Spatial Inequality and Informality in Kenya's Firm Network *

Verena Wiedemann[†] Benard K. Kirui[‡] Vatsal Khandelwal[§] Peter W. Chacha[¶]

August 2024

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 overall domestic trade patterns in Kenya — a context with a high prevalence of informal economic activity. We first document a series of stylized facts and show that informal economic activity is not randomly distributed across space and sectors, with a higher incidence of informality in downstream sectors and smaller regional markets. We then link granular transaction-level data on formal firms with data on informal economic activity to estimate a structural model and predict a counterfactual network that accounts for informal firms. We find that formal sector data overstates the spatial concentration of aggregate trade flows and under accounts for trade within regions and across regions with stronger social ties. Additionally, the higher the informality in a sector and region, the more we underestimate its vulnerability to domestic output shocks and overestimate its vulnerability to import shocks.

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 Raphael Bradenbrink, Banu Demir, Kevin Donovan, Douglas Gollin, Justice Tei Mensah, Luke Heath Milsom, Sanghamitra Warrier Mukherjee, Solomon Owusu, Nina Pavcnik, Simon Quinn, Gabriel Ulyssea, Alexander Teytelboym, Christopher Woodruff, Hannah Zillessen and participants of the Oxford firms discussion group, the STEG junior research workshop, the World Bank's DaTax Group, the IFC lunchtime seminar, the CSAE research workshop, the KIPPRA seminar series, and the Jeune Street seminar series 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 qualitative data collection was approved by the Department of Economics Research Ethics Committee at Oxford (protocol no.ECONCIA21-22-48), the Strathmore University Institutional Scientific and Ethical Review Committee (protocol no.SU-ISERC1480/22), and the Kenyan National Commission for Science, Technology and Innovation (protocol no.NACOSTI/P/22/20556). 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.

[†]Department of Economics, University of Oxford &
Economic Research Unit, International Finance Corporation

[‡]Privatization Authority Kenya

[§]Department of Economics and Merton College, University of Oxford

[¶]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 the bias caused by ignoring informality is typically very 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 further employ a structural model to predict a counterfactual firm network that accounts for informality. We first show that informal economic activity is not evenly distributed across regions and sectors. We then find that extrapolating from formal sector data leads researchers to mismeasure the firm network and consequently mispredict key economic indicators such as the aggregate impact of economic shocks, the degree of spatial inequality, and the importance of urban hubs. For example, formal sector data assigns disproportionate weight to firms in urban hubs and those

[1] Detailed transaction-level survey data (see e.g. Startz, 2021) or administrative industry-specific data (see e.g. Hansman et al., 2020) are a popular complements to using tax records that often contain a greater depth of observable characteristics at the firm and transaction level. Given the high cost of collecting such data, researchers have typically focused on specific sectors when collecting tailored survey data.

[2] For example, the informal sector has been documented to play a crucial role in an economy's 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.

linked to international markets. Our results indicate that incorporating data on informality alongside formal firm-to-firm trade datasets can improve policy prescriptions.

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.[4] 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.[5]

We proceed as follows. We first establish a series of stylized facts about both domestic trade patterns of formal firms as well as the informal sector. Based on these insights, we estimate a network formation model that accounts for heterogeneity in firm size, sector, and geography. This model allows us to investigate whether including informal firms alters the spatial inequality observed in firm-to-firm links documented in tax records. We then analyze the network predicted by the model and use simulations of random domestic output and import shocks to demonstrate how accounting for informal firms, rather than extrapolating from formal sector data, changes predictions about the pass-through of such shocks.

We begin by documenting four stylised facts about firm-to-firm trade among formal sector firms. First, trade among formal firms is substantially more concentrated around Kenya's metropolitan areas compared to both population distribution and aggregate economic activity. Second, the concentration in aggregate trade flows is the result of spatial inequality along the extensive margins of the firm network i.e. the location of firms and trading relationships, rather than transaction volumes. In fact, 90% of the variation in aggregate trade volumes across counties can be attributed to a combination of the location of firms and inequality in the number of firm-to-firm links across regions. Third, upstream linkages (to suppliers) are more equally distributed across space than downstream linkages (to customers). Fourth, linking patterns vary by firm size, with small firms sourcing more from intermediaries and sourcing more locally.

Having established these stylized facts, we then explore whether the observed patterns are a

[4]With the exception of VAT data from India and Turkey, similar administrative records have predominately been used in smaller countries with less geographic variation and/or a smaller population for the purpose of research.

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

result of our limited view due to the systematic selection of firms into the administrative data or if they reflect the underlying structure of the economy. Importantly, we document that informality is not evenly distributed across space but varies systematically across sectors and regions of the country. 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. By comparing spatial inequality across various data sources, we show that the spatial concentration of economic activity is largely a formal sector phenomenon. As a result, we expect that accounting for the informal sector can systematically alter the structure of the observed production network.

To do this, we introduce and estimate a network formation model with heterogeneous node types following Bramoullé et al. (2012) that allows us to show how accounting for informal firms could alter the structure of the observed network and predict a counterfactual network. In our adaptation of the model, we classify firms based on their sector of operation, location, and size, reflecting the substantial heterogeneity along these three dimensions as 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). We first estimate this network formation model to predict the Kenyan firm network as it is. 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.[6]

We then predict a counterfactual network that accounts for informal firms by combining the model with 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

[6]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.

as a proxy for informal firm linking patterns, we address concerns that informal firms might link more within their own locality and source more from intermediaries relative to formal firms operating in the same sector and location (e.g. due to internal economies of scale (Grant and Startz, 2022)).[7]

We now use the counterfactual network to answer the question of interest: How do spatial patterns of trade 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 counterfactual network relative to the baseline network. The spatial inequality in outlinks declines by 7% and the prominence of Nairobi falls substantially. 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.[8] 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. These patterns have implications on the predicted pass-through of domestic and international trade shocks across space.

We simulate the pass-through of domestic and import shocks to sector-regions using both the network estimated by the structural model and the counterfactual network obtained after accounting for informal firms. When relying on the counterfactual 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 extrapolate from formal sector data. 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 4 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 ignored. This discrepancy arises because import shocks primarily affect larger formal firms, which carry less weight in the overall firm network once informality is taken into account.

Our paper contributes to the literature on macroeconomic development, informality, firm net-

[7] In the stylized facts, we provide empirical evidence showing that small firms link more locally and buy more from intermediaries relative to their larger peers. While some barriers that small formal firms face in linking with larger firms or firms outside their locality or sector are similar to those faced by informal firms, informal firms might encounter additional obstacles. These can include wedges introduced by the VAT system itself (De Paula and Scheinkman, 2010; Gadenne et al., 2022). These additional barriers do not impact our results as long as they generate similar sectoral and geographic linking patterns for both informal and small formal firms.

[8] Social connectedness is proxied by the likelihood of two randomly selected individuals being friends on a popular social media platform (Bailey et al., 2021).

works, 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).[9]

We document stylised facts about the informal sector in Kenya, which reveal that informality systematically occurs in downstream activities and in smaller markets. Understanding the implications of this non-random selection of firms into administrative records is particularly important, given the growing reliance on such data in the literature. We hence use a structural model to show how ignoring this sector can have substantial implications for the structure of the production network, the estimated impact of shocks, and patterns of spatial concentration. Our approach to employ a structural model to bridge gaps in our understanding of informal firm dynamics 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.[10] 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, 2023; Arkolakis et al., 2023), shock propagation in firm networks (Baqaee, 2018; Huneeus, 2018; Carvalho et al., 2021; Baqaee and Farhi, 2024), and urban primacy (as published in Jefferson (1989), Jefferson, 1939; Memon, 1976; Ades and Glaeser, 1995; Henderson, 2002; 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. As a result, this oversight may cause researchers to underestimate the economic connectedness and vulnerability of

[9]Studies like Startz (2021) circumvent the issue of informality by collecting granular transaction-level records from wholesale traders in Nigeria through a survey.

[10]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).

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 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 instead of studying the decision to formalise of marginal firms.

Finally, we contribute to the literature on estimating network statistics and reconstructing networks in the presence of missing data. For instance, Chandrasekhar (2016) provides two ways to correct for biases in network statistics that can arise due to missing data. Our technique is similar in spirit to the graphical reconstruction technique proposed in their paper. We reconstruct the network by estimating a structural model using the data that we can observe. However, it is important to keep in mind, that in our context, nodes are missing in a non-random manner. The preferences for network formation of these nodes can be systematically different from those of the nodes that we observe. We account for this by exploiting the heterogeneity in firm size, location, and sector that we observe in the data. The key assumption here is that conditional on an informal sector firm being of the same type as a formal firm (i.e. being small and operating in the same sector and location) they are going to form links similar to the observed firm in the formal sector. This assumption then allows us to predict a counterfactual network that is useful to understand the implications of ignoring informality in the absence of alternative sources of data. To establish the potential validity of this assumption we document how small formal firms behave similar to our expected behaviour of informal firms in that they source more from the same locality and buy from intermediaries such as retailers and wholesalers rather than manufacturing firms.

Our paper is structured as follows. In Section 2, we describe the administrative data used to

measure the formal firm network. We document stylised facts about this network in Section 3. We discuss the role of the informal sector and map its spatial and sectoral composition in Section 4. Finally, we tie the two together in Section 5 where we discuss and estimate a network formation model with preferential attachment (Bramoullé et al., 2012). We present the results of the counterfactual in which we include informal firms in Section 6 and predict the impact of simulated domestic and trade shocks in Section 7. We conclude in Section 8.

2 Administrative data

2.1 Description of data sources

Our analysis draws on micro data from value-added and pay-as-you-earn tax returns. We utilise the Kenya Revenue Authority's tax registry to compile basic, self-reported information on each firm, namely the 4-digit sector classification, the business type, the start date of its operations, and the headquarters location. All data sets can be linked using anonymised firm identifiers. Amongst the tax reports, the key data set are monthly value-added tax (VAT) returns. The VAT returns include details on firm-to-firm transactions between registered firms. Sales to and purchases from non-registered parties (e.g., exempt parties, non-registered businesses, final consumers) are recorded as an aggregate monthly figure.[11]

VAT applies to individuals and 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 filter the data set for entities that identify as private companies or partnerships in their tax-registration form. In doing so, we exclude all government-owned firms, government agencies, international organisations, NGOs, trusts, and clubs.We 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, which we observe in the data.[12]

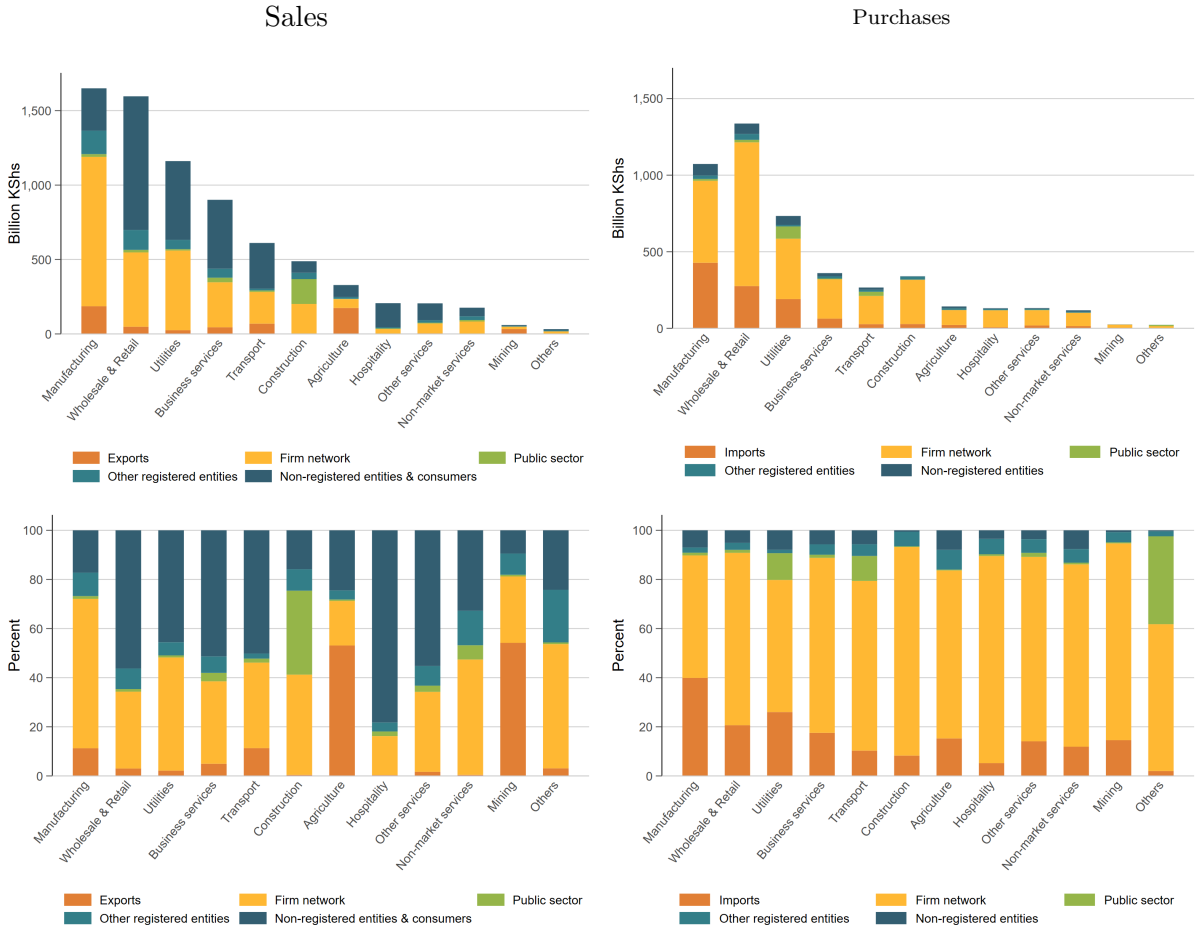
Figure 1 plots the sector composition and the respective sales and input channels of firms covered in the administrative records. Manufacturing and wholesale and retail firms together account for almost half of the sales we observe in the tax records.[13]

[11] Throughout this paper, we use the term "non-VAT paying" firm to refer to private sector entities that are either not VAT-registered due to their size, exempt from VAT payments due to the products and services they sell or do not comply with the tax law.

[12] We apply the VAT threshold to exclude firms that registered for VAT to bid for tender but were never operational.

[13] Firms providing financial and educational services, and to a significant degree, those trading in agricultural products and pharmaceuticals, are exempt from filing VAT returns.

Figure 1: 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 sales to and purchases from registered vs non-registered parties as a percentage of total sector-level sales and purchases.

3 The spatial concentration of domestic firm-to-firm trade

We start by analysing the geography of Kenya's firm network of formal firms. We document four stylized facts. First, firm-to-firm trade flows within the network are highly spatially concentrated around Kenya's metropolitan areas. Second, the majority of this spatial concentration (90%) in economic activity can be attributed to the extensive margin of trade: firm location and the number of firm-to-firm links. Third, upstream linkages (to suppliers) are more equally distributed than downstream linkages (to buyers). Fourth, linking patterns differ by firm size. Small firms are more likely to source locally and are more likely to buy from intermediaries.

We present these stylized facts to highlight the core features of the data, which also motivate the setup of our model. We will revisit the first fact once we take informality into account. The second fact, that firm location and firm-to-firm relationships explain the bulk of variation in

trade volumes across locations, motivates the modelling choice to focus on these two extensive margins of trade rather than trade volumes. Due to the third fact that shows that downstream linking patterns vary more across space, we will focus on modelling the outdegree of firms (i.e. the number of buyers) and target the outdegree distribution of the real-world network in our structural estimation. Finally, consistent with the fourth stylised fact, our model incorporates firm size as a key dimension influencing linking patterns, alongside sector and geography.

3.1 Urban primacy in Kenya's firm network

Kenya's firm network is strongly concentrated around its metropolitan areas Nairobi and Mombasa.[14] As much as 68% of the sales volume within the network of formal firms is generated by Nairobi-headquartered firms. Notably the city's role in the firm (or trade) network is disproportionate relative to its population and even its contribution to aggregate GDP.[15] In 2019, as little as 9% of Kenya's population lives in Nairobi County[16] and the city contributes 37% of Kenya's GDP outside the agricultural sector (see Table 1).

[14] Although, no exact figures for comparison are reported in the respective papers, Huneeus (2018) and Cardoza et al. (2023) find a stark geographic concentration of trade flows around metropolitan areas in Chile and the Dominican Republic.

[15] Nairobi and Mombasa first emerged as Kenya's primate urban centres 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.

[16] We map firm headquarter locations and population density in Appendix Figure A2.

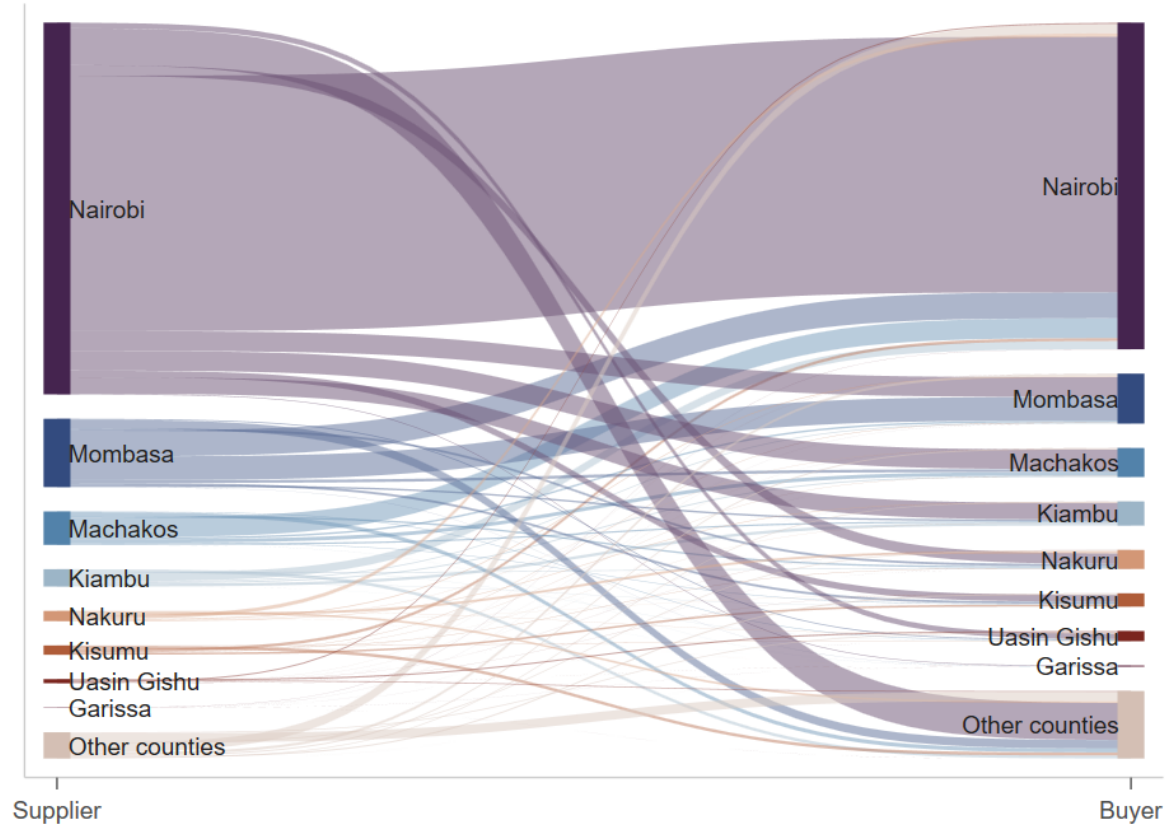
Table 1: Geographic concentration of economic activity in Kenya

	Nairobi	**Mombasa**	**Pareto exponent**	
	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. 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
Network sales	68	13	0.35	0.02
Network purchases	60	9	0.43	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): $\log \text{rank} = \log A - \alpha \log x$.

The county-to-county trade flows plotted 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, segments on the right are proportional to purchases. The colouring of the trade flows aligns with the county of origin. 89% of the over 21.5 million firm-to-firm transactions in 2019 involved at least one firm based in Nairobi or Mombasa. Moreover, trade between Nairobi-based firms themselves accounts for 47% of the total trade volume. The graph further reveals that trade flows out of Nairobi and Mombasa are larger than inflows into the two cities.

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.

Moving beyond Nairobi and Mombasa, how concentrated is economic activity if we consider the entire distribution? The distribution of both firm and city sizes is often well-approximated by a Pareto Distribution (Gabaix, 2009). Under this premise, the Pareto exponent can be considered a measure of inequality for the dispersion of population and economic activity (Gabaix, 2009; Soo, 2005; Gabaix and Ioannides, 2004). In Table 1, we compare the Pareto exponent α for the regional distribution of population and gross value added (KNBS, 2022) to a series of measures derived from the administrative data. [17] The Pareto exponents for both total county-level GDP and aggregate income generated outside the agricultural sector is close to unity. At the same time, Kenya's population is more evenly distributed across counties than economic

[17] The α for each indicator is obtained via rank-size regressions (Gabaix and Ioannides, 2004). I.e. a county-level regression of each county's rank (log) on the respective measure x (log): $\log \text{rank} = \log A - \alpha \log x$. A lower value indicates a flatter slope and hence more inequality across counties.

activity. Turning to the firm network, we find values of α that are substantially lower than one, indicating a high degree of spatial inequality. An exponent of 0.63 suggests that the number of VAT-paying firms is still fairly evenly distributed across counties - despite the concentration of firms in Nairobi. Meanwhile, the αs for employment, value added, sales and purchases are 57%-76% lower than the exponent for overall economic activity (GDP aka the Gross County Product). Comparing the α for network sales, i.e. trade flows out of a county (0.35), versus network purchases, i.e. trade flows into a county (0.43), shows that downstream trade flows are much more concentrated than upstream trade flows. Put differently, a smaller number of counties supplies disproportionate amounts of inputs to the rest of the country.

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.2 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 quantify which margins of aggregate trade flows drive the spatial trade patterns, we use the granular transaction-level records to decompose them into sub-components.

3.2 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 N, the number of relations R per firm, the number of transactions c per relationship, and the trade volume v per transaction. In a nutshell location o's sales to the firm network τ can be summarised as:[18]

$$\tau_o = N_o \times \frac{R_o}{N_o} \times \frac{c_o}{R_o} \times \frac{v_o}{c_o} \tag{1}$$

Table 2 summarises the share of the variance attributed to each term in both upstream (purchases) and downstream (sales) trade flows.[19] The number of firms operating in each county alone accounts for 67% of the variance in purchases across counties.[20] The number of supplier relationships other have with the county accounts for yet another 22%. This leaves a little over

[18] The same is true for purchases.

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

[20] This includes purchases the firms of a respective county make within their own county or from firms outside the county.

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. [21]

In the next two stylized facts, we document how firm-to-firm linking patterns vary across geographies and by size.

Table 2: 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

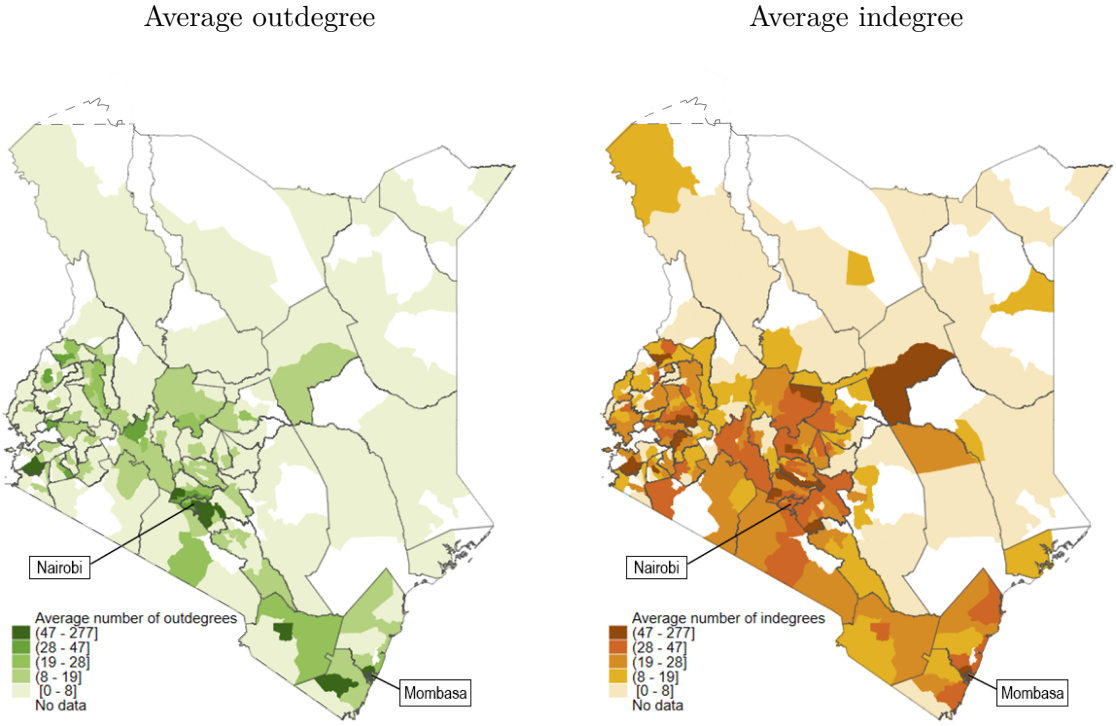
3.3 Upstream linkages are more equally distributed than downstream linkages

Our observation that firm-to-firm relationships are a more important margin for cross-county network sales rather than purchases is consistent with the pattern that links to suppliers are more evenly distributed among firms than links to buyers. Consistent with this, Figure A1 shows that the out-degree distribution is more unequal and has a heavier tail, indicating a higher proportion of high-degree nodes compared to the in-degree distribution. In other words, there is greater heterogeneity in the sales channels utilised by firms than in their input channels. While virtually all business models require some form of material input, firms can have diverse customer bases, including other businesses, final consumers, or the public sector (see Figure 1).

We show that this pattern replicates across space by plotting the average county-level in- and outdegree in Figure A3 and map averages at the subcounty level in Figure 3.The comparison between outdegrees mapped on the left in Figure 3 and indegrees mapped on the right, shows that the average indegree is much more equally distributed across space. Nairobi and Mombasa based firms on average have 30 buyers, while in the rest of the country firms have only 21 buyers

[21] The variance decomposition is a useful exercise to track the respective margins of trade. It, however, does not identify the relative importance of selection of entrepreneurs into certain regions versus the place effect of a region on an entrepreneur's ability to form relationships.

Figure 3: Average in- and outdegrees across space



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

on average.[22] Turning to the average indegree (suppliers), Nairobi and Mombasa no longer stand out as much. Firms in the two cities have 28 suppliers on average, while the average across all other counties is 26. This pattern also aligns with the higher level of spatial inequality found in network sales relative to the location of firms in Table 1.[23]

3.4 Linking patterns differ by firm size: smaller firms source locally and from intermediaries

Lastly, we consider differences in linking patterns between firms of different sizes across space and sectors. We look at the relevance of firm size for two reasons. First, a sizeable literature has documented the close link between network links and firm size (Bernard et al., 2019, 2022; Arkolakis et al., 2023). Second, 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 that 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

[22] Here we refer to firm-level averages, while the map plots subcounty averages. Spatial concentration in links can also be concentrated within the county, e.g. in Nairobi's central business district.

[23] In Table 1, the Pareto exponents for the number of firms suggest that firm locations are more evenly distributed than sales volumes and the number of firm-to-firm relationships within the network.

to exhibit different sourcing patterns. For instance, take wholesalers in Garissa county. Large wholesalers in Garissa county might source directly from manufacturers in Nairobi, while smaller wholesalers within the same region might rely on other local wholesalers instead.[24] Indeed in the data (see Tables 3) we find that small buyers within the same sector and county are less likely to directly source from manufacturing firms, but instead are more likely to source from retailers or wholesalers. Further, they are less likely to source from Nairobi-based suppliers and more likely to source locally.

Table 3: 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 buyers (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.

4 The role and position of the informal sector

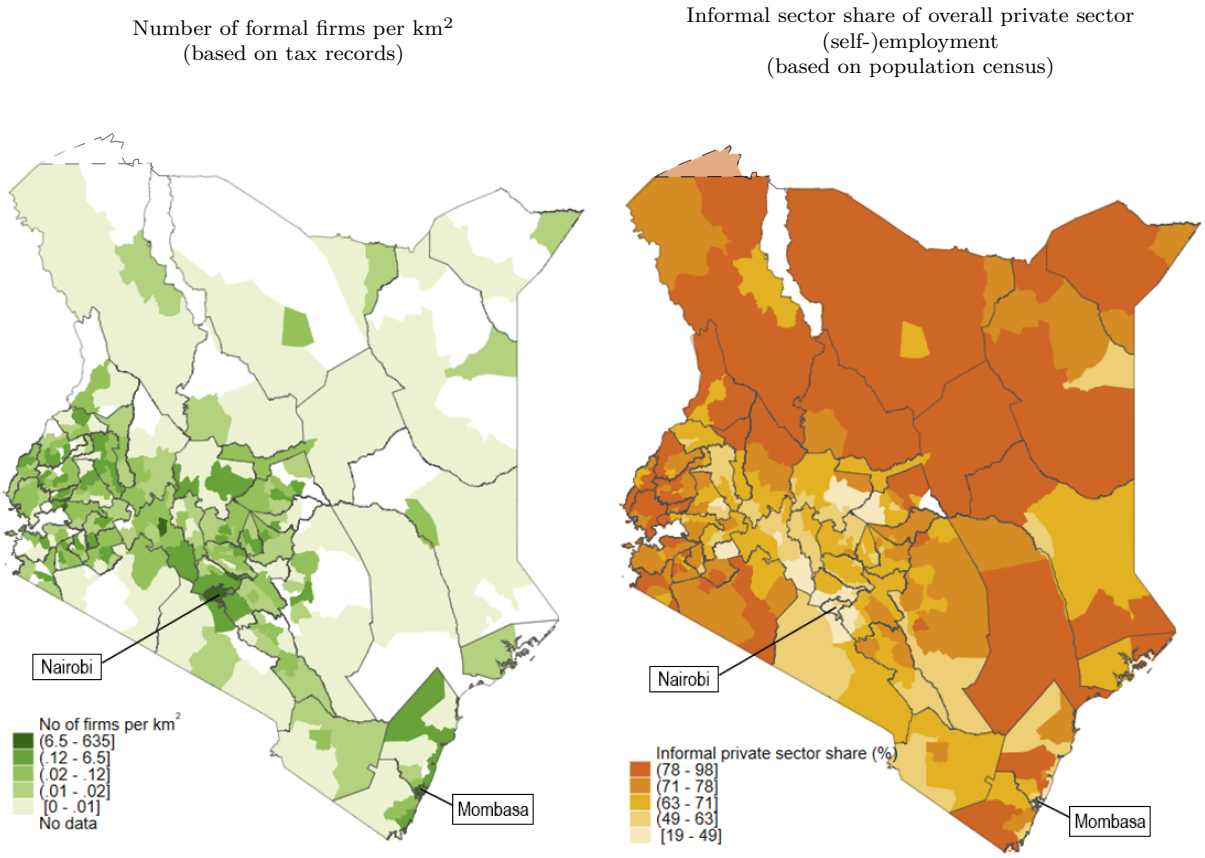
In this section, we empirically explore to what extent the above spatial patterns could be driven by the presence of the informal sector. Do we have reason to suspect that some of these patterns do not apply to the economy as a whole, but arise only because significant parts of the economy are not visible in administrative records?

Broad spatial patterns like the concentration of firms correlate strongly with population density (see Figure A2). However, as documented in the previous section, trade between formal firms is much more concentrated in space than both population and aggregate non-agricultural GDP (see Table 1). In Figure 4, we map the number of formal firms per km^2 and the share of people being employed in the informal sector.[25] Both measures correlate negatively with each other. Locations with a high density of formal firms have a low informal sector employment share and vice versa.

[24] In line with the definition we will use to estimate the model, we define small firms as those in the bottom sales quartile within a sector and county.

[25] Note that the measure of informality does not rely on the administrative data, but is based on the population census. Hence the correlation is not mechanical.

Figure 4: Location of formal firms and informal sector employment shares



The right 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 - also at the subcounty level. Sub-counties represent the second administrative layer. The borders of Kenya's 47 counties, the first administrative layer are outlined in grey.

In Table 4, we show that the number of firm-to-firm links observed in the administrative data correlates with local formal sector shares as well, even when controlling for population and travel time to the metropolitan areas.[26] This motivates us to have a closer look at whether links of firms in sectors and regions with high informality are captured in a complete manner. Whether or not they are can have substantial implications for the spatial pass-through of domestic and trade shocks.

[26] 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.

Table 4: 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.

In this section, we utilise third-party data collected by the Kenya National Bureau of Statistics (KNBS) to assess the scale of the informal sector beyond the VAT-system and to analyze the sectoral and geographic distribution of informality in Kenya. Our approach is structured as follows. First, we briefly discuss margins of informality in the context of firm networks and related measures of informality. We then document three essential stylised facts that will inform how we account for unobserved informal firms in our subsequent counterfactual exercise where we attempt to predict a network that accounts for informal firms. This model-based counterfactual will shed light on the relevance of unobserved firms for the structure of Kenya's firm network.

4.1 Margins of informality in firm networks

Many plausible definitions of informality exist and can be applied even within the same setting (La Porta and Shleifer, 2014). Firms can be formally registered entities (extensive margin of informality), but engage in informal activities (intensive margin of informality), e.g. by hiring informal workers (Ulyssea, 2018). A wholesaler we interviewed in Nairobi's Central Business District[27] explains 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, they downplay their output VAT (i.e. sales). Some customers will purchase with receipt and output VAT on it. Some customers*

[27] The firm's customers cover the whole range of potential sales channels: other wholesalers, retailers, public institutions, and some individual consumers.

will purchase without a receipt."

Table 5 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 unit is too small to be taxed.[28]

Table 5: Margins of informality in firm networks

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

The extensive margin helps identify who the informal firms are. These can be either (i) small firms who never crossed the annual revenue threshold for VAT or (ii) larger non-reporters, i.e. firms with revenues above the VAT threshold, but which do not file VAT. If either of the two trading parties is informal, we do not observe transaction-level information on their interaction.[29] For the purpose of this paper, we exclusively focus on whether or not firms pay national taxes like VAT. Some of the firms we classify as informal might be formal according to alternative definitions of informality.[30]

The intensive margin considers informality at the transaction level conditional on both parties being VAT-registered firms. Here informality in administrative records can either occur because (i) transactions fall below a reporting threshold specified in the tax code (for transactions rather than for firms) or (ii) non- or under-reporting of transactions that firms are required to declare. The first aspect is not a concern in the Kenyan context.[31] The Kenya Revenue Authority requires firms to record transactions of any size, conditional on both parties being VAT-registered. Omission of transactions between two formal firms or the misreporting of trade volumes on the other hand remains a concern for us.[32] We are able to recover some of the omitted transactions

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

[29]In our setting, we observe sales of formal firms to non-registered entities, including non-VAT businesses, but only as a monthly aggregate figure and not at the transaction-level. We are unable to distinguish whether those sales go to final consumers or informal firms.

[30]Specific to Kenya's legal context, a lot of firms we miss out on in the national tax records do pay sub-national business license fees. These license fees are collected by county governments. VAT-paying firms are thus a sub-set of the universe of firms with a county license.

[31]The application of transaction thresholds varies substantially across contexts. In Belgium, Costa Rica, and Turkey, for example, a trading pair of firms does not need to report their transactions if their annual bilateral volume falls below 250 €, $4800, and $2650 respectively (Dhyne et al., 2015; Alfaro-Urena et al., 2018; Demir et al., 2022).

[32]It is partly mitigated by asymmetric incentives of buyers and suppliers to report transactions correctly. Buyers might want to overstate purchases to claim refunds for input VAT, while suppliers have an incentive to downplay the volume of their sales to reduce their output VAT liability. Almunia et al. (2022) show that despite this built-in VAT enforcement mechanism firms in Uganda still misreport trade volumes, sometimes even against their own

or under-reported trade volumes by relying on information from a firm's trade partner when processing the data. Any residual misreporting feeds into the data as informal trade flows along the intensive margin.

We now turn towards a discussion of available measures of informality.

4.2 Measuring informality

All our proposed measures represent formal sector shares, i.e. the proportion of overall economic activity that can be traced back to the formal sector. Table B1 provides an overview of the data sources we draw on, while Table 6 summarises how we compute the various measures of informality. Importantly, we not only measure informality as the gap between what is captured in the administrative data and overall economic activity, but further rely on several measures of informality generated independently of the administrative records. This is crucial for the interpretation of insights from the administrative data. It allows us to establish that gaps between the administrative data and aggregate economic activity are reflective of economy-wide dynamics between the formal and informal sector and do not capture idiosyncratic patterns that are specific to the VAT system.

Table 6: Measures of informality

Unit	**Numerator** (formal sector)	**Denominator**	**KNBS**	**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 B1. The term "all firms" refers to both licensed and unlicensed businesses based on KNBS (2016) estimates and county government records.

Throughout the remainder of this section, we draw on three different indicators for economic activity, namely employment figures, the number of firms, and value added (sales - purchases). Our default measure of informality will be an employment-based measure that draws on the 2019 population census (KNBS, 2019). The two alternatives based on the number of firms and value added rely on estimates of the universe of businesses in KNBS (2016)[33] and on estimates of the regional economic size captured by the Gross County Product (KNBS, 2022) respectively. While the employment-based measure is likely to primarily capture the extensive margin of informality, the value-added-based measure can be considered as an aggregation of all margins. The number of firms will provide a more nuanced picture on the extent of potential non-compliance by firms

interest.

[33] 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.

on the extensive margin. The key advantage of the population census data is that they allow us to dis-aggregate employment records both at the sectoral and regional level at the same time. This aspect is missing for measures based on the number of firms.[34]

We exclude any employment or firms in the agricultural sector or non-market services when measuring informality at the regional level.[35] In both cases, the tax records cover a small and very specific sub-population of firms and employees only. In the case of agricultural firms, the administrative data cover large-scale commercial agriculture only. The vast majority of these firms are export oriented (Chacha et al., 2024). For non-market services, the VAT data mostly capture a small number of firms which operate in sectors dominated by non-profit organisations and the government. In addition, the majority of for-profit firms in this sector enjoy VAT exemptions and hence do not appear in the network.

Our preferred measure of informality considers formal sector employment as per the 2019 population census. It correlates strongly (ρ= 0.83, Table B2[36]) with its counterpart based on the administrative records (also see Figure B1).The employment based measure will also be a key input for our counterfactual firm network in the next section.

4.3 Three stylized facts on informality in space

We document three stylised facts about the informal sector. The first fact highlights the importance of accounting for the informal sector in terms of its contribution to the Kenyan GDP. The second stylised fact shows how the incidence of informality is not randomly distributed in the economy but varies by sectors, geography, and the position of the firm along the supply chain. This also helps us motivate the assumptions of the model that we present in the next section. The third stylised fact shows that the spatial allocation of economic activity is not as unequal once we begin to consider informal firms. The model and estimation that follows will help us further understand the economic implications of ignoring informal firms.

[34]Further, we can distinguish between employment in the private compared to the public sector. None of the other measures allows for this distinction.

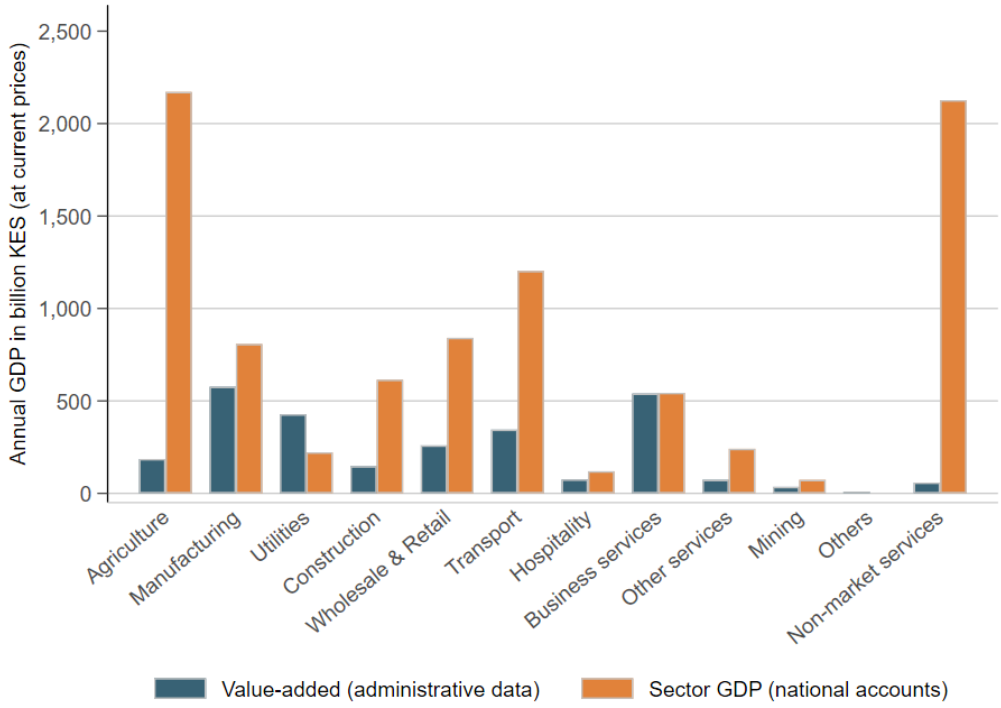
[35]We do so wherever possible. The county-level statistics in the MSME report (KNBS, 2016) and the Census of Establishments (KNBS, 2017) do not allow us to abstract from those sectors.

[36]Table B2 summarises the correlation coefficients of the alternative informality measures.

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

The gross value added generated by VAT-paying firms corresponds to 36% of Kenya's annual GDP[37] (see Appendix Table B3).[38] The gap between value added in the administrative data and aggregate GDP figures arises for two reasons: The first reason is differential treatment of sectors in the tax code, in particular the treatment of financial services, non-market services, and agriculture.[39] The second reason is informality.

Figure 5: Value added by VAT firms vs GDP



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

If we exclude sectors that are to a large extent exempt from VAT (non-market services, agriculture) or have special reporting rules applied to them (financial services), the VAT sector accounts for 67% of residual economic activity.[40] 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 the 29% estimated by Hassan and Schneider (2019) for 2013. Both of these

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

[38] Average for the period between 2015 and 2019. The value added recorded in the VAT data fluctuates between 40% (2016) and 28% (2019) of Kenya's GDP.

[39] We discuss the relevance of sectors that are not well-represented in the data for reasons to do with special tax treatment in Appendix B.2.

[40] To arrive at this number, we exclude financial services, non-market services, and agriculture from total GDP. Appendix Table B3 details how the GDP share of the VAT sector changes if each of them is added or removed sequentially.

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 step 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 informality for firms. Having established that we are missing out on 33% of the Kenyan economy by looking at the formal sector only begs the question - where to find those informal firms?

4.3.2 Fact 2: Incidence of informality varies by sector, geography, and position along the supply chain

The incidence of informality varies substantially across sectors.

We now turn our full attention to sectors where gaps arise due to informality rather than tax exemptions. Comparing the value added of each sector (based on the administrative data) to the sectors' contribution to Kenya's GDP (based national accounts) in Figure 5, shows that manufacturing and business services are best represented in the administrative data. Both sector predominantly interact with other formal firms (see Figure 1). We observe a larger disparity between value added and GDP contributions in downstream sectors (see Figure B2), closer to the final consumer, where VAT self-enforcement is weakest (Naritomi, 2019). We observe a similar trend when concentrating solely on the extensive margin of informality, which indicates a high occurrence of informal firms in the wholesale and retail sectors.

The gap between value added and GDP is an aggregate of all margins of informality discussed in Section 4.1. We are unable to precisely quantify the extent to which each margin contributes to aggregate informality. To get an idea of how the extensive margin of informality varies across sectors, we compare the number of VAT-paying firms to a number of alternative firm counts by KNBS. We distinguish between firms with revenues above the VAT-reporting threshold and smaller firms that are too small to be asked to file VAT. Both measures suggest that the number of wholesale and retail firms is much larger than the VAT data would suggest.

The first type of extensive margin informality arises due to non-reporting firms with revenues above the VAT threshold. Figure B3 plots both the number of VAT paying firms and firms in the Census of Establishments (CoE) (KNBS, 2017) with revenues of KShs 5 million and above in 2016. The distribution of firms across sectors mirrors each other quite well.[41] Wholesale & retail stands out as the sector with the largest number of non-compliant firms.[42]

[41] The number of firms in the administrative records can be higher because the CoE tends to under-count firms without a highly visible establishment. Further, sector classification is self-reported in the administrative data and hence might not align perfectly with the classification in the CoE.

[42] Conversely, education and health sectors' discrepancies largely stem from VAT exemptions, with public es-

The vast majority of firms in Kenya, however, are too small to be VAT-registered. The 2016 KNBS (2016) report estimated 7.4 million businesses in Kenya, only one-fifth of which are licensed, and a mere fraction captured in VAT or CoE data.Of the licensed businesses the CoE and the VAT data capture 9% and 2.5% respectively. The bottom graph of Figure B3 plots the business count by data source and sector. For agriculture, utilities, and construction, the overall number of licensed businesses aligns closely with the number of businesses captured in the CoE or the VAT data. In all other sectors, the number of licensed businesses is substantially higher.

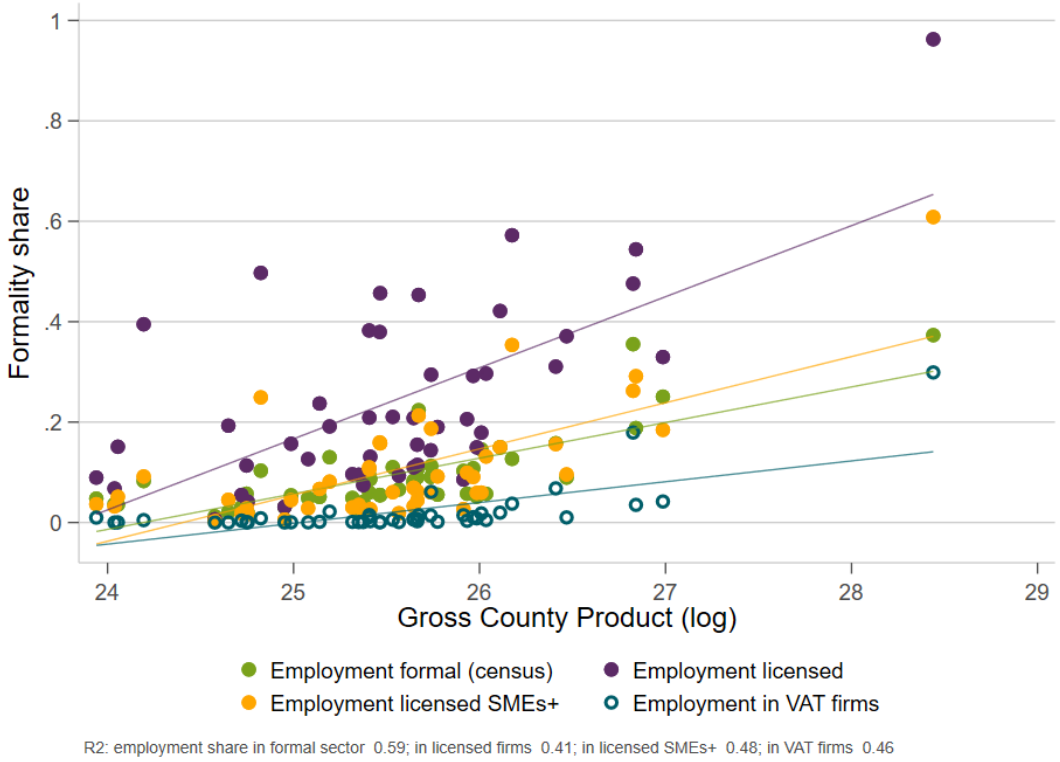
Turning to the intensive margin of informality, the prevalence and possibility of informal transactions involving registered firms becomes evident when considering the construction sector. Only 24% of this sector's GDP contribution is captured in VAT records, despite most firms being VAT-registered (see Figure B3). This suggests that much of the value-added gap stems from underreporting in VAT filings.

The VAT-sector share correlates positively with regional economic size and income. Moving on from sectors, we explore the spatial distribution of informal firms, which predominantly reside in smaller markets. We find a strong correlation between a county's economic size and income level (Gross County Product per capita) and its share of the formal sector. Economic size accounts for 35% to 52% of the variation in formal sector shares across counties, as shown in Figure B4. In this figure, we correlate economic size and income levels with measures of informality using three indicators for which we have KNBS benchmark data: employment (turquoise), value added (orange), and the number of firms (green). Each marker in the scatter plot represents one of the 47 counties. It is notable that the slope for value added is steeper than for employment. This aligns with findings from cross-country studies suggesting that the informality of output declines less steeply with income levels compared to the share of informal workers (Kose et al., 2019). Moving on from sectors, where do we find informal firms in space? Here the answer is: in smaller markets. A county's overall economic size and income level (Gross County Product per capita) correlates strongly with its formal sector share. Economic size explains between 35% and 52% of the variation in formal sector shares across counties (see Figure B4). In Figure B4 we correlate economic size and income levels with measures of informality based on each of three available indicators for which we have KNBS benchmark data: employment (turquoise), value added (orange), and the number of firms (green). Each marker in the scatter plot represents one of the 47 counties. We note, however, that the slope for value added is steeper than for employment. This finding is in line with the general notion previously shown in cross-country studies that informality of output declines less steeply with income levels than the share of informal workers (Kose et al., 2019).

tablishments like schools included in the CoE.

To verify whether the positive correlation between market size and formal sector share is merely an artefact of the administrative records, we also correlate three other employment-based formality measures with the Gross County Product in Figure 6. While the slope becomes flatter for measures that apply a more stringent definition of informality, with the VAT-based measure being the most stringent one, the R^2 barely changes. This indicates that the variation in county-level informality explained by economic size is consistent across different informality measures.

Figure 6: Share of formal sector employment and regional market size



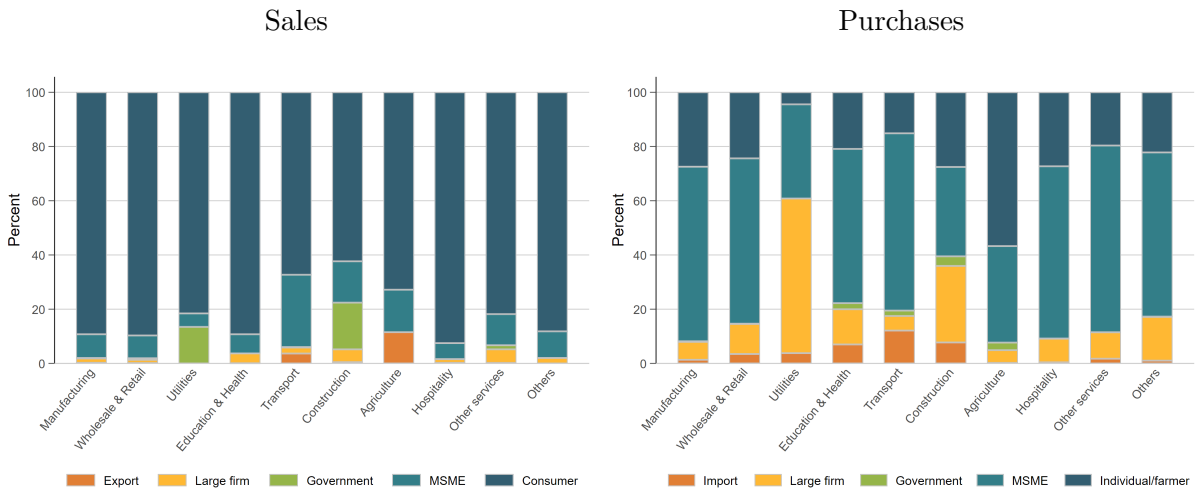
The first measures uses the formal sector employment share according to the 2019 population census, the second measure considers the number of employees in licensed businesses, the third uses the same measure but disregards micro-enterprises, and the fourth measure considers employment in the tax records. 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 6.

Informal firms are located downstream of larger firms. While larger gaps between value added and GDP in more downstream sectors, along with a high number of non-VAT paying firms in wholesale and retail, indicate the downstream positioning of informal firms, we utilize survey data to further document their placement along the supply chain.

We find that informal firms are located mostly in consumer facing roles and downstream of large formal firms. In other words, large firms provide inputs for informal firms, while informal businesses in turn often take on the role of distributors in the economy with consumers as their

main source of demand.[43] 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 sell to large firms, while 14.5% acquire inputs from them.[44] Figure 7 shows that the pattern holds across sectors.[45] Our 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 West Bengal, India (Zhou, 2022) respectively.[46]

Figure 7: 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). The survey asks each firm for their main input sources and their main customer type. We restrict the sample to participating firms with an annual revenue below the VAT registration cut-off. Note that the 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.

The higher 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. This enforcement mechanism incentivizes downstream firms to ask their suppliers for receipts in order to claim input VAT they can then deduct from the output VAT they have collected on behalf of the revenue authority. The weak link of any such system are consumers or VAT-exempt entities, who are not eligible for VAT refunds and hence do not have an incentive to ask for a receipt Naritomi (2019). Put differently, we expect

[43] Cordaro et al. (2022), for example, show how microenterprises subsidise the distribution of fast-moving consumer goods of a multinational in Kenya.

[44] KNBS (2016) defines non-MSMEs/large firms as entities with more than 99 employees.

[45] The survey responses can be interpreted as a lower bound estimate of the interaction between the VAT-registered and non-VAT-registered sector. The main trading partners for MSMEs are other MSMEs, and the survey does not differentiate between micro, small, and medium enterprises. Kenya's Micro and Small Enterprises Act No.55 of 2012 defines small enterprises as those with up to 50 employees and KShs five million annual turnover (KNBS, 2016), thus medium-sized enterprises often surpass the VAT threshold.

[46] See Meagher (2013) for a literature review.

a larger share of economic activity to take place outside the VAT system in more downstream sectors.

Moreover, formal firms in sectors and counties with higher levels of informality are more likely to transact with non-registered entities, such as consumers or firms outside the VAT system (final column of Table 4). This result could well be an outcome of differences in sales channels of firms. At the same time, similarly to the steep decline in average outdegrees outside the metropolitan areas, both could be driven by informality along supply chains as we might simply not observe links with informal firms.

4.3.3 Fact 3: Kenya's spatial concentration of economic activity is predominantly a feature of the formal sector

As a next step, we revisit the question of spatial concentration of economic activity. To achieve this, we expand Table 1 from Section 3 with additional measures of economic activity, presented in Table 7. We observe that spatial concentration becomes more pronounced as we move from less formal to more formal economic activities. 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. This pattern aligns with Obudho (1997)'s discussion of spatial concentration in economic activity back in 1992 when Nairobi accounted for 73% of formal sector employment in Kenya.

Table 7: Geographic concentration of economic activity by degree of formalisation

	Nairobi	**Mombasa**	**Pareto exponent**	
	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
Network sales	68	13	0.35	0.02
Network purchases	60	9	0.43	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): $\log \text{rank} = \log A - \alpha \log x$.

What have we learned from the stylized facts? First, we have shown that the VAT-paying sector only accounts for a third of the Kenyan GDP implying that the informal sector is significant enough to warrant further analysis. Next, we show that the incidence of informality varies systematically across sectors and geographies, rather than being evenly distributed across the economy. Informal firms can be found in downstream activities, which in turn are relatively more important in smaller markets. This suggests that a model-based approach that accounts for sectors and geography can be useful. Third, we have provided suggestive evidence that the spatial concentration described in the previous section might just be an artefact of the formal sector. We would therefore expect that accounting for the informal sector would systematically alter the structure of the observed production network. Correcting for this can have implications on our predictions about how connected specific sectors and counties are and how the economy responds to both domestic and trade shocks. To predict a counterfactual network that accounts

for informal firms, we introduce a theoretical framework that we estimate using the available data.

5 A network formation model with heterogeneity in sectors, regions, and firm size

We present and estimate the network formation model presented in Bramoullé et al. (2012). This model allows us to (i) predict the formal firm network as observed in the data and (ii) estimate a counterfactual network that accounts for informal firms.

5.1 Model motivation

Bramoullé et al. (2012) 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 extensive margin of trade, firm location and firm-to-firm links. In the stylized facts we showed that these two components account for the bulk of 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 allows us to study the question of spatial inequality. The size-dimension incorporates the widely documented positive correlation between firm size (measured using firm sales) and the number of firm-to-firm links (Bernard et al., 2022) as well as potential differences in the geographic and sectoral composition of the supply chains of small firms.

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 widely documented extreme heterogeneity in outdegrees across firms (Bernard et al., 2019; Panigrahi, 2022; Bernard et al., 2022; Bernard and Zi, 2022; Arkolakis et al., 2023).[47] We first present the dynamic network formation model proposed by Bramoullé et al. (2012) below and then discuss how we estimate it. The power law distribution is generated by preferential attachment (Barabási, 2014) i.e. link formation via existing links such that a new firm is more likely link to an existing firm that has more preexisting connections. Our framework is flexible and allows us to estimate the share of firm-to-firm links formed via preferential attachment versus undirected search (often referred to as *random search* in the networks literature (Jackson and

[47] Using this framework, we abstract away from additional complexities like endogenous firm entry and exit or the decision of a firm to formalise. We use empirical firm entry patterns to proxy for entry probabilities. An extension could micro-found the entry behaviour of firms. Micro-founding firm entry by itself would only add nuance to our current analysis if one were to further introduce an endogenous response in linking patterns to changes in entry.

Rogers, 2007; Bramoullé et al., 2012; Chaney, 2014)). The motivation at the core of this type of set up is that searching via existing suppliers allows firms to overcome information asymmetries about future supplier's quality type (Chaney, 2014).

Finally in this framework, we target the heterogeneity in firms' outdegree. By doing so, we hold the firms' indegree fixed at the empirically observed mean degree, while the firms' outdegrees evolve over time. This modeling choice is motivated by our stylized facts which show that the number of indegrees is distributed fairly evenly across firms and localities, while informal firms are more likely to operate downstream of formal firms. Hence we expect informality to primarily underestimate the observed number of outlinks a firm has.

5.2 Model setup

Consider an economy with a set of firms 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 pairs. i.e. all firms in the same sector, county, and size group are classified as the same type.

The network formation process is as follows. In every period t, a new buyer firm of type θ enters with probability $p(\theta)$. In order for its operations to be viable it needs to source inputs from suppliers through a fixed number of links m. It first chooses a sector-county-size pair (i.e. a type) with probability $p(\theta, \theta')$ for all θ' in Θ. Then, it forms m links with firms of the chosen type(s). The probabilities $p(\theta, \theta')$ represent the firm's bias in terms of sectors, regions, and firm size types it wants to link with. 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 (aka 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.

Let us note two important aspects of this process. First, while a firm's number of buyers evolves over time, the number of suppliers that a firm chooses is fixed to m and does not change as the network grows. 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. It is consistent with the third stylized fact from Section 3, documenting that the number of inlinks is more evenly

distributed across firms and localities relative to the number of outlinks. Second, the model includes "biases" about firm-to-firm linking patterns. 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 $p(\theta)$.

Table 8: Model parameters

Parameters	**Description**	**Source**	**Proxy**	**Value**
$p(\theta)$	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	Time periods	Data	Number of firms in admin data	56822
r	Share of suppliers via random search	Estimated	-	0.45

At the aggregate level, we are interested in the outdegree of each sector-county-size type. 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:

$$\pi_{t_0}^{t} = m\frac{r}{1-r}(f(t, \mathbf{B}) - \mathbf{I}) \tag{2}$$

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 $\pi_{\mathbf{t_0}}^{\mathbf{t}}$ 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. At the same time, the framework's intention is not to study the dynamics themselves, but rather consider the network's steady state properties.

5.3 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 from the data (see Table 8 for an overview). These include all probabilities $p(\theta) \in \Theta$ that a firm enters in a given sector, county, and size group as well as all interaction probabilities $p(\theta, \theta')$ between 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.[48] The only parameter we need to estimate is r, the fraction of input links a firm obtains via undirected search independent of the network environment.

First, we classify firms into types defined by unique sector-location-size combinations. Sectors refer to 13 aggregate sectors,[49] the location is 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.[50] 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 inlinks that it forms with type θ'.[51] 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 of the network. The variable t denoting time is found by dividing the total number of links in the 2019 network with the average indegree and is equal to 50,897.[52]

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 $r \in [0, 1]$. However, we face two concerns. First, note from Equation 2 that $\pi^t_{t_0}$ only tells us the expected outdegree of types born in t_0 evaluated at time t. Since a new firm is born in every period up until period t, we need to aggregate these matrices across all time periods leading up to t to get the type-by-type adjacency matrix of the network as a whole. The matrix of connections at time t is given by $\pi_t = \sum_{t_0}^{t} (\mathbf{p} \cdot \pi^t_{t0}{}')'$ where $\mathbf{p}$

[48] In particular, we use all firms and their linkages in the year 2019. We exclude a small proportion of firms that do not report buying from any other firm. This is because the model requires all entering firms to form m buying links with existing firms.

[49] Namely, agriculture, mining, manufacturing, utilities, construction, transportation and logistics, wholesale, hospitality, retail, business services, non-market services, other services, miscellaneous (incl. international organisations and non-classified firms).

[50] 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.

[51] The model also allows for self links. Wholesale firms in Nairobi, e.g., are able to buy from other wholesalers in the city.

[52] This is because the model predicts that m links are formed by buyers in every period implying that the total number of links in the network must be mt.

is a column vector with the probability that each type is born. For example, 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 $\mathbf{p} \cdot \pi^t_{t0}$. Then, we repeat the process again to compute the probability that a node of a certain type is born in time t_{0+1} and it's 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 must compute t such matrices and add them up to give us the type-by-type degree distribution at time t.

It is computationally difficult to compute π^t for t= 56822 in every iteration while looping through different candidate values of r during computation. As a result, in every iteration, we only compute π^t_{t0} for 500 representative time periods over which we then aggregate to obtain π_t. We space the sample of 500 time periods equally out between our first period $t_0 = 1$ and our final period $t_0 = 56822$. This approach ensures that we do not disproportionately sample from either older or younger nodes and hence bias our results. For example, sampling from nodes born in the first 500 periods will lead us to predict the type-by-type outdegree distribution only for firms at 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 is because older nodes are likely to be more connected. This can bias our estimation of r as we will be matching the predicted distribution of such firms with all firms observed in the data. As a result, we compute π_t= $\sum_{t_0=1:100:56822}(\mathbf{p} \cdot \pi^t_{t0}{}')'$. This implies that we will under-predict the average degree of the network as our model ignores firms born between specific time periods. At the same time, it ensures that our estimate of r is not dependent on including or excluding specific types of older or younger firms. Even if the network is scaled down in terms of number of firms, the features of the network are kept intact. Second, based on Bramoullé et al. (2012)'s formula for predicting π_t also 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 the matrix has very small entries afterwards.

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 r^*. Each moment is weighted by the probability with which we observe a specific sector-region type in the data.[53] In particular, r^* is defined as follows:

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

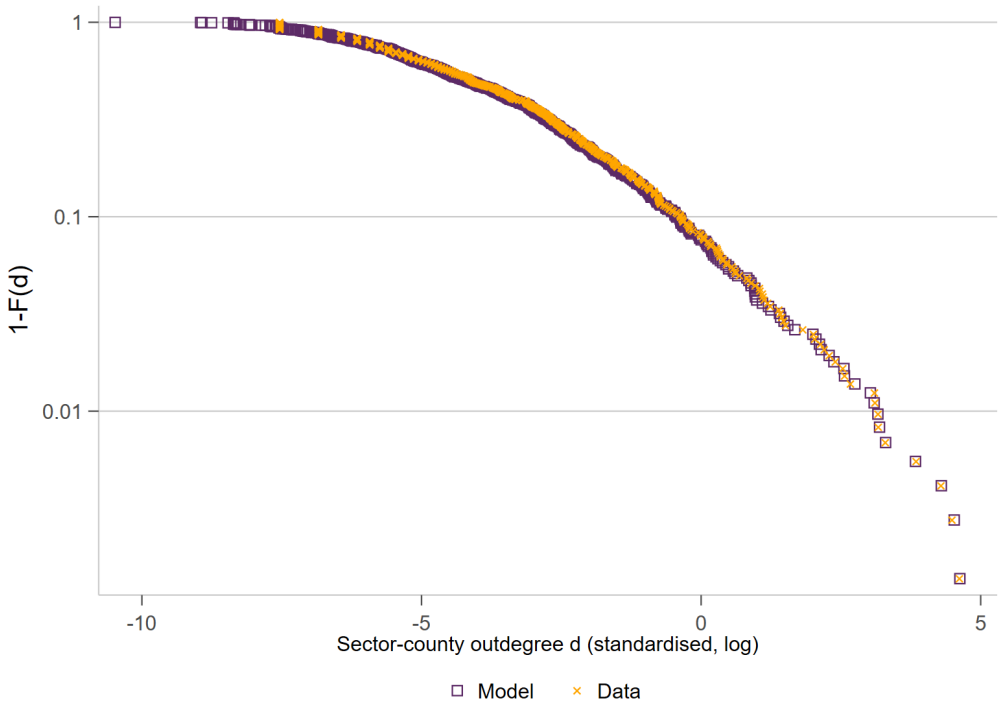
[53]In doing so, we assign greater weight to more common sector-region-size types whose probabilities tend to be more stable over time.

r^* is obtained by minimising the distance between the model predicted matrix of type-by-type interactions and the corresponding matrix obtained from the data. We estimate r using simulated annealing. Having to only estimate a single parameter comes with the advantage that we can plot the above objective function for various values of r to ensure that our estimated value is indeed the global minimum (see Figure C1).

5.4 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 randomly, 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. 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 for information asymmetries.

Figure 8: Model Fit: Actual and predicted outdegree distribution



This 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.

To assess how well our model does in fitting the targeted outdegree distribution, 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 8 shows that the key properties of the outdegree distribution are replicated by the model's predictions.[54] Model and data match particularly well in the right tail of the distribution and hence the part that is specifically targeted by the preferential attachment framework. Estimating the Parto exponent, 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.

6 Spatial inequality and unobservable informal firms - predicting a counterfactual network

With an estimated model at hand, we are now able to tackle the question of informal firms and spatial inequality in network links. 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 implement this counterfactual we rely on updated information on the spatial and sectoral dispersion of economic activity that accounts for the informal sector. In model terms, our counterfactual shifts the probabilities $p(\theta)$ with which we observe nodes of certain sector-region-size types θ to be born. We will use our population census-based measures of informality from Section 4 to update our estimates for all $p(\theta) \in \Theta$. Specifically, we attribute all informal economic activity to the types with small firms. 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.

The proposed counterfactual hinges on the assumption that the linking bias of informal firms follows the linking patterns of small firms in the administrative data. We assume that the type-by-type linking biases, i.e. $p(\theta, \theta')$, do not change even after informal firms are included in the model. Put differently we assume away any endogenous relationship between $p(\theta)$ and $p(\theta, \theta')$. Given informality correlates strongly with firm size, we allocate all informal firms to the sector-county-size cell of small formal firms. By doing so, we assume that conditional on the sector of operation and geography, informal firms have similar preferences about which other sectors and geographies they source from. As discussed in the stylized facts on firm-to-firm trade patterns, small formal firms tend to source more locally and from intermediaries. This finding is in line with multi-intermediary supply chains and sourcing patterns we would accordingly also expect from informal firms.

[54] As discussed, our predicted network will have a lower average degree than the real-world data. To compare the predicted degree distribution to the degree distribution in the data, we therefore standardise the outdegrees by dividing them by the mean of the respective degree distribution.

6.1 Predicting the sector-county profile of non-VAT firms

To incorporate informal firms into the network, we 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, we rely on the number of people who work in each sector-county-size cell to compute our alternative entry probabilities $p(\theta_a)$. :

$$p(\theta_{a,\,large}) = \frac{e_{sc}^{\text{formal}}}{\sum_{c}^{47}\sum_{s}^{13} o_{sc} + e_{sc}} \times \frac{1}{\overline{x}_{sc}^{\text{formal}}}$$

$$p(\theta_{a,\,informal}) = \frac{o_{sc}^{\text{formal}} + o_{sc}^{\text{informal}} + e_{sc}^{\text{informal}}}{\sum_{c}^{47}\sum_{s}^{13} o_{sc} + e_{sc}} \times \frac{1}{\overline{x}_{sc}^{\text{informal}}}$$

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

The denominator is equivalent to the total number of people who work in the private sector (employed e_{sc} and self-employed o_{sc} - both formally and informally)[55] across Kenya. For both $p(\theta)$ and $p(\theta_a)$ the sector-region-size probabilities sum to one.[56] This is important to keep in mind for the interpretation of the counterfactual: $p(\theta_a)$ captures 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. 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 $\overline{x}_{sc}^{formal,informal}$. We obtain the average number of employees of bigger formal sector firms from the administrative data.[57] For small formal and informal firms, we use the 2016 Medium, Small and Mirco Enterprise survey to compute the average number of employees (KNBS, 2016).

In addition to accounting for heterogeneity in firm size across sectors and geographies, we apply an exception for agriculture and non-market services: We estimate their $p(\theta_a)$ drawing only on

[55] The census allows us to distinguish between four different types of employment. Formal sector wage-employment, formal sector self-employment, informal sector wage employment, and informal sector self-employment.

[56] 13 and 47 refer to the 13 aggregate sectors and 47 counties respectively.

[57] If big formal firms employ informal casual workers (Ulyssea, 2018) not captured in the administrative data, we understate their size and hence the probability of big formal firms in the network. As a result our counterfactual network would be biased towards the baseline formal network.

formal private sector employment as formal VAT-paying firms occupy a very specific niche in both cases (see discussion in Section 4.2). Amending the $p(\theta)$s of agriculture and non-market services using records on total employment (instead of formal private sector employment only) would greatly overestimate the number of firms that are operating in these sectors and participate in a similar manner in the private sector network to their peers.

What are our expectations about the differences between the probability $p(\theta)$ that a formal firm enters in a given sector-county-size cell versus the alternative probability $p(\theta_a)$ that takes informal ones into account? For sectors and regions with high levels of informality we expect $p(\theta_a)$ for small formal and informal to be larger than $p(\theta)$. Sectors and counties with a high degree of formality will have a $p(\theta_a)$ that is smaller than $p(\theta)$. Put differently, their importance in the economy is overstated by the administrative data. Figure C2 shows the correlation between updates to firm entry probabilities and the respective sector-region formality shares. 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).[58] We plot both our preferred measure accounting for heterogeneity in firm size as well as a plain vanilla version that solely relies on employment shares (the first term of the above equations). The only notable difference is that correcting for heterogeneity in firm size yields higher (but not significantly) entry probabilities in sector-county-size cells with the lowest formal sector share.

In a nutshell, our suggested counterfactual accounts for informal firms being born into the network based on their sector-region profile. Rather than thinking of the counterfactual as adding new firms, we adjust the weights of each sector-region-size type. At the same time, due to lack of granular data on linking preferences of informal firms, we rely on linking probabilities $p(\theta, \theta')$ for small formal firms. We thereby treat $p(\theta, \theta')$ as a fundamental production technology conditional on a sector-region-size profile. This assumption is plausible if the barriers for informal firms to link with firms from other sectors and regions result in linking patterns similar to those of small formal firms. For example, consider a small formal retailer and an informal retailer both wanting to purchase soap. Neither can source directly from the manufacturer in Nairobi. The small formal retailer might purchase soap from a larger formal wholesaler in the same town. However, the wholesaler might not trade with the informal retailer due to concerns about the firm's tax status, so the informal retailer buys from another wholesaler in town. Despite accessing different trading partners, both the informal and the small formal retailer exhibit similar sectoral and regional buying patterns.

[58] To estimate the slope in Figure C2, 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 sectors-county pairs are included.

6.2 Counterfactual results

How does the predicted counterfactual network that accounts for informal firms compare to the baseline network? First, firm types in sectors and counties with a high incidence of informality are predicted to have a relatively larger increase in outdegrees (Figure C3). Second, the variation in outdegrees across counties reduces by 7.5% if we account for informal firms (Table 9). We visualise this reduction in inequality by plotting the Lorenz curve for county-level outlinks in Figure C5.

Table 9: County-level changes in the dispersion of outdegrees

County outdegree	Δ sd/mean (in %)
Default	-7.5
Default without Nairobi & Mombasa	-18.0
Alternative	-5.6
Alternative without Nairobi & Mombasa	-11.8

The above table reports the difference in outdegrees between the original and the counterfactual 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 (see Section 6.1). The first two lines capture the results from the default counterfactual. The difference between one and two is that we exclude the outdegrees of Nairobi and Mombasa when we compute the coefficient of variation in the second row. Row three and four report the results for alternative distributions of $p(\theta)$ that only rely on employment information in the population census and do not adjust for heterogeneity in firm size across sectors and regions when computing $p(\theta)$.

What are key drivers of this shift in inequality in outdegrees? First, trade shifts away from Nairobi and Mombasa. In Figure 9 we plot the row-normalised adjacency matrices, before and after accounting for informal firms, at the county and sector level respectively.[59] 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 counterfactual. This is in line with the stylized facts in the previous section showing that smaller firms are more likely to source from the same county.[60] 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 C6 where we compare the change in both inter- and intra-county links for the baseline and counterfactual 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.[61]

[59] Each row sums to 1 in the normalised adjacency matrix.

[60] If informal firms purchase an even higher share of their inputs locally, the predicted shift towards intra-county trade represents a lower bound.

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

If counties are selling less to Nairobi and Mombasa, where do inter-county trade links shift? We find that the number of bilateral trade links now becomes more sensitive to social ties between two counties. In Table C1, we regress the number of links between two counties on both travel distance and social connectedness (Bailey et al., 2021).[62] The share of inputs sourced from another county now correlates more strongly with the strength of social ties between the two counties.

Both the increased importance of within county trade and trade with counties with stronger social ties give rise to a network with a stronger 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 if it contains groups of nodes that have more links among each other than one would predict if links were formed at random. We compute the modularity of the weighted adjacency matrix at the sector-county level. We find that modularity in the counterfactual network with informal firms increases by about 60%, which suggests that the counterfactual network exhibits a more pronounced group structure. We apply a community detection algorithm to further understand this group structure using the trade flows between counties as predicted by the model and counterfactual. As illustrated in Figure C4, we find that the group structure in the counterfactual network is significantly correlated with Kenya's geography i.e. geographically proximate counties are now more likely to be clustered in the same group.

Similar to this shift in the patterns of inter-county trade, we also detect a change in inter-sector trade patterns as seen in Figure 9. Sectors with a lot of informal activity like other services, retail, and wholesale now gain in prominence in the network acting as buyers. Consequently other sectors start selling more to them. Overall manufacturing, wholesale and mining gain the most number of new links in relative terms.

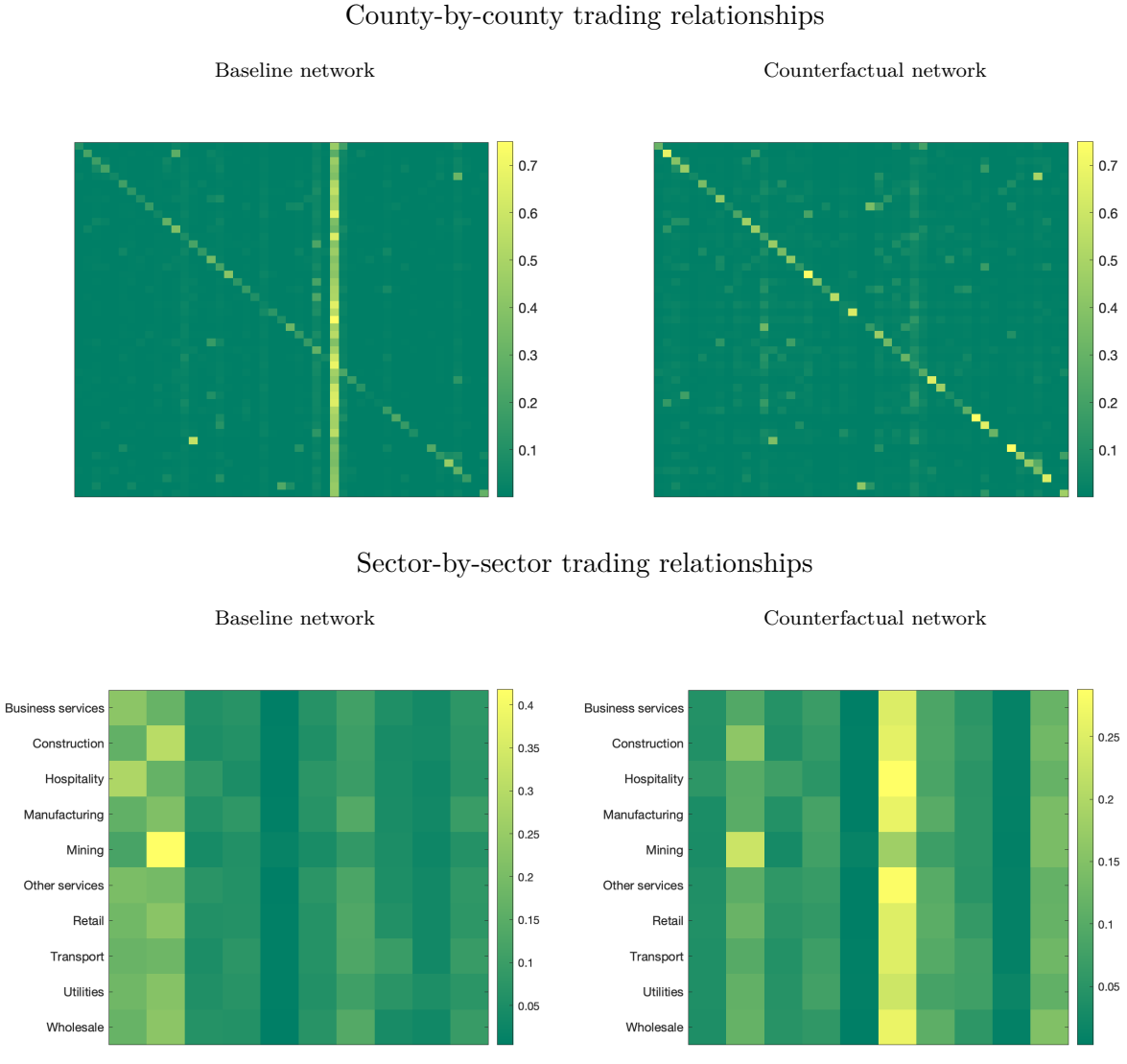
Finally, we explore in which types of counties we tend to underestimate the number of links the most. To do so, we regress the change in county-level outlinks between baseline and counterfactual on key county characteristics like the aggregate level of formality, population, Gross County Product, and market access[63] in Table C2. We find that relying on formal sector data only in our baseline network, we tend to particularly underestimate how connected firms in smaller counties and counties with a high market access are. The correlation with a county's aggregate share of formality is no longer statistically significant once we account for these other characteristics.

counties in the counterfactual network.

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

[63] We rely on county-by-county travel distance and county population to compute market access.

Figure 9: Baseline vs counterfactual 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).

7 Simulating the effect of economic shocks

As a natural 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. Are sectors and regions with more informality more or less vulnerable to shocks than the administrative data would suggest? How does the predicted impact of the shock depend on whether we account for informality? 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 common international supply shocks that affect all firm types given their exposure to international markets. We discuss both the results for both domestic and international shocks below.

7.1 Domestic shocks

Following the supply side version of classic input-output models (Sargent and Stachurski, 2022),[64] firm type j's output y_j in period t as a function of the sum of intermediate inputs it purchases from other types i plus payments to other factors of production (value added) v_{jt}:[65]

$$y_{jt} = \sum g_{ij} y_{it-1} + v_{jt} \tag{3}$$

The intermediate inputs purchased from other firm types in turn are the product of the supplier's total output in the previous period y_{jt} and the fraction it sells to type i i.e. g_{ij}. g_{ij} captures the share of inputs j obtains from i. It represents the individual cells of matrix π. We normalise the rows of π by dividing each entry in a row by the sum of that row – this implies that a firm's total output is equal to a weighted average of the outputs of its suppliers. We assume that v_{it} is an independent draw from a uniform distribution $U[-10, 10]$ for every type i in every time period t. We also assume that each type starts with a randomly drawn output drawn from the distribution $U[0, 100]$ in $t = 0$. Using this set-up, we first simulate the output process without any shock and then simulate the output process after a negative output shock to sector-region-size type i in the first time period.[66] We repeat this exercise for all types i.

To study the relevance of unobserved informal firms, we conduct our simulation exercise twice. In the first scenario, we use the matrix π derived from administrative records. In the second scenario, we use an alternative version of π that incorporates the predicted type-by-type linkages, accounting for the presence of informal firms.[67] Our primary question is: how does the shock impact a type j when informal firms are considered versus when they are not? For a fixed type i that is shocked in the first period, we compute the following measures for each $i \neq j$: (i) the absolute reduction in output of i using the original adjacency matrix (excluding informal firms) and (ii) the absolute reduction in output of i using the new adjacency matrix (including informal

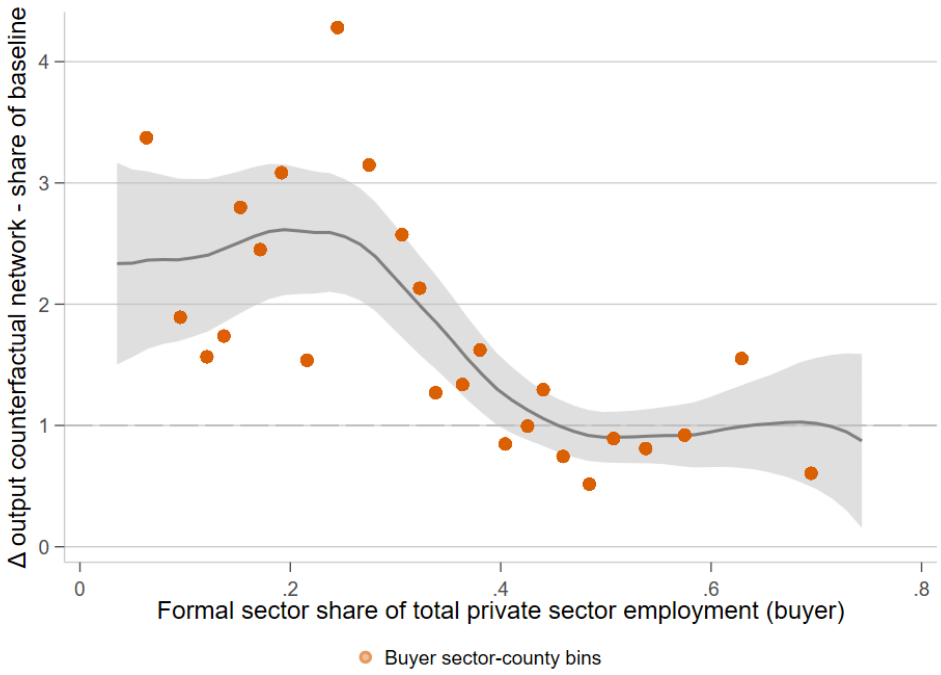
[64] Note, for this simple illustration we do not account for any endogenous network adjustments (see e.g. Panigrahi, 2022; Arkolakis et al., 2023)

[65] Alternatively, v_{jt} can also be interpreted as a type-specific and period-specific shock to output. We remain agnostic about the interpretation as it does not affect the core result of this section.

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

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

Figure 10: How do output shocks pass-through in a counterfactual network that takes into account informal firms? - % Difference in output drops and the level of formality



The outcome of interest measures the % change in the output reduction in response to an adverse shock if we account for informal firms vs rely only on the administrative data. The above graph plots the output of our shock simulation under the two different networks on the y-axis. The x-axis captures the respective sector-region type's formal sector shares. The scatters represent formalisation bins. The plotted polynomial is estimated using the underlying type-level data.

firms).[68]

In Figure 10, we plot the output reduction accounting for informal firms as a share of the baseline scenario without informal firms on the y-axis, and the formal sector share at the sector-region level on the x-axis. The figure illustrates that relying solely on formal sector data may lead to underestimating the exposure of sectors and regions with high levels of informality to shocks. Specifically, the higher the incidence of informality in a sector and region, the more we underestimate the adverse impact of a domestic output shock. Notably, a one percentage point decrease in the formal sector share corresponds with a 3.8 percentage point larger drop in the sector-region's output (Table C3).

Figure C7 presents the distribution of the ratio of the shock impact using the baseline network versus the counterfactual network. We observe that this ratio exceeds 1 for 42% of the types, indicating an underestimation of the shock impact in these cases. Notably, among the 42% of types where the shock impact is underestimated when informality is not accounted for, 73% are types with small firms. This finding suggests that the underestimation is primarily driven by the omission of small, informal firms in the domestic economy.

[68] For both measures, we consider the average output reduction across all time periods.

7.2 Import shocks

In addition to a domestic shock, we consider the impact of a reduction in output as a result of an adverse shock to import markets of Kenyan firms. As before, firm j's output can be written as follows:

$$y_{jt} = \sum g_{ij} y_{it-1} + m_{jw} y_w + \upsilon_{it} \tag{4}$$

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 for various firm types.

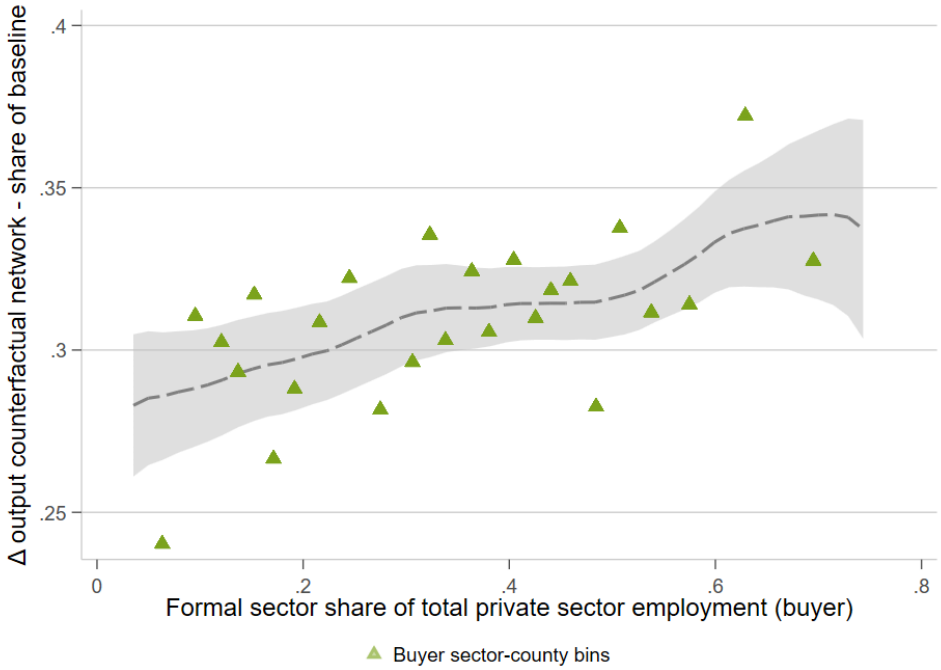
The bottom graph of Figure C7 plots the impact of the shock relying on the counterfactual network as a proportion of the impact using the baseline network. Unlike domestic shocks, our findings for import shocks indicate that extrapolating from data on the formal sector network to the overall economy leads to an overestimation of the reduction in output. In fact, the impact of the shock is always lower (i.e., less negative) when considering the counterfactual network compared to the original network.

Furthermore, this effect is particularly pronounced in sectors and regions where informal economic activity constitutes a larger share (see Figure 11). Table C3 shows that a 10 percentage point increase in the informal sector share corresponds to a 1 percentage point overestimation of the output reduction. This effect is driven by formal firms operating in predominantly informal markets which are now less exposed once we take into account their connections which we do not observe directly in the administrative data.

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, accounting for informality reduces the prominence of sectors with a high share of formal activity. These types in turn are more likely to have higher import shares and greater exposure to international markets. 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.[69]

Figure 11: How does a shock to import markets pass-through in a counterfactual network that takes into account informal firms? - Output drops and the level of formality



The outcome of interest measures the % change in the output reduction in response to an adverse shock if we account for informal firms vs rely only on the administrative data. The above graph plots the output of our shock simulation under the two different networks on the y-axis. The x-axis captures the respective sector-region type's formal sector shares. The scatters represent formalisation bins. The plotted polynomial is estimated using the underlying sector-county data.

8 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 informal economic activity is not evenly distributed across space and sectors. Informality is more prevalent in downstream sectors and smaller regional markets within Kenya. Consequently, relying solely on formal sector data leads to an overestimation of the spatial concentration of overall domestic trade flows. We show this both in empirical data as well as using a structural model that allows us to predict a counterfactual network that accounts for the dispersion of informal activity across space and sectors.

Our findings indicate that formal sector data underrepresents trade within counties and trade between counties that have strong social ties. This has implications, for example, for predic-

[69] 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.

tions about the propagation of shocks. In a simulated output shock, we show that the higher the incidence of informality, the more we might underestimate the economic impact of an adverse reduction in economic output. Conversely, when considering international trade shocks, extrapolating from formal sector data about aggregate output consequences might overstate the impact. This is because it assigns too much weight to formal firms with strong links to international markets, whereas, in reality, the overall economy's ties to import markets are much weaker.

An important question for future research, beyond the scope of this paper, is whether the observed spatial concentration of formal sector firm networks results from market frictions or is a feature of structural transformation (Gollin, 2008). Understanding the drivers of this spatial concentration will enable policy-relevant statements about the optimal spatial distribution of formal economic activity.

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-Urena, A., Fuentes, M. F., Manelici, I. and Vásquez, J. P. (2018), 'Costa rican production network: Stylized facts', *Research Paper Banco Central de Costa Rica* **2018**(2).
URL: *https://jpvasquez-econ.github.io/files/Costa$_R$ican$_P$roduction$_N$etwork$_S$tylized$_F$acts.pdf*

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.

Almunia, M., Hjort, J., Knebelmann, J. and Tian, L. (2022), 'Strategic or confused firms? evidence from "missing" transactions in uganda', *Review of Economics and Statistics* pp. 1–35.

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.
URL: *https://doi.org/10.1146/annurev-economics-081919-041554*

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.

Baqaee, D. R. and Farhi, E. (2024), 'Networks, barriers, and trade', *Econometrica* **92**(2), 505–541.

Barabási, A.-L. (2014), *Network Science: Degree Correlation*, CC BY-NC-SA 2.0.
URL: *https://barabasi.com/f/620.pdf*

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

Bernard, A. B. and Zi, Y. (2022), Sparse production networks, Working Paper 30496, National Bureau of Economic Research.
URL: *http://www.nber.org/papers/w30496*

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.

Cardoza, M., Grigoli, F., Pierri, N. and Ruane, C. (2023), 'Worker mobility in production networks', *Working paper* .

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., Michalski, T. K. and Ors, E. (2022), 'Financial Constraints and Propagation of Shocks in Production Networks', *The Review of Economics and Statistics* pp. 1–46. **URL:** *https://doi.org/10.1162/rest_a_01162*

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

Dhyne, E., Magerman, G. and Rubínova, S. (2015), The belgian production network 2002-2012, Technical report, National Bank of Belgium.

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.

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.

Gabaix, X. and Ioannides, Y. M. (2004), The evolution of city size distributions, *in* 'Handbook of regional and urban economics', Vol. 4, Elsevier, pp. 2341–2378.

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**.

Henderson, V. (2002), 'Urban primacy, external costs, and quality of life', *Resource and Energy Economics* **24**(1-2), 95–106.

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'.

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.
URL: *https://www.knbs.or.ke/?wpdmpro=2019-kenya-population-and-housing-census-volume-i-population-by-county-and-sub-county*

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

Kose, M. A., Ohnsorge, F., Yu, S., Amin, M., Celik, S. K., Kindberg Hanlon, G. J., Islamaj, E., Kasyanenko, S., Okou, C., Sugawara, N., Taskin, T. and Collette, W. (2019), *Global economic prospects: A World Bank Flagship Report*, World Bank, chapter Growing in the Shadow: Challenges of Informality, pp. 128–195.

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, Technical report.

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. (2023), 'Matching and agglomeration: Theory and evidence from japanese firm-to-firm trade', *Working Paper* .

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* .

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.

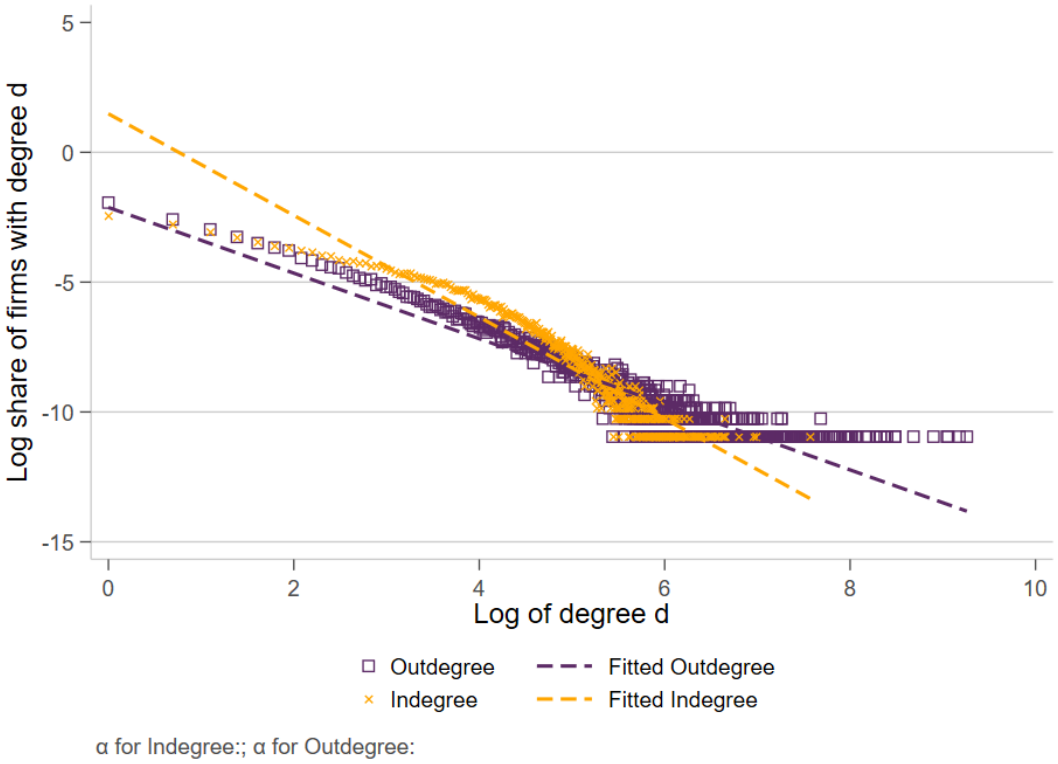
Zhou, Y. (2022), The value added tax, cascading sales tax, and informality, *in* M. Bussolo and S. Sharma, eds, 'Hidden Potential: Rethinking Informality in South Asia', World Bank Publications, chapter The Value Added Tax, Cascading Sales Tax, and Informality, pp. p. 61–90.

Appendices

Appendix A Complementary material for Section 3 on spatial trade patterns

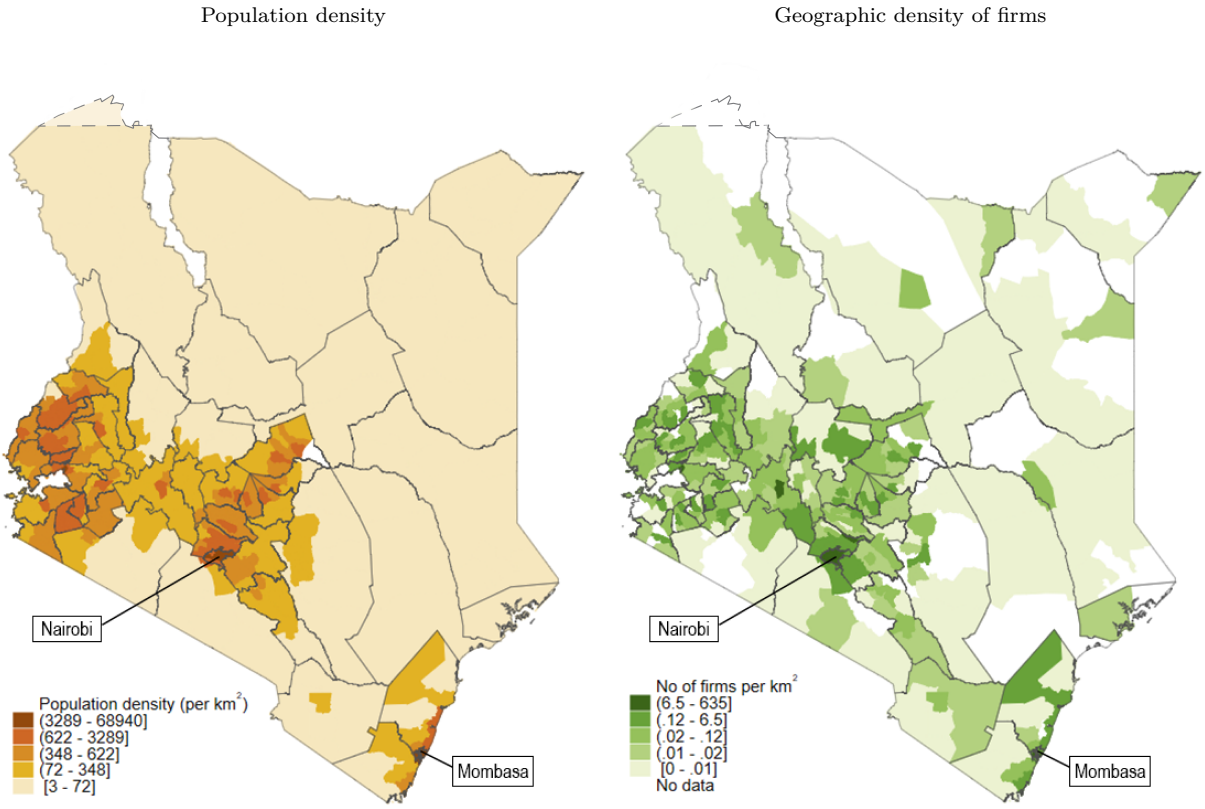
A.1 Additional graphs and tables

Figure A1: Degree distributions



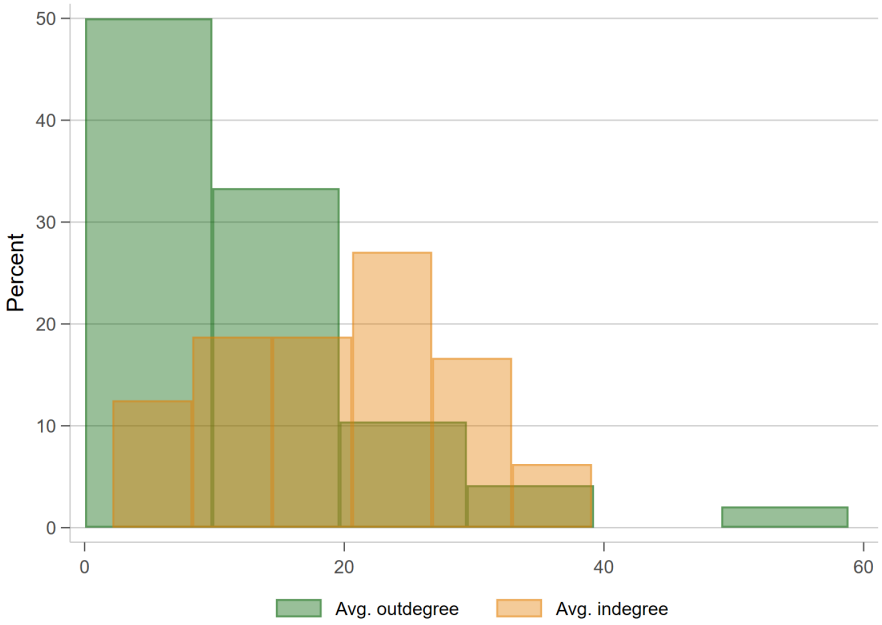
The figure plots the log-log plot of the probability density function (PDF) against firm outdegree and indegree respectively. The coefficients α shown at the bottom of the plot correspond to the power law exponent indicating the existence of a heavier tail for the outdegree distribution.

Figure A2: Firm headquarter locations and population density



The right map shows the density of firm headquarter locations at the sub-county level, i.e. the number of firms per km^2. The left map shows the population density - also at the sub-county level. Sub-counties represent the second administrative layer. Their size varies between 3 and 19,837 km^2 with a median size of 1,738 km^2 and an average size of 421 km^2. We therefore chose to map the density of firms rather than absolute numbers. Sub-counties are much more comparable in terms of population. The median sub-county has a population of 143,156 people, while the average sits at 129,263. The borders of Kenya's 47 counties, the first administrative layer are outlined in grey.

Figure A3: County-level average in- and outdegree



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

A.2 Exploring the robustness of spatial concentration with respect to 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 A1. 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[70] in mining, manufacturing and utilities, while KNBS (2010) covered 2,252 firms five years earlier.

48% of all firms involved in industrial production are located in Nairobi County generating as

[70] 2015 is the earliest year for which the VAT records have been fully digitised. 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.

much as 61% of total sales in 2010.[71] 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 higher spatial concentration observed among VAT-registered firms compared to overall economic activity.

Table A1: 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). 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): $\log \text{rank} = \log A - \alpha \log x$. The Census of Industrial Production was carried out by KNBS (2010).

[71] 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).

Appendix B Complementary material for Section 4 on informality

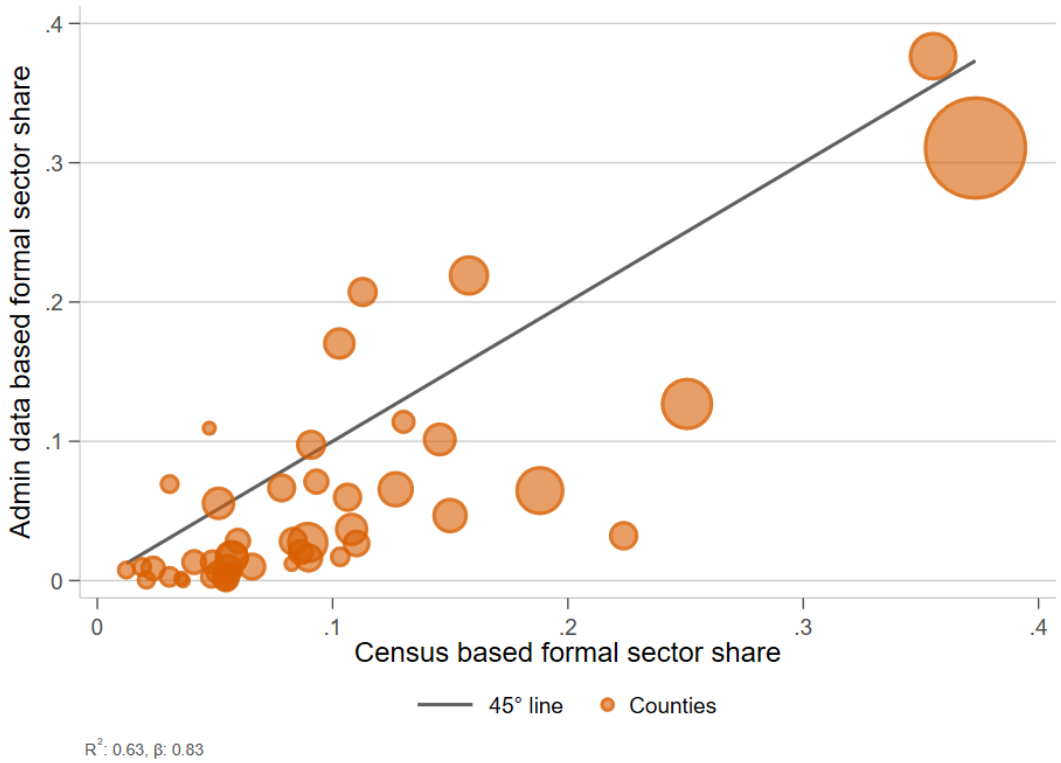
B.1 Measures of informality

Table B1: Overview of benchmark data

Name	Year	(Dis-)Aggregation	Key indicators
Small & Medium Sized Enterprises Survey (MSMEs)	2016	firm-level	main input source and buyer
Census of Establishments (CoE)	2017	sector OR county	# of formal sector establishments
Gross County Product (GCP)	2019	sector AND county	gross county product
Population & Housing Census (Census)	2019	sector AND (sub)county	formal & informal employment

All data are collected and published by the Kenya National Bureau of Statistics. **Sources:** Small & Medium-Sized Enterprises Survey https://statistics.knbs.or.ke/nada/index.php/catalog/69 KNBS (2016); Census of Establishments https://www.knbs.or.ke/download/report-2017-kenya-census-establishments-coe/ KNBS (2017); Gross County Product https://data.humdata.org/dataset/kenya-gross-county-product-gcp-by-economic-activities-per-county and KNBS (2022); 2019 Kenya Population & Housing Census https://www.knbs.or.ke/publications/# *2019 Kenya Population and Housing Census Volume IV: Distribution of Population by Socio-Economic Characteristics* KNBS (2019).

Figure B1: 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.

Table B2: 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.

As documented in Table B2, 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. Employment in licensed businesses (second row) is likely to be concentrated in the large firms of this population and hence aligns more strongly the estimates based on the administrative data.

B.2 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. Figure 5 shows that only a fraction of the sector's GDP is captured in the VAT data.
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 being estimated using non-market prices (see penultimate column of Table B3). Figure 5 highlights another sizeable gap for "others", which includes international organisations, unclassified firms, and financial services.

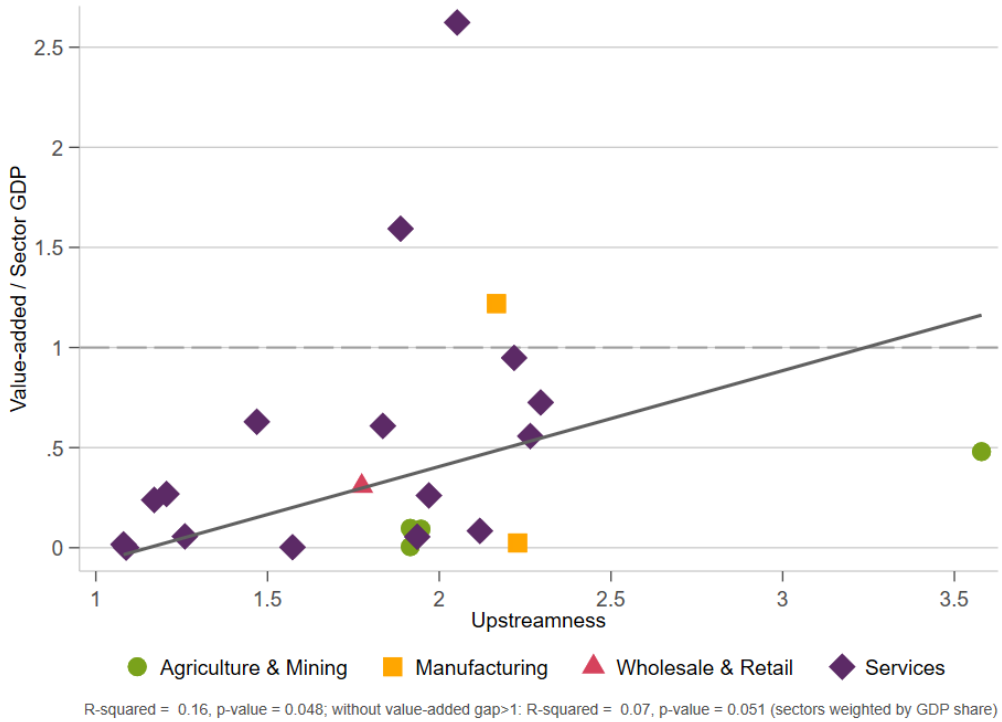
Table B3: Share of GDP covered in the administrative records

	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 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.

Table B3 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. Firstly, 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. Secondly, 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.

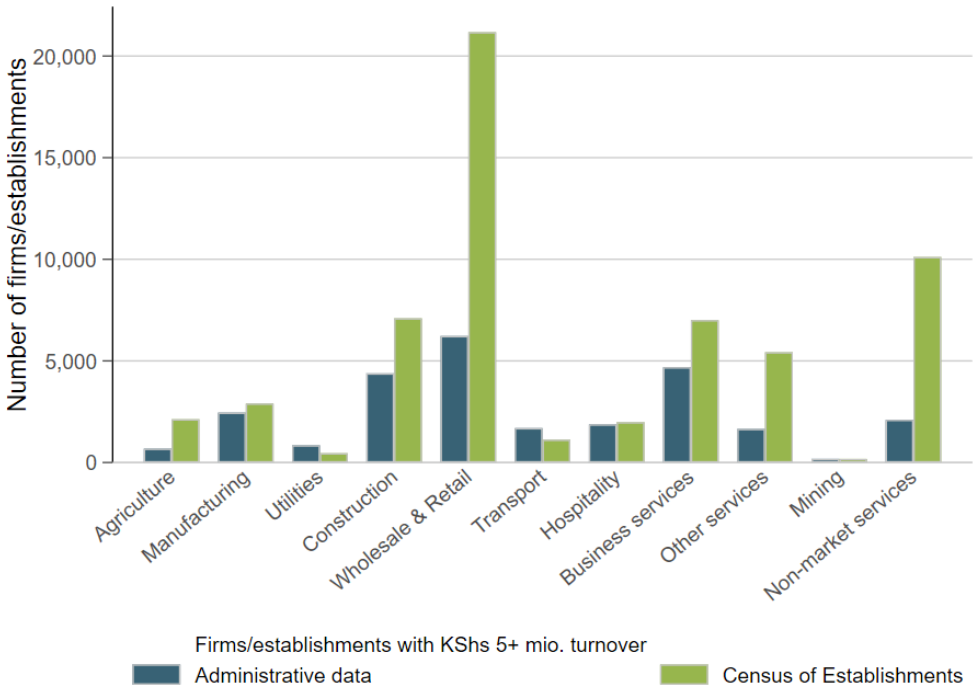
Figure B2: The GDP/value-added gap and upstreamness



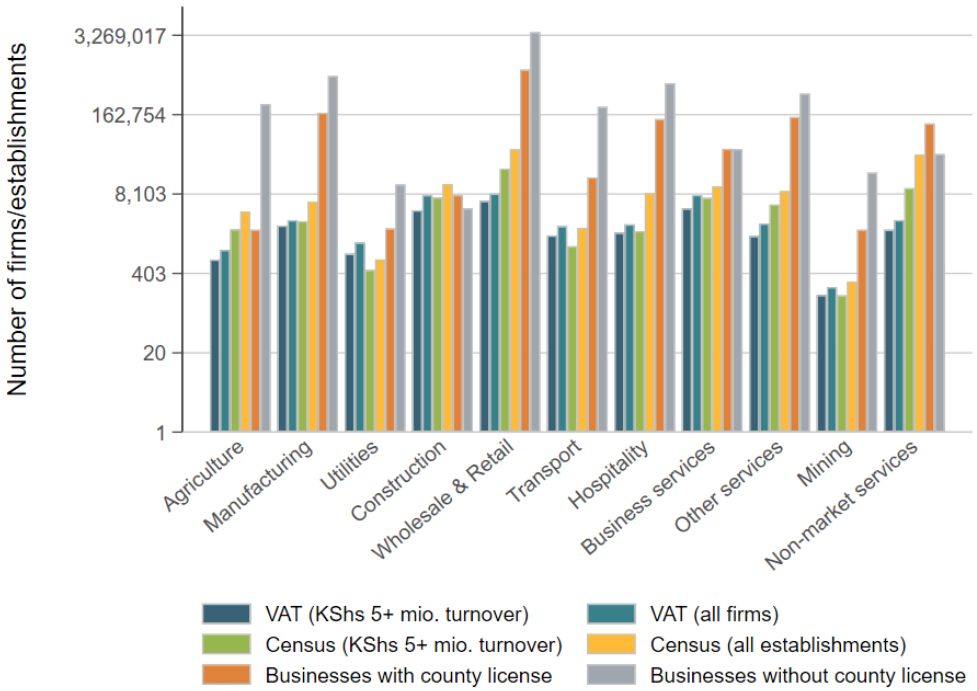
We plot the gap between value-added in the VAT and national accounts figures at the sub-sector level for the most granular sector classification reported in national accounts. We correlate it with a measure of upstreamness (Antràs et al., 2012), which captures how removed a sector is from final consumers (it takes a value of 1 if the sector sells everything directly to final consumers).

Figure B3: The extensive margins of informality - in which sectors do informal firms operate?

Firms above the threshold, but which don't report



Firms below the VAT threshold (log scale)



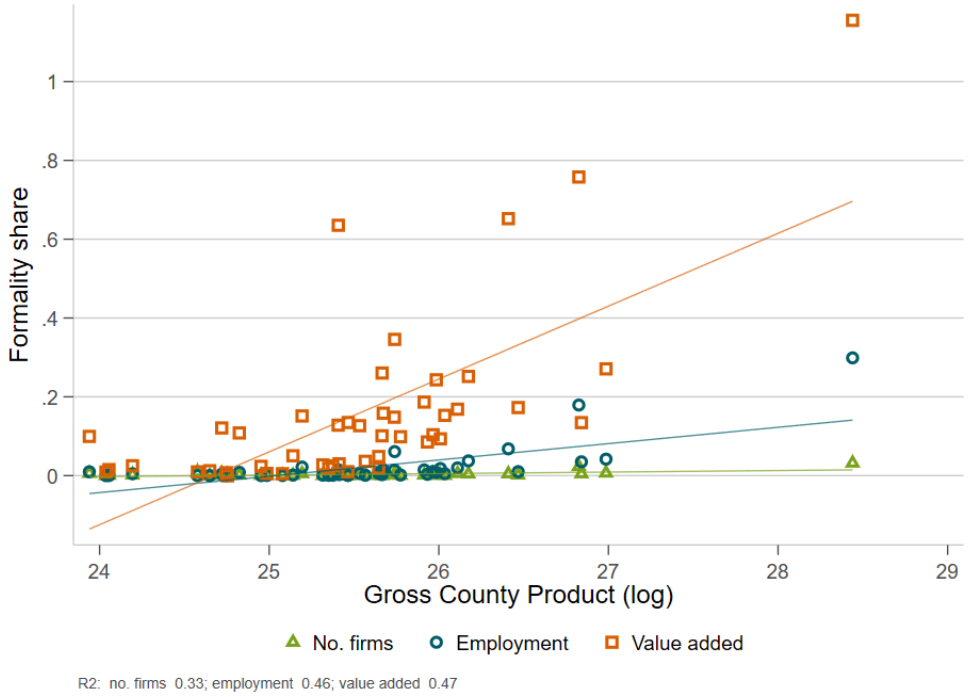
The top graph compares the number of firms covered in the administrative data and had an annual revenue of over KShs 5 million in 2016 to the number of firms with annual revenues above KShs 5 million in the 2016 Census of Establishments (CoE) (KNBS, 2017). The bottom graph compares the two groups of firms to all firms in the VAT data and the CoE, irrespective of their performance in 2016, and the number of licensed and unlicensed businesses reported by KNBS in KNBS (2016).

B.3 Informality, market size, and income levels

Figure B4: Informality, market size, and income levels

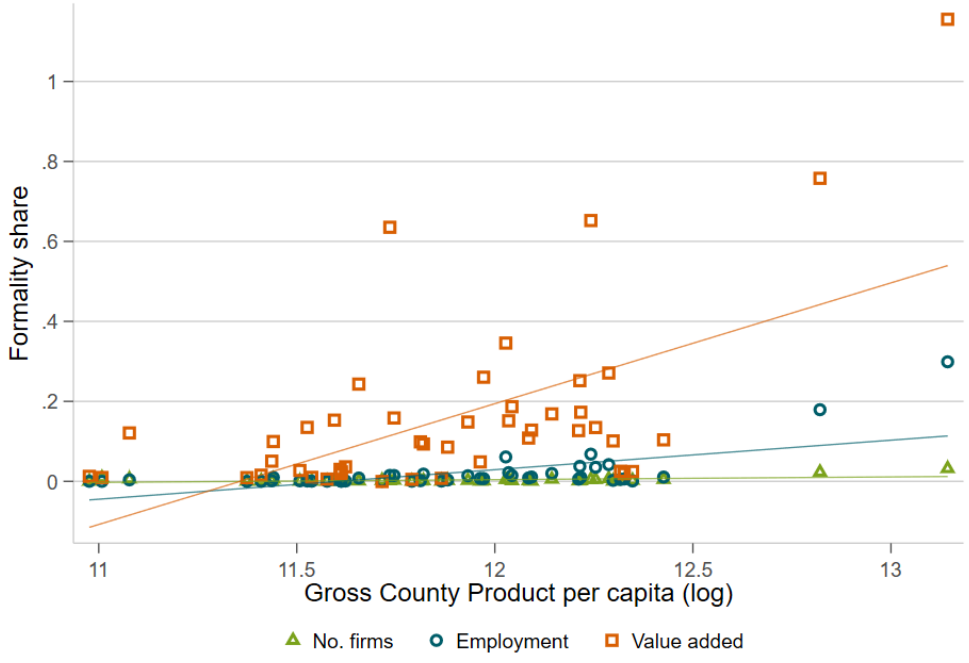
Correlation of the formal sector share and ...

... Gross County Product



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

... Gross County Product per capita

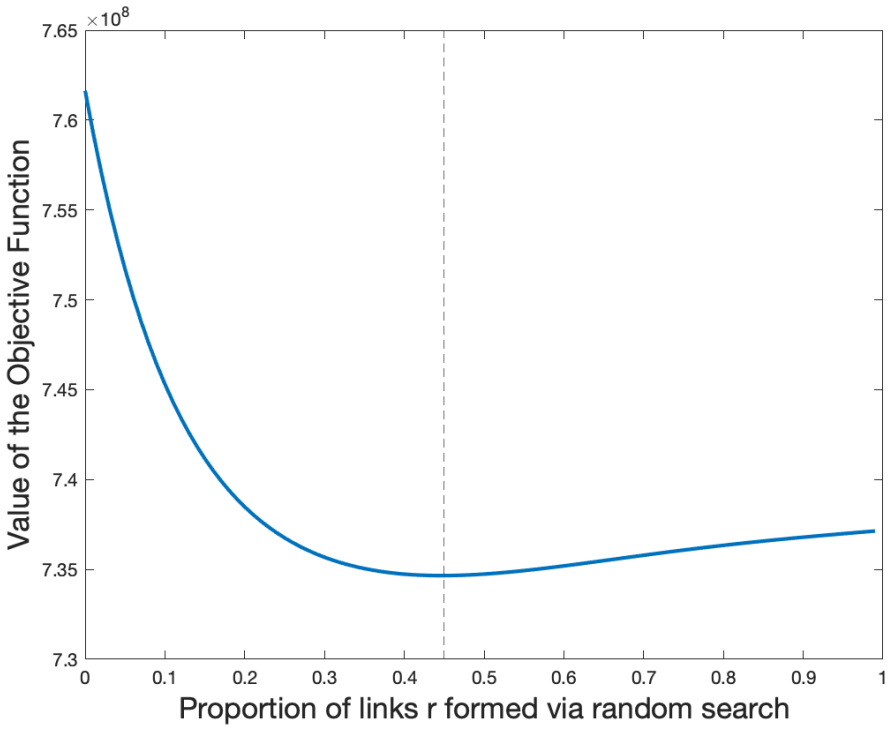


R2: no. firms 0.28; employment 0.40; value added 0.34

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.

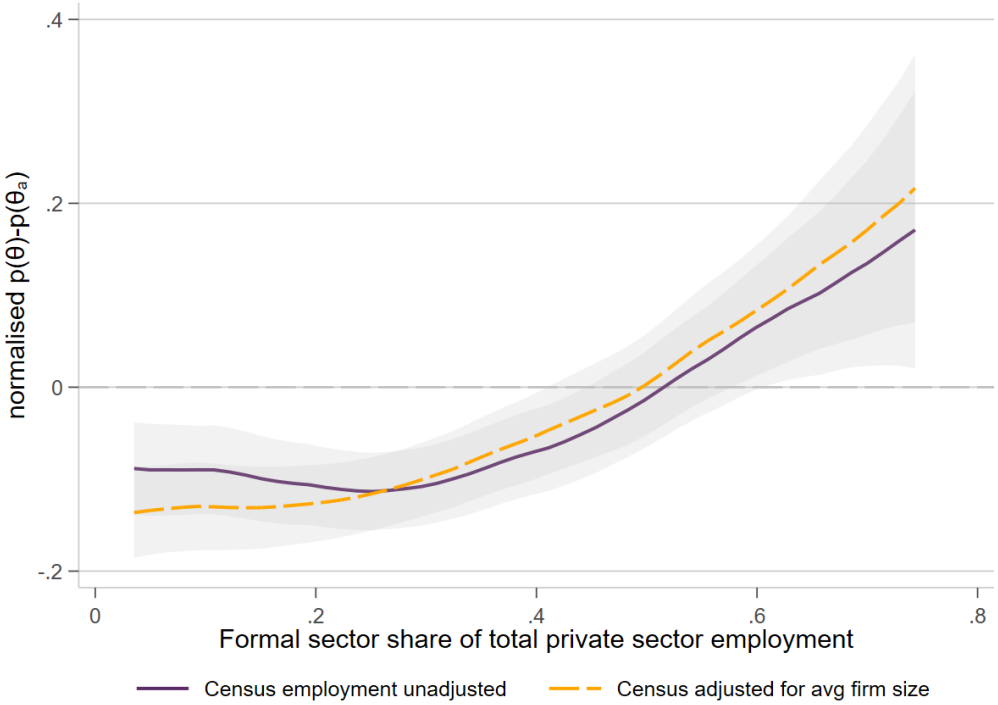
Appendix C Complementary material for Sections 5 and 6 on the model and counterfactual

Figure C1: Objective function for various values r



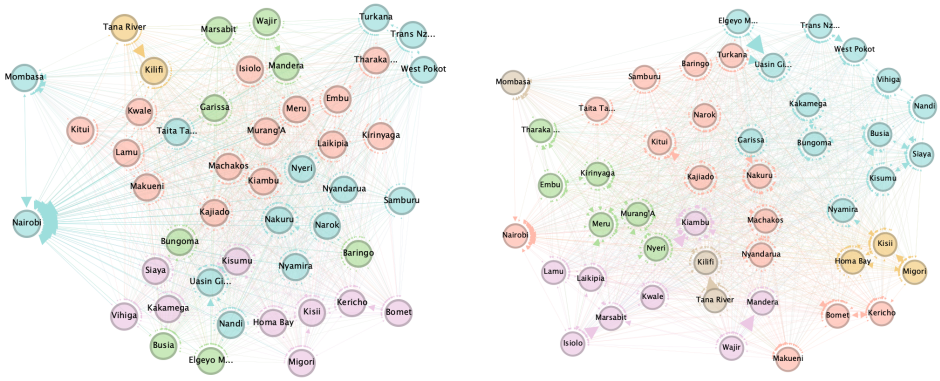
This figure plots the sum of the squared difference between each element of the model predicted interaction matrix π and the matrix π directly observed in the data, for various values of the parameter $r \in [0, 1]$. The figure shows that $r^* = 0.45$ obtained via simulated annealing minimises the objective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

Figure C2: Sector-county-size probabilities and formal sector shares



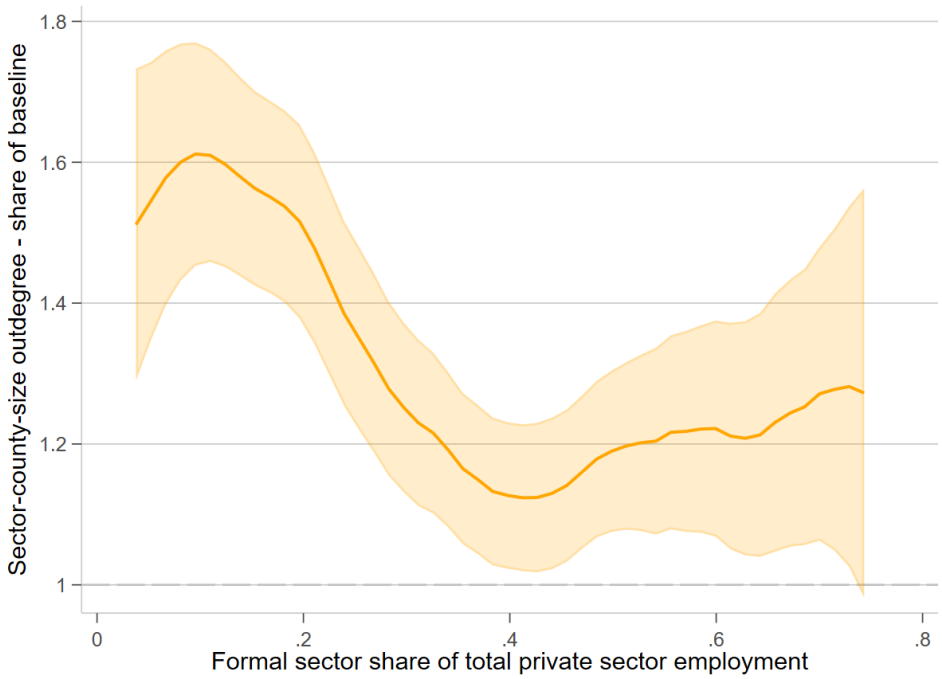
The graph plots each sector-regions formality share against the normalised difference between the baseline $p(\theta)$ and the augmented version $p(\theta_a)$ that takes into account informal firms. $p(\theta)$-$p(\theta_a)$ is reported in terms of standard deviations.

Figure C4: Model and counterfactual network plots at the county level



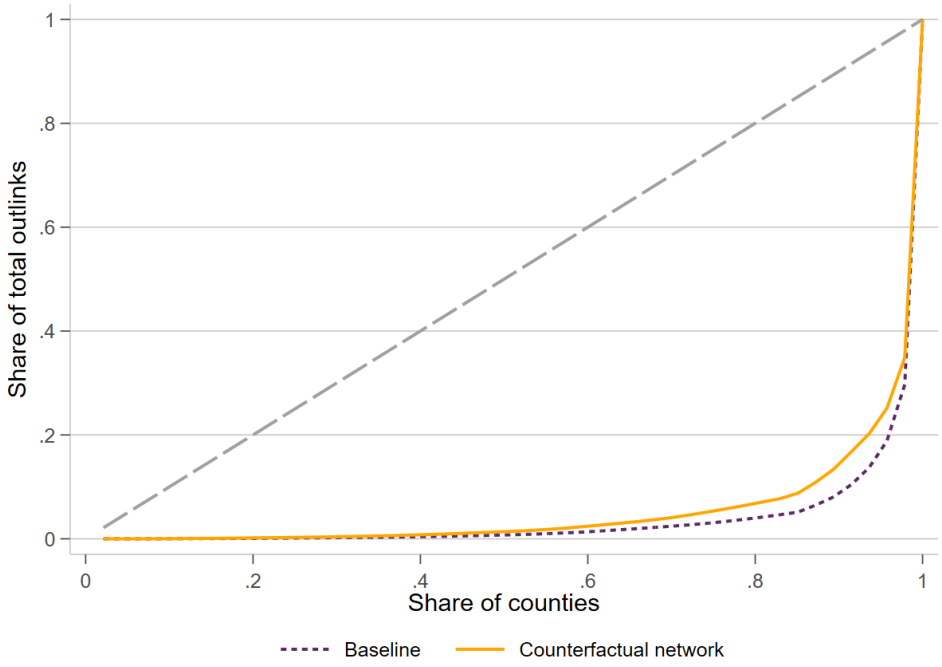
The graph on the left plots the network at the county level as predicted by the model while the graph on the right plots the network as predicted by the counterfactual using the augmented probabilities. We use the row-normalised adjacency matrix to construct the plots where arrows indicate links from suppliers to buyers. Counties in the same color are grouped together using a community detection algorithm.

Figure C3: Predicted change in type-level outdegree and formal sector shares



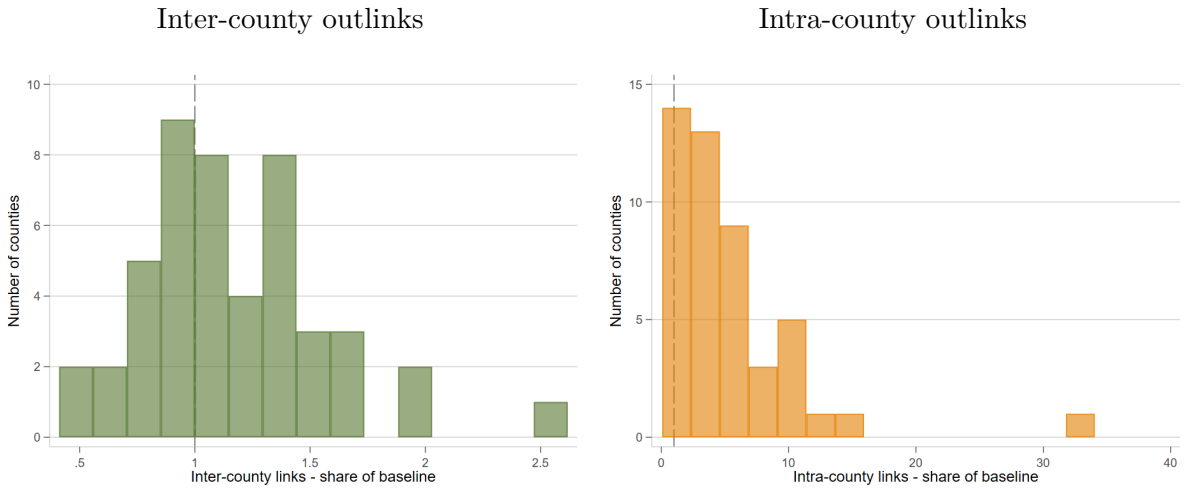
We plot the formal sector share at the sector-county level against the relative change in the type-level outdegree comparing the baseline network with formal firms only to the counterfactual network that also includes informal firms.

Figure C5: Inequality in county-level outlinks in the baseline and counterfactual network



To visualise the change in inequality between the baseline and the counterfactual network, we plot the Lorenz curve for the number of outlinks at the county level.

Figure C6: Inter- and intra-county trade patterns in a counterfactual network



The figure plots the ratio of the supplier to buyer links for the counterfactual 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: Social connectedness, travel time and county-by-county-links

	Any		Without within county trade	
	Baseline	Counterfactual	Baseline	Counterfactual
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 C2: 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

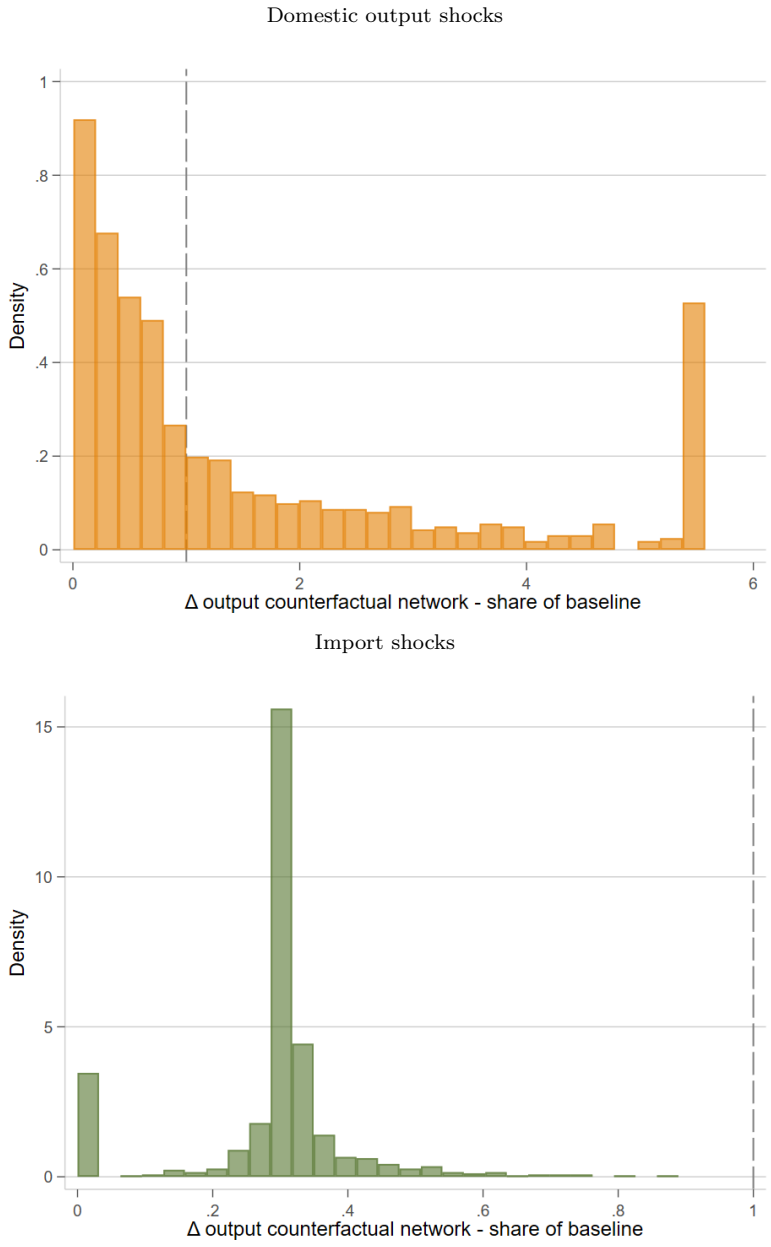
We regress the county-level change in outdegrees on various county characteristics including the formal sector share, the Gross County Product, and market access.

Table C3: Differences in simulated output reduction for counterfactual 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	-3.781***	-4.070***	-3.013***	0.089***	0.228***	0.148***
	(0.778)	(1.184)	(0.835)	(0.025)	(0.040)	(0.055)
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 vs rely only on the administrative data. The ratio is larger than 1 if we underestimate the impact of the shock and smaller than 1 if we overestimate it, if we do not account for informality. The above table shows the results from regressing this outcome at the sector-county level on the formal sector share measured at the sector-county level.

Figure C7: Δ output counterfactual network - share of baseline



We plot the ratio of the impact response to an adverse shock if we account for informal firms vs rely 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.