by the Kenya National Bureau of Statistics.

A.4 Firm Location and Relationships Drive Spatial Concentration in Trade Flows

The extensive margins of the firm network, firm location, and firm-to-firm relationships, account for 70%-90% of the variation in aggregate trade volumes. Using transaction-level data, we are able to distinguish between four different sales margins: the number of firms, the number of relationships with buyers per firm, the number of transactions per relationship, and the average trade volume per transaction. The same is true for purchases. Table A4 summarises the share of the variance attributed to each term in both upstream (purchases) and downstream (sales) trade flows.[34] The number of firms operating in each county accounts for as much as 67% of the variance in purchases across counties. This includes purchases the firms make within their own county and what they buy outside the county. The number of supplier relationships other counties have with the county accounts for yet another 22%. This leaves a little over 10% of the variance to be picked up by the intensive margins for trade, i.e. the number of transactions between firm pairs and the average transaction volume. Turning to downstream trade flows, i.e. the decomposition of the variance in sales across (sub-)counties, the location of firms plays a slightly less important role. Instead the number of firm-to-firm relationships now accounts for one third of the variance in network sales.

[34] Our decomposition follows Klenow and Rodriguez-Clare (1997); Eaton et al. (2011); Panigrahi (2022).

Table A4: Geographic concentration of economic activity in Kenya

Purchases

Aggregation	No. firms	No. relationships/firm	No. transactions/relation	Avg. volume/transaction
County	0.67	0.22	0.14	-0.04
Subcounty	0.53	0.29	0.16	0.06

Sales

Aggregation	No. firms	No. relationships/firm	No. transactions/relation	Avg. volume/transaction
County	0.60	0.31	0.12	-0.00
Subcounty	0.39	0.34	0.15	0.16

A.5 Spatial Concentration of Economic Activity and Multi-Establishment Firms

A potential concern with the VAT data is that it may overstate spatial concentration because firms are only required to report their headquarters' locations, which are often situated in major cities like Nairobi or Mombasa. To assess the sensitivity of measures of spatial concentration to multi-establishment firms, we use micro-data from the 2010 Census of Industrial Production (KNBS, 2010), which includes the mining, manufacturing, and utilities sectors. We compare the spatial distribution of sales and firm locations for all firms, including those with multiple branches, to that of single-establishment firms in Table A5. Firms covered in the Census of Industrial Production overlap closely with the group of VAT-paying firms we observe in the tax records. A 1:1 mapping is not possible due to the anonymous nature of the data sets. However, the overall number of industrial firms observed in each of the two data sources aligns closely. In 2015, we observed 4,064 VAT-paying firms[35] in mining, manufacturing, and utilities, while KNBS (2010) covered 2,252 firms five years earlier.

Of all firms involved in industrial production, 48% are located in Nairobi County generating as much as 61% of total sales in 2010.[36] When we limit the data from the Census of Industrial Production to single establishments, the overall concentration of firm locations does not change. The concentration becomes even slightly more unequal once we consider sales instead of purely counting the number of firms. We, however, overstate the concentration of sales in Nairobi by six percentage points if multi-establishment firms are in the sample and their sales are aggregated geographically based on headquarter information only (i.e. the measure we obtain from the VAT data by default). Despite this, the discrepancy is not large enough to fully account for the

[35]The earliest year for which the VAT records have been fully digitised is 2015. A later Census of Industrial Production is available for 2018. However, the data set published by KNBS does not include any information on firm locations. Further, information on sales is missing for over half of the firms.

[36]The figures for Kenya are similar to the concentration of formal manufacturing firms reported by Storeygard (2016) for Tanzania. Dar es Salaam, Tanzania's primate city, accounts for 8% of its population (Storeygard, 2016) - a very similar figure to Nairobi's population share in Kenya (KNBS, 2019).

higher spatial concentration observed among VAT-registered firms compared to overall economic activity.

Table A5: Geographic concentration of industrial activity

	All firms		**Single est. firms**	
	Nairobi (%)	α	Nairobi (%)	α
Census of Industrial Production (2010)				
		N = 2252		
No. firms	48	0.54	48	0.54
Sales	61	0.32	55	0.30
Industrial firms in admin data (2015)				
		N = 4064		
No. firms	64	0.50	-	-
Sales	69	0.21	-	-

The columns for Nairobi report their share of the respective national aggregate figures (e.g., the share of industrial establishments located in Nairobi). α is the estimated coefficient from a county-level rank regression of each county's rank (log) on the respective measure x (log): $\log \text{rank} = \log A - \alpha \log x$. The Census of Industrial Production was carried out by KNBS (2010).

B Material for Model Setup and Estimation

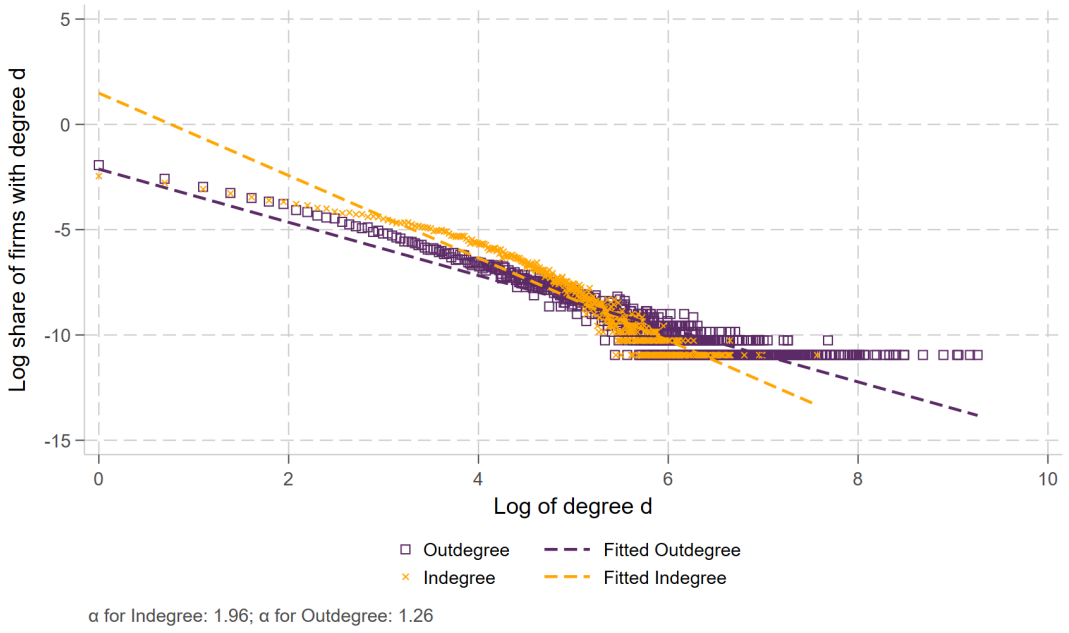
B.1 Supplementary Graphs and Tables

Table B1: Linking patterns of small buyers

	Manufacturing	Wholesale	Retail	Nairobi	Mombasa	Same county	Bigger supplier	Final demand
Small buyer	-0.023***	0.011	0.038***	-0.046***	-0.003	0.037***	0.002	0.030*
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.02)
No. observations	892	892	892	892	892	892	892	850
R2	0.585	0.593	0.568	0.721	0.860	0.872	0.477	0.637
Sector-county FE	✓	✓	✓	✓	✓	✓	✓	✓

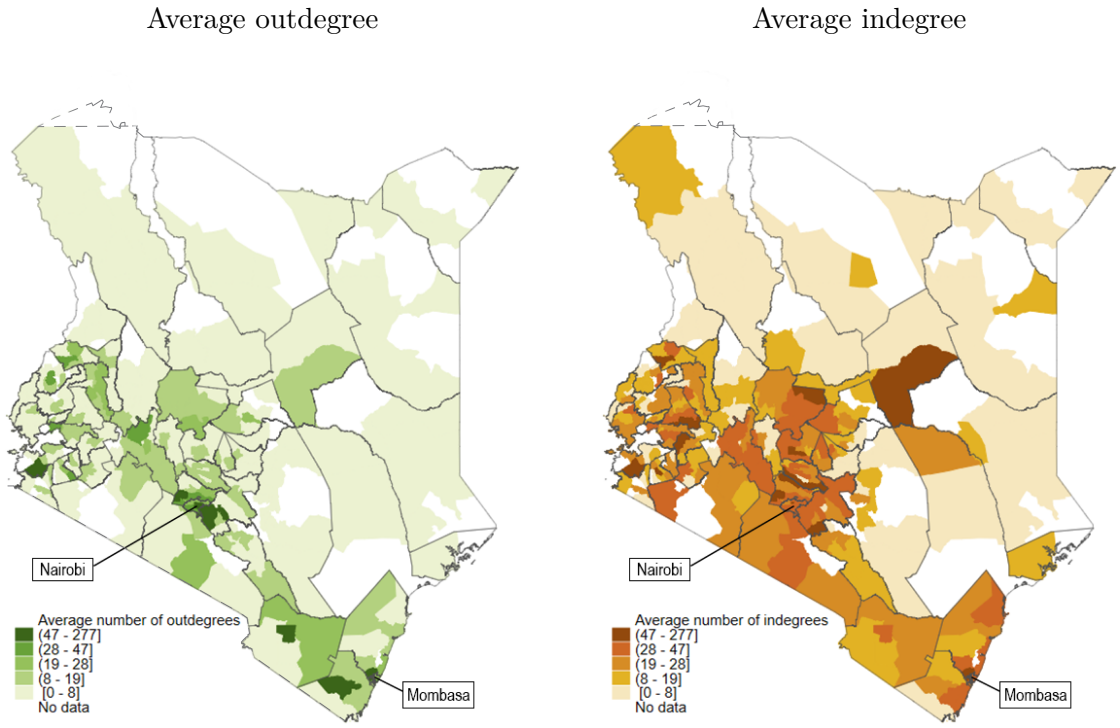
We group firms by sector, county, and size. Small firms represent the bottom sales quartile of a sector and county. We then compute the share of overall links the firm has with another sector-county-size group. We then aggregate the share for suppliers with specific characteristics (e.g. any wholesaler, irrespective of location) for each type of buyer (sector-county-size). The column titles list the characteristics of the suppliers. Finally, we regress the respective sum of shares on whether or not the buyer type is a small buyer type.

Figure B1: Degree distributions



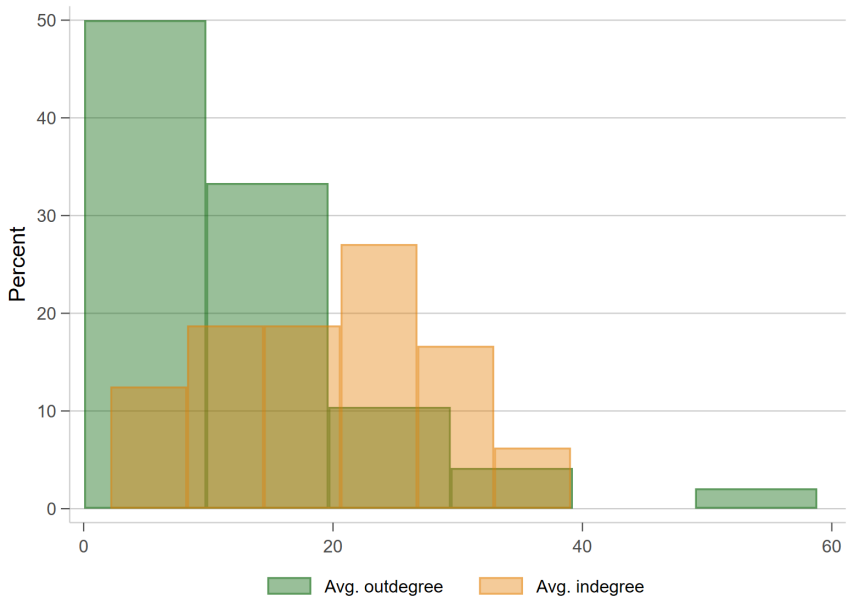
The figure plots the log-log plot of the probability density function (pdf) against firm outdegree and indegree respectively. The coefficients α shown at the bottom of the plot correspond to the power law exponent indicating the existence of a thicker tail for the outdegree distribution.

Figure B2: Average in- and outdegrees across space



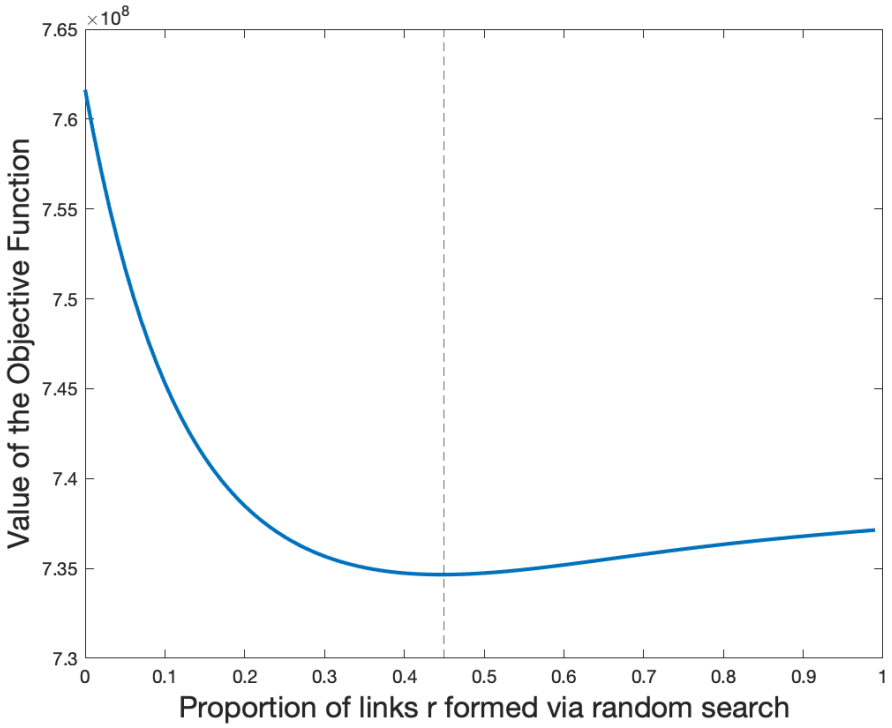
The above map plots the average in- and outdegree of firms for each sub-county. The borders of Kenya's 47 counties, the first administrative layer, are outlined in grey.

Figure B3: County-level average in- and outdegree



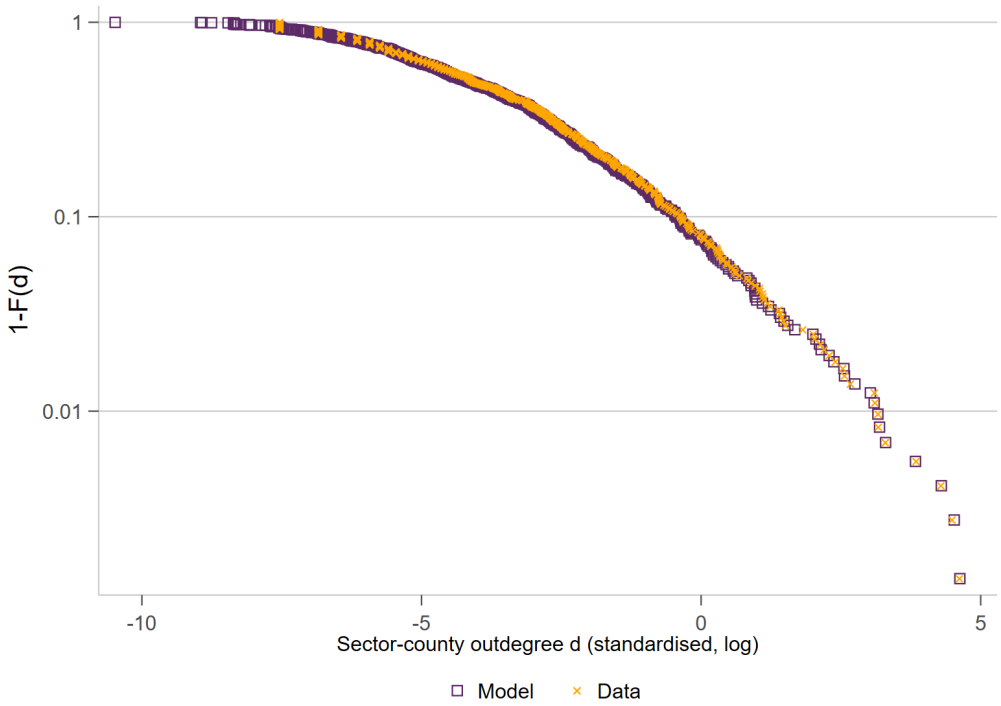
The histogram plots the average in- and outdegree across firms in each county.

Figure B4: Objective function for various values r



The graph plots the sum of the squared difference between each element of the model predicted interaction matrix π and the matrix π directly observed in the data, for various values of the parameter $r \in [0, 1]$. The figure shows that $r^* = 0.45$ obtained via simulated annealing minimises the objective function.

Figure B5: Model Fit - actual and predicted outdegree distribution



The figure plots inverse CDF for the actual and model-predicted total outdegree for each type (i.e. sector-county-size cell). The number of outdegrees is standardised. Note the log scale on both the x- and the y-axis.

B.2 Constructing Type-by-Type Matrix $\pi_{model}(\theta, \theta'; r)$

Each iteration of the estimation requires two additional steps to construct π_t.

First, based on Bramoullé et al. (2012)'s formula, predicting π_t requires us to compute a geometric series of matrix $\mathbf{B}$. For ease of computation, we restrict this to the first five entries of the geometric series as subsequent matrix entries become negligible.

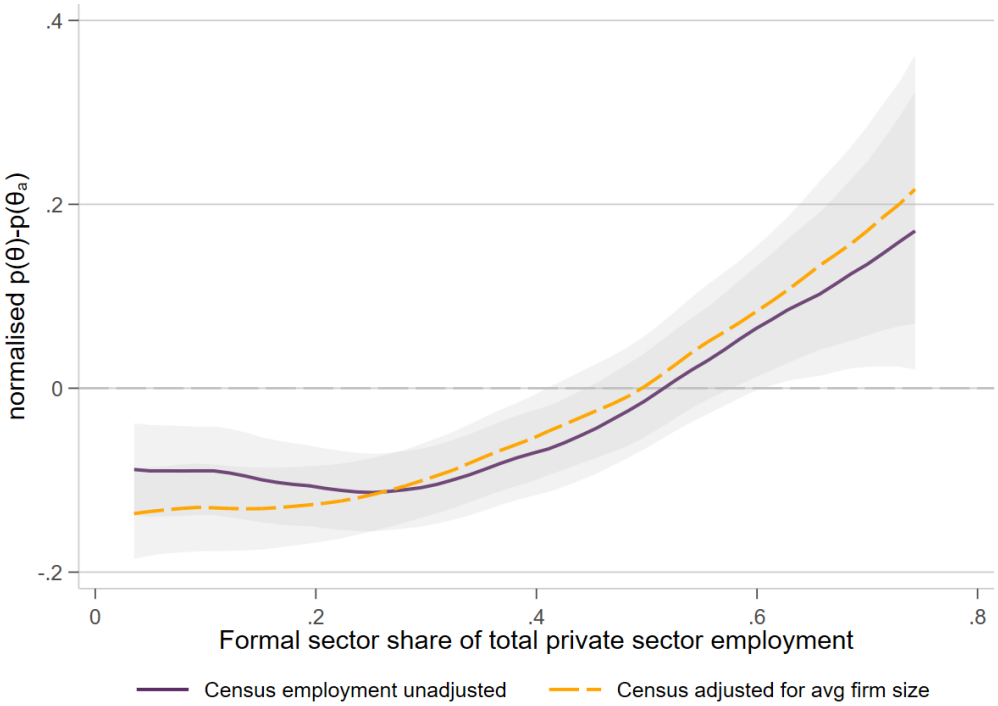
Second, note from Equation 1 that $\pi^t_{t_0}$ represents only the expected outdegree of types born in t_0 evaluated at time t. Since new firms are born in every period up until period t, we need to aggregate these matrices across all time periods leading up to t to obtain the type-by-type adjacency matrix of the entire network. The matrix of type-by-type links at time t is given by $\pi_t = \sum_{t_0}^{t} (\mathbf{p} \cdot \pi^t_{t0}{}')'$ where $\mathbf{p}$ is a column vector containing the probabilities for each type to be born. We compute the probability that a node of a certain type is born in time t_0 and its expected links in time t with every other type to get $\mathbf{p} \cdot \pi^t_{t0}$. We then repeat this process to compute the probability that a node of a certain type is born in time t_{0+1} and its expected degree in time t to get $\mathbf{p} \cdot \pi^t_{t0+1}$. We have to undertake this exercise for all time periods leading up to t. In other words, we need to compute t such matrices and add them up to obtain the type-by-type degree distribution at time t.

Computing π^t for t= 56822 in each iteration while looping through different candidate values of r is computationally intensive. Therefore, in every iteration, we compute π^t_{t0} for 500 'representative' time periods that we then aggregate to obtain π_t. We space these 500 periods equally between our first period $t_0 = 1$ and final period $t_0 = 56822$. As a result, we compute π_t= $\sum_{t_0=1:100:56822}(\mathbf{p} \cdot \pi^t_{t0}{}')'$. This implies that the network is scaled down in terms of firm count. However, this approach ensures that we do not disproportionately sample from either older or younger nodes and thereby bias our results. For example, sampling from nodes born in the first 500 periods would lead us to predict the type-by-type outdegree distribution only for firms in the right tail of the firm degree distribution if the observed network happens to exhibit preferential attachment, since preferential attachment results in older nodes having a higher chance of being more connected. This can bias our estimation of r as we will match the predicted distribution of such 'older' firms with all firms observed in the data. Our sampling strategy prevents this by ensuring a balanced representation of firms across different time periods and ensures that the essential features of the network formation process and network structure remain intact.

C Material on the Revised Network

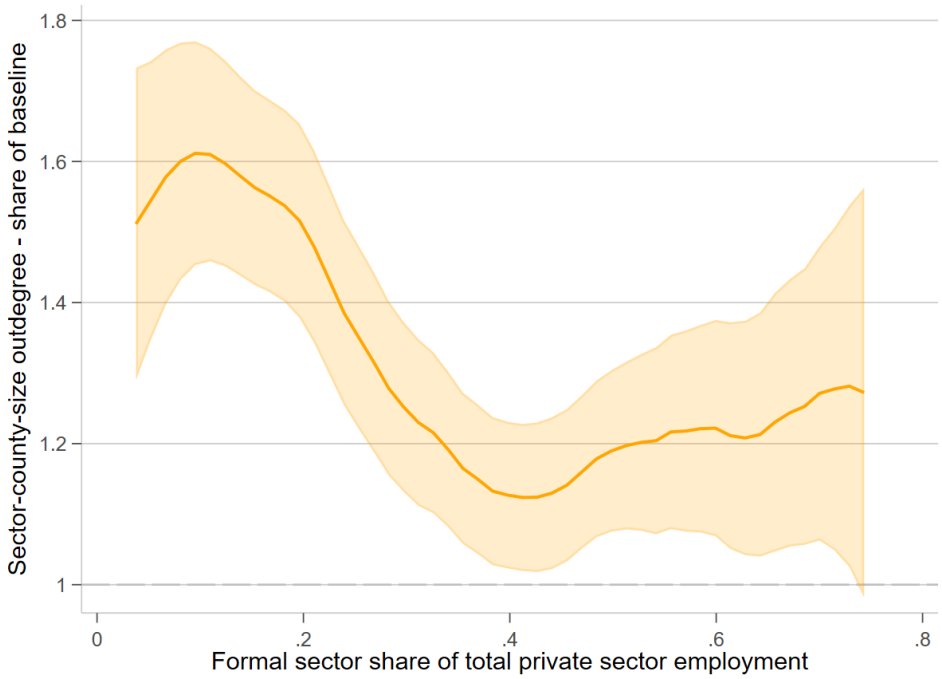
C.1 Supplementary Material for Revised Network

Figure C1: Sector-county-size probabilities and formal sector shares



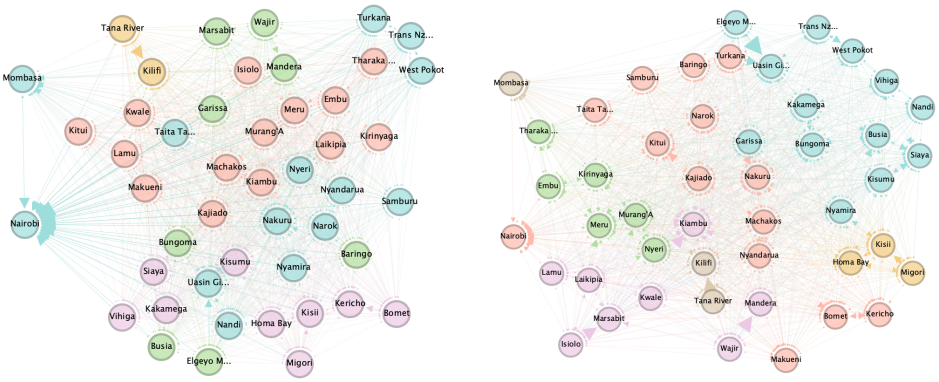
The graph plots each sector-regions formality share against the normalised difference between the baseline $p(\theta)$ and the augmented version $p(\theta_a)$ that takes into account informal firms. $p(\theta)$-$p(\theta_a)$ is reported in terms of standard deviations. A 10 percentage point increase in formality leads to an increase of $p(\theta)$-$p(\theta_a)$ by half a percentage point (0.35 standard deviations). To estimate the slope, we exclude eleven sector-county-size types which are adjusted by more than two standard deviations. All of the eleven types are Nairobi-based. Nine are large types, plus small firms in business services and construction. The slope becomes a little more than twice as steep if the five sector-county pairs are included.

Figure C2: Predicted change in type-level outdegree and formal sector shares



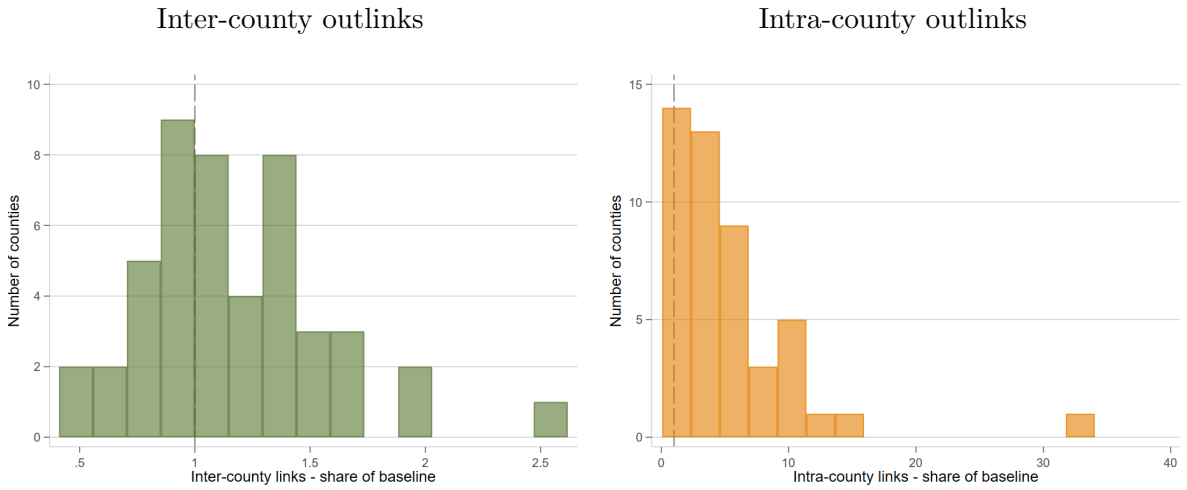
The figure plots sector-county formal sector shares against changes in type-level outdegree. The change in outdegree is measured as the difference between the revised network (including informal firms) and the baseline network (formal firms only), expressed relative to the baseline.

Figure C3: Model versus revised county-level network



The left and right panels show the baseline and revised county-level networks, respectively. We use the row-normalised county-level adjacency matrix to construct the plots with arrows indicating links from suppliers to buyers. Colours indicate county groupings identified by a community detection algorithm.

Figure C4: Inter- and intra-county trade patterns in a revised network



The figure plots the ratio of the supplier-to-buyer links for the revised network relative to the baseline for each county, distinguishing between trade links between counties (inter) and within counties (intra). To the left of the dotted line, at a value of one, are counties that experienced a decline in outlinks in the respective type of trade linkages.

Table C1: County-level changes in the dispersion of outdegrees - alternative scenarios

Scenario		Δ sd/mean (in %)
Informal firms ≈ small firms		
Default	all counties	-7.5
	w/o NBO & MSA	-18.0
$p(\theta)$ using employ. share only	all counties	-5.6
	w/o NBO & MSA	-11.8
Alternative linking patterns for informal firms		
0% sales, 50% input to/from formal	all counties	-16.0
	w/o NBO & MSA	-33.5
0% sales, 75% input to/from formal	all counties	-11.1
	w/o NBO & MSA	-25.7
25% sales, 50% input to/from formal	all counties	-16.1
	w/o NBO & MSA	-33.7

The above table reports the difference in outdegrees between the original and the revised network - aggregated at the county level. We look at the coefficient of variation as the key metric. Adjusting for the mean accounts for the fact that the change in the number of outlinks predicted by the model needs to be interpreted in relative rather than absolute terms. We exclude the outdegrees of Nairobi and Mombasa when we compute the coefficient of variation in every other row. The first two scenarios assume similar linking patterns for informal and small formal firms, conditional on their sector and county. In the second scenario, we use a simple version of the updated entry probabilities $p(\theta)$ that does not account for differences in firm size across sectors and locations. Scenarios three to five rely on the default assumptions on how to update $p(\theta)$ to incorporate informal firms. Instead, assumptions about $p(\theta, \theta')$ are modified: Scenarios three and four assume that informal firms do not sell to formal firms at all, but buy 50% or 75% of their inputs from the formal sector, respectively. Scenario five maintains the assumption that 50% of inputs are sourced from the formal sector and further allows 25% of the informal firms' sales to go to formal firms.

Table C2: Social connectedness, travel time and county-by-county-links

	Any		Without within county trade	
	Baseline	Revised	Baseline	Revised
Social connectedness (log)	0.004	0.009**	0.007**	0.014***
	(0.002)	(0.004)	(0.003)	(0.003)
Travel time (log)	-0.023***	-0.043***	-0.012**	-0.009***
	(0.003)	(0.005)	(0.005)	(0.002)
No. observations	2,209	2,209	2,162	2,162
R2	0.876	0.565	0.901	0.349
Origin FE	✓	✓	✓	✓
Destination FE	✓	✓	✓	✓

We regress matrix of county-by-county outlinks, more precisely the share of inputs a given county purchases from another county, on social connectedness and travel time between the two counties. Standard errors are clustered at the origin and destination level. Social connectedness captures the probability of two random individuals being friends on a popular social media platform (Bailey et al., 2021), conditional on their present location.

Table C3: County-level changes in outdegree and county characteristics

	Outlinks counterfactual/outlinks baseline			
Formal sector share	-3.525***	-1.569	0.561	0.839
	(1.303)	(1.711)	(2.293)	(2.244)
Population (log)		-0.365*	0.323	0.871
		(0.213)	(0.542)	(0.613)
Gross County Product (log)			-0.545	-1.063**
			(0.395)	(0.485)
Market access (distance, log)				0.330*
				(0.187)
No. observations	47	47	47	47
R2	0.140	0.194	0.228	0.281

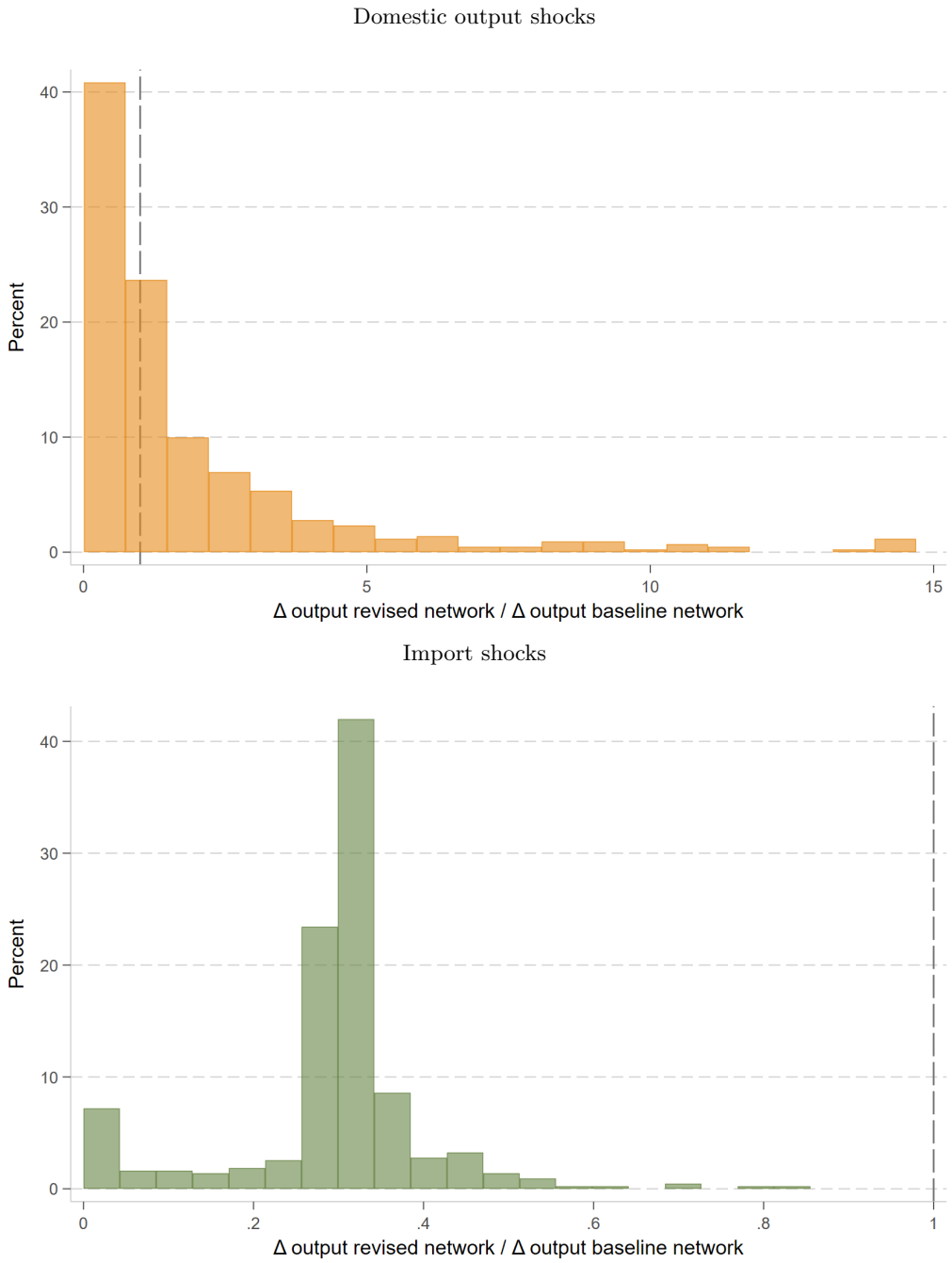
We regress the county-level change in outdegrees on various county characteristics including the formal sector share, the Gross County Product, and market access. We weight observations by Gross County Product.

Table C4: Differences in simulated output reduction for revised network with informal firms versus baseline network with formal firms only

	Domestic output shocks			Import shocks		
	(1)	(2)	(3)	(4)	(5)	(6)
Buyer sector-county formal employment share	-4.948***	-4.803***	-4.066***	0.115***	0.238***	0.053
	(0.931)	(1.458)	(1.001)	(0.032)	(0.052)	(0.034)
No. observations	431	431	431	431	431	431
Sector FE	-	✓	-	-	✓	-
County FE	-	-	✓	-	-	✓

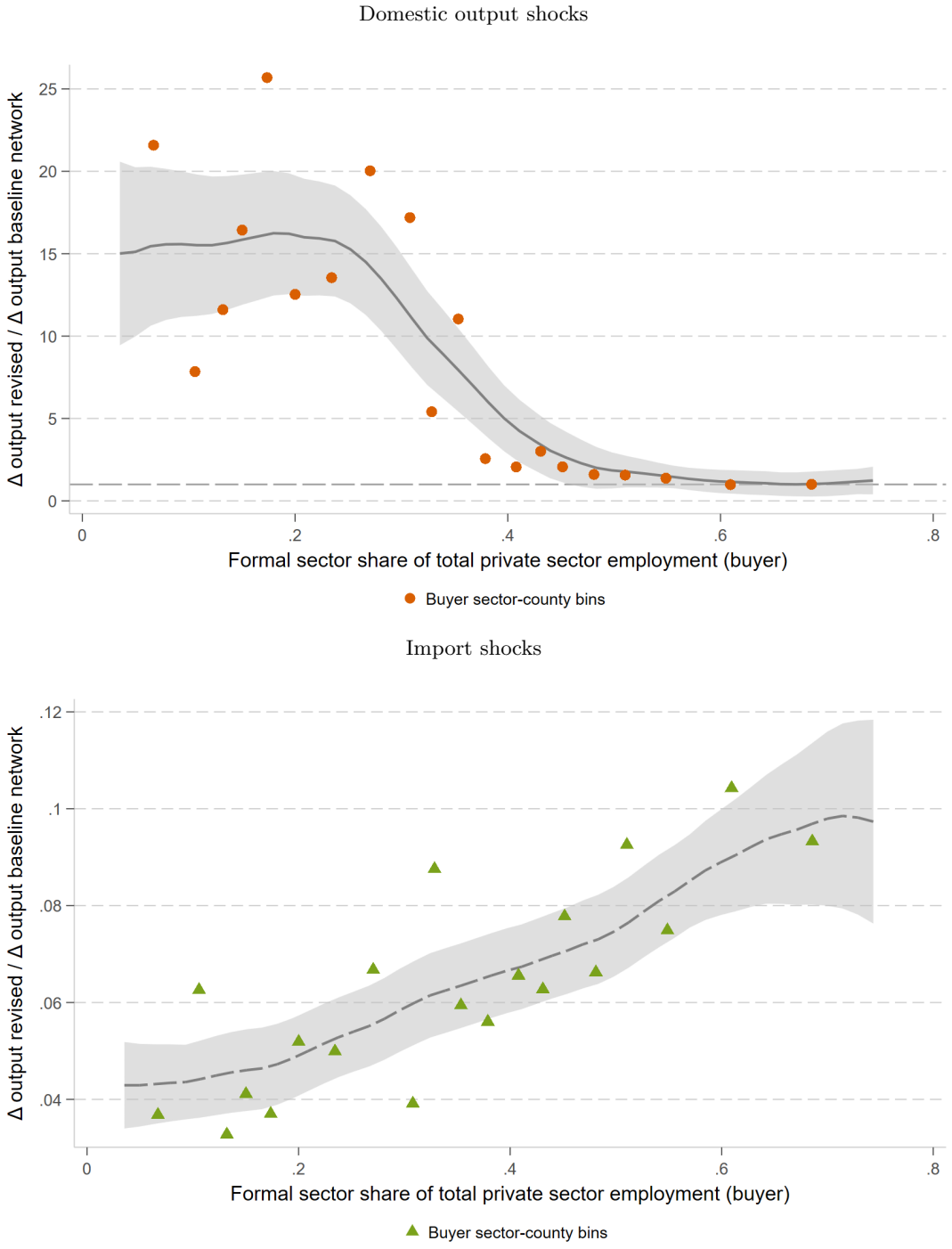
The outcome of interest measures the ratio of the impact response to an adverse shock if we account for informal firms versus relying only on the administrative data. The ratio is larger than one if we underestimate the impact of the shock and smaller than one if we overestimate it by not accounting for informality. The above table shows the results from regressing this change in output reduction at the sector-county level on the sector-county formal sector share.

Figure C5: The ratio of Δ output (revised network) and Δ output (baseline network)



We plot the ratio of the impact response to an adverse shock if we account for informal firms versus relying only on the administrative data, for the domestic and trade shocks respectively. The ratio is larger than one if we underestimate the impact of the shock and smaller than one if we overestimate it, if we do not account for informality. We aggregate the impact of the shock at the sector-county level where we weight the impact for each size and formality type using its entry probability $p(\theta)$.

Figure C6: % change in output drops
Scenario: no sales of informal firms to and 50% inputs sourced from the formal sector



The above graphs plot the percentage change of the output reduction in response to domestic and international output shocks for two scenarios: the baseline network using only administrative data and the revised network assuming that informal firms do not sell to the formal sector and source 50% of their inputs from formal firms. We aggregate the output reduction across buyer types at the sector and county level, weighted by the entry probability $p(\theta)$ for each size and formality type. The x-axis shows the formal sector share for each sector-region pair.

C.2 Updating Entry Probabilities for Scenarios that Distinguish Between Small Formal and Informal Firms

To differentiate informal firms from small formal firms, we simply split the first term in Equation 4 into two components: self-employment in the formal sector (small formal firms) and employment in the informal sector (informal firms). As outlined in Section 6.1, our proxy for informal firms' linking patterns still relies to some extent on small formal firms' connections to capture the sector-region composition of formal-informal linkages. Consequently, we can only introduce informal types where corresponding small formal types exist in the administrative data. This approach excludes 53 potential types, representing sector-county cells that account for 9% of private sector employment (excluding agriculture and non-market services). We do not introduce informal types for the agricultural and non-market service sectors. The resulting classification yields 1,376 distinct firm types: 419 informal, 493 small formal, and 464 large formal.