Spatial Inequality and Informality in Kenya's Firm Network *

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

June 2024

Preliminary

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 leverages transaction-level tax records to study spatial patterns of domestic firm-to-firm trade in Kenya and explores how these may be shaped by the presence of an unobserved informal sector. First, we document stylised facts about formal firms in this setting, revealing a high degree of spatial concentration in the production network, over and above the concentration of aggregate economic activity. Then, using data from the population census and national accounts, we show that informality is particularly prevalent in downstream economic activities and smaller regional markets. We structurally estimate a network formation model and find that accounting for informal firms in a counterfactual network increases the outdegree of firms in regions with the highest levels of informal activity and reduces the dispersion in outdegrees across counties. Further, we find that the higher the informality in a sector and region, the more we underestimate its vulnerability to domestic shocks and the more we overestimate its vulnerability to import shocks.

*We thank the Kenya Revenue Authority (KRA) for the outstanding collaboration. Romeo Ekirapa, Simon Mwangi, and Benard Sang provided excellent technical support and advice. We thank Elizabeth Gatwiri and Daniela Villacreces Villacis for excellent research assistance. We thank Andrea Bacilieri, Raphael Bradenbrink, Banu Demir, 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 CSAE research workshop, 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

Leveraging domestic markets and reducing inequalities within them are gaining traction as ways to promote economic development in low- and middle-income countries (Goldberg and Reed, 2023). The limited scope of export-led growth models has also led to an increased focus on domestic supply chains and market integration (Atkin and Donaldson, 2015; Bustos, Garber and Ponticelli, 2020; Grant and Startz, 2022). The structure of these supply chains can also influence economic development by affecting the dispersion of welfare gains from trade shocks, infrastructure investments, the propagation of adverse shocks, spillover effects of FDI, and the exposure of the network to systemic risk (e.g. Acemoglu, Carvalho, Ozdaglar and Tahbaz-Salehi (2012); Baqaee (2018); Liu (2019); Alfaro-Urena, Manelici and Vasquez (2022); Arkolakis, Huneeus and Miyauchi (2023); Castro-Vincenzi, Khanna, Morales and Pandalai-Nayar (2024)). However, studying domestic firm networks is empirically challenging due to a lack of granular and comprehensive data.

A rapidly growing literature has turned to using value-added tax (VAT) returns to track firm networks (Atkin and Khandelwal, 2020), thus far focusing largely on high- and upper-middle-income economies with relatively lower levels of informality.[1] By definition, transaction-level tax records do not capture details on informal firms and their role in inter-firm trade, which could be crucial for our understanding of how shocks pass through the entire economy and how welfare gains of government policies are dispersed across space. This paper asks how the presence of a sizeable informal sector can bias what we can learn from administrative records on formal firms and the degree to which this bias can affect our predictions about the remoteness and resilience of the economy. We study this question in the context of Kenya, combining transaction-level administrative records on over 76,000 formal firms and close to six million supplier-buyer relationships with granular data on the sectoral and spatial dispersion of informal sector activity from population census data and national accounts.

We ask the following questions: How can we characterize informal firms in the Kenyan economy in terms of their spatial and network characteristics? How does the presence of a sizeable informal sector shape the observed spatial inequality in a network of formal firms? How does taking informal firms into account alter predictions about the propagation of domestic and import shocks and the economic vulnerability of specific sectors and regions? The Kenyan context is particularly well-suited to answer these questions. First, the VAT paying sector only accounts for 34% of Kenya's GDP. Second, as East Africa's largest economy, Kenya boasts a substantial

[1] Notable exceptions are recent papers by Panigrahi (2022); Gadenne et al. (2022); Fujiy et al. (2022); Castro-Vincenzi et al. (2024) in India, Chacha et al. (2024) in Kenya, Spray (2019) in Uganda and Spray and Wolf (2018) in Rwanda.

domestic market with vast geographic and socio-economic cross-regional heterogeneity. Third, a series of granular data sets allow us to estimate economic activity both at the sectoral and regional level – features that can otherwise be difficult to quantify in many contexts as statistical bureaus tend to focus on national aggregates (Chacha, 2019).

We proceed as follows. First, we establish four stylized facts concerning the spatial patterns of domestic trade among formal firms in Kenya. Subsequently, we utilize census and national accounts data to document three additional stylised facts related to the informal sector. Based on these insights, we estimate a network formation model that accounts for heterogeneity in firm sizes, sectors, and locations to investigate whether taking into consideration informal firms alters the spatial inequality in firm-to-firm links documented in the tax records. We analyse the network predicted from the model and then use simulations of random domestic and import shocks to show how the pass-through of such shocks changes using a predicted network that accounts for informal firms.

With regard to our findings on firm-to-firm trade among formal sector firms: First, we find that trade among formal private sector firms is substantially more concentrated around Kenya's metropolitan areas than both population and aggregate economic activity. For example, the Pareto exponent for the spatial dispersion of trade flows within the network of formal firms is 57%-76% lower than the exponent for regional GDP.[2] Firms from the country's capital Nairobi and the port city Mombasa participate in as much as 88% of all firm-to-firm transactions as either buyers or suppliers. Second, these aggregate trade flows are a result of spatial inequality along the extensive margins of the firm network i.e. driven by firm locations and trade relationships rather than trade volumes. In fact, 90% of the variation in aggregate trade volumes across counties can be attributed to the extensive margin: the location of firms and a stark inequality in the number of firm-to-firm links across regions. The intensive margin, i.e. the number of transactions and average trade flows per shipment, plays a minor role. Third, upstream linkages (to suppliers) are more equally distributed across space than downstream linkages (to customers). Fourth, we find that linking patterns differ by firm size in that small firms source more locally.

Having established these stylised facts, we then ask whether the observed patterns are a result of our limited view due to the systematic selection of firms into the administrative data or whether it is reflective of the economy's underlying structure. We document three stylised facts about the informal sector. First, we show that the VAT-paying sector only accounts for 34% of Kenya's GDP implying that ignoring the remainder of the economy while studying the production network can impose substantial costs. Second, we show that informality is not randomly distributed but is located in specific sectors, geographies, and positions along the

[2] A lower Pareto exponent indicates a higher degree of inequality.

supply chain. For instance, informal firms are usually located downstream of large (formal) firms and informality correlates negatively with regional economic size and income. Third, we show that the spatial concentration in economic activity is largely a result of the concentration of formal sector activity. As a result, we expect that accounting for the informal sector can 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 import shocks.

To predict a counterfactual network that accounts for informal firms, we introduce and estimate a network formation model with heterogeneous node types by Bramoullé et al. (2012) to our setting. In our adaptation of the model, we classify firms into types based on their sector of operation, location, and size, since the stylised facts indicate substantial heterogeneity along these dimensions.[3] The network formation process is as follows. A new-born 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[4] and the remainder via preferential search i.e. it chooses a certain proportion of suppliers independent of its network environment, but the remainder from the pool of the suppliers of these suppliers.[5] The model provides predictions for the number of links between firms of different sector-location-size types. 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. 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.

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. We use the counterfactual network to answer the question of interest: how do spatial patterns of trade change when informal firms are accounted for? The key assumption is that informal firms are similar in terms of their linking probabilities to the smallest quartile of formal firms

[3] Conditional on the sector and county, we distinguish between large firms in the top three sales quartiles and small firms, firms in the bottom quartile.

[4] The network formation literature refers to this search pattern as "random" search (Jackson and Rogers, 2007; Bramoullé et al., 2012; Chaney, 2014).

[5] This search behaviour can be rationalised by underlying information frictions about the quality of potential suppliers (Chaney, 2014).

observed in our data.[6][7] The counterfactual results show that regions with the highest levels of informality and fewer ex-ante outlinks are predicted to be more spatially integrated if we account for unobserved informal firms. Hence, the spatial inquality in outlinks declines by 5% and the prominence of Nairobi also falls substantially. At the same time, we find that the predicted network is more partitioned in that it has more clusters with links among them rather than links across different clusters.[8] These predictions have implications on the predicted pass-through of domestic and trade shocks across space. By considering the formal network only, smaller regions with a higher degree of informality appear less exposed to shocks to network hubs like Nairobi.

Next, we simulate the pass-through of domestic and import output shocks to individual sector-regions using both the network observed in the administrative data and the counterfactual network that accounts for informal firms respectively. In the scenario with the counterfactual network, we find a larger adverse impact of those output shocks on sector-regions with a higher level of informality. Our results suggest that a 1 percentage point decrease in the formal sector share results in us underestimating the output reduction of a domestic shock by 4 percentage points if we do not account for informal firms. On the other hand, we find that the economy is less exposed to import shocks than we would predict if we ignored informal firms. This is because unlike domestic shocks, import shocks affect types with a higher import share as they are more exposed to world markets. These types are relatively less connected in the network than before, once we account for informality. Thus, after accounting for informality, we find that the economy is more resilient to trade shocks than one would otherwise predict. This complements existing literature that shows that the informal sector can act as a buffer against international shocks Dix-Carneiro et al. (2024).

Our paper is structured as follows. In Section 2, we discuss our contributions to the existing literature. In Section 3, we describe the content of the administrative data used to map the firm network. We document the spatial patterns of Kenya's formal firm network in Section 4. We discuss the role of the informal sector and map its spatial and sectoral composition in Section 5. Finally, we tie the two together in the Section 6 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 7. Section 9 concludes.

[6]This is because we keep constant the probability with which types interact with types of other sectors, sizes, and locations but update the probability of a type being born in the first place.

[7]A potential concern could be that informal firms may not link in the same way as formal firms in that they might link more locally. Incorporating firm size, location, and sector while defining the types accounts for such concerns. We also back this up with evidence on linking patterns of small firms that are in line with the notion that small and informal firms source more locally and more from intermediaries (e.g. due to internal economies of scale (Grant and Startz, 2022)).

[8]This, for example, could be because informal firms are more likely to link within the same county rather than across.

2 Contributions

Our paper contributes to the literature on firm networks, informality, and the dispersion of economic activity within countries.

Our first contribution is to describe the spatial mapping of the formal firm network in a context with a sizable informal sector. The empirical facts on Kenya's firm network presented in our paper build on and extend the seminal paper by Bernard et al. (2019), who were among the first to describe key features of domestic production networks. Our detailed mapping of the spatial network structure complements recent studies by Panigrahi (2022); Miyauchi (2023); Arkolakis et al. (2023), who draw on similar data from five Indian states, Japan, and Chile respectively. These papers focus specifically on modelling and micro-founding the endogenous network formation process. Our focus instead is a more granular mapping and quantification of the spatial patterns in firm networks with specific attention to the potential bias that arises due to informality. In doing so, we also contribute to the literature on urban primacy and disproportionate agglomeration of economic activity (as published in Jefferson (1989), Jefferson, 1939; Ades and Glaeser, 1995; Henderson, 2002; Soo, 2005). We show that the extent of urban primacy that one predicts using data from formal firm networks can be overstated due to the missing informal firms.

Second, despite the large size of the informal sector in various contexts, informality along supply chains remains understudied to date. We document stylised facts about the informal sector in Kenya and show how ignoring this sector can have substantial implications on the structure of the production network, predicted impact of shocks, and patterns of spatial concentration and remoteness predicted using the formal data. Exceptions to this include De Paula and Scheinkman (2010); Böhme and Thiele (2014); Zhou (2022) who use firm survey data and structural models to look at the implications of a VAT regime for informality along supply chains and Gadenne et al. (2022) who can observe firm-to-firm linkages between VAT-paying and non-VAT paying, but registered, firms. We complement these papers by exploring the relevance of informality for observed spatial supply chain structures. Methodologically our paper sits in between the two groups of papers in that we are also able to observe the formal firm network at a very granular level (Gadenne et al., 2022), but do not observe linking probabilities between formal and informal firms and therefore rely on a network formation model and complementary data sources to look at the sensitivity of spatial concentration in firm-to-firm trade to accounting for informal firms.

Relatedly, we also 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). Relying on cross-country regressions, this literature has documented that the relative size of the formal economy increases as income levels rise (Brandt, 2011; La Porta and Shleifer, 2014; Ulyssea, 2018). We show that this pattern extends to Kenya's domestic economy: 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 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).

Finally, we contribute to the large literature on estimating network statistics and reconstructing networks in the presence of missing data. For instance, Chandrasekhar (2016) provide 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 nodes are missing in a non-random manner in our case and 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 and updating the probability that a node exists accordingly while keeping the linking probabilities constant. The key assumption here is that conditional on an informal sector firm being of the same type (eg: the same size, sector, and location) they are going to form links similar to the observed firm in the formal sector.

3 Administrative data

3.1 Description of data sources

Our analysis draws on micro data from value-added and pay-as-you-earn tax returns. We further 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.[9] We deflate all variables denoted in monetary terms using the monthly Consumer Price Index (CPI) published by the

[9]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.

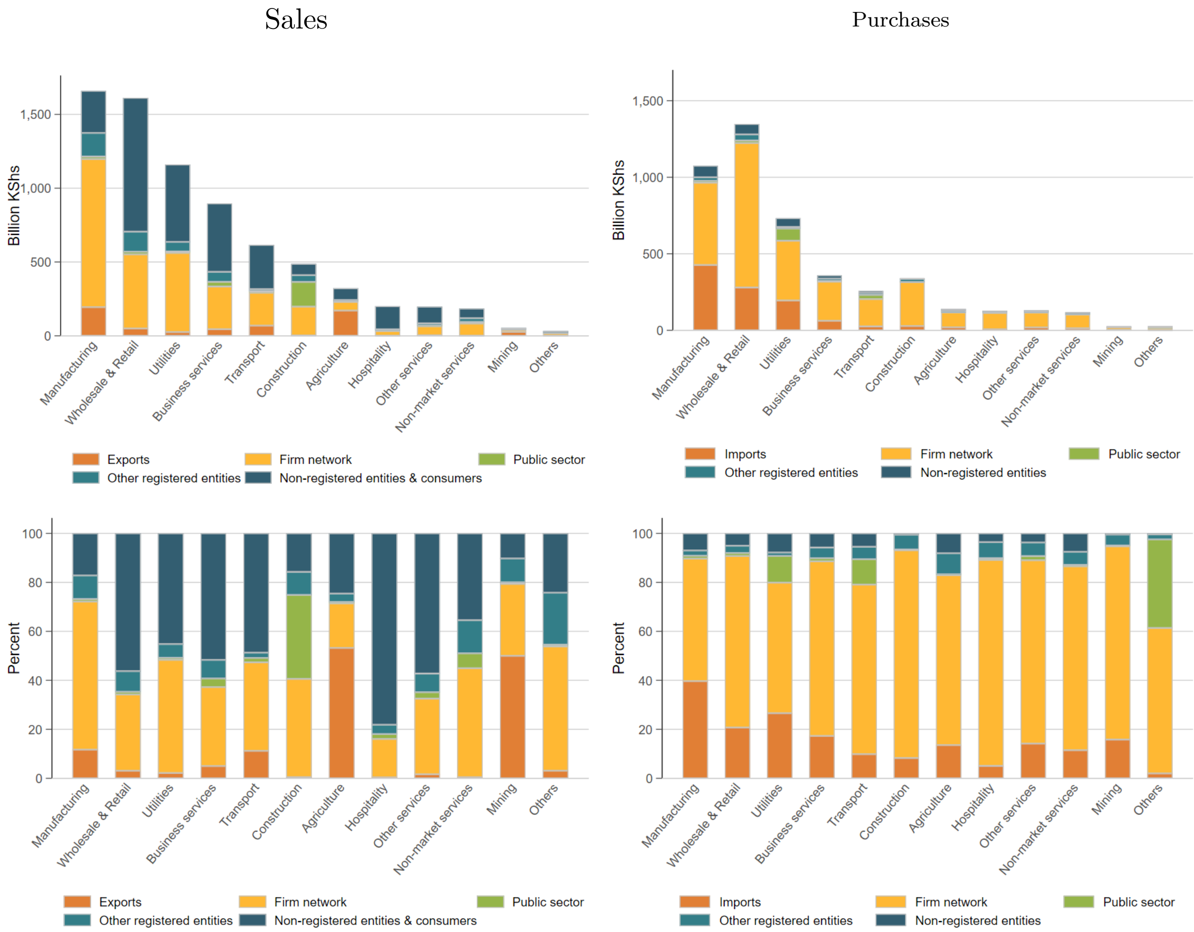
Kenya National Bureau of Statistics (KNBS).

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 exclude firms operating in the financial sector (most of which are banks and insurance companies). 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. We apply the VAT threshold to exclude firms that registered for VAT to bid for tender but were never operational.

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 account for almost half of the sales we are able to track in the tax records.

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.

Firms below the threshold, those offering financial and education services, and to a large extent, firms dealing in agricultural goods and pharmaceuticals, are exempt from VAT. They are not required to register for VAT and do not submit a monthly return. Sales to firms below the threshold and exempt firms are thus both captured in the category of sales to non-registered parties. The same category exists for purchases; however, it only allows for the reporting of purchases of VAT-exempt goods and hence excludes a lot of potential inputs from the non-registered sector. While this leads to a bias in what type of economic activity is captured in the data, we use firm survey data collected by KNBS (2016) in Section 5 to show that the bias is potentially not as severe given small firms are more likely to purchase from larger VAT-registered firms rather than the other way round.

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

We begin by examining the geography of Kenya's firm network of formal firms. We show three stylized facts: First, firm-to-firm trade flows within the network are highly spatially concentrated around Kenya's metropolitan areas. Second, 90% of this geographic concentration in economic activity is driven by the extensive margin of trade: firm location and the number of firm-to-firm links. Third, upstream linkages are more equally distributed than downstream linkages. Fourth, linking patterns differ by firm size. Small firms are more likely to source locally and buy from intermediaries.

These stylized facts will later motivate our model set-up. Fact 2, firm location and relationships explaining the bulk of variation in trade volumes across locations, motivates our choice of a network model that is centred around the extensive margin of trade, where firm location and the number of firm-to-firm relationships are the two key ingredients. Due to fact 3, downstream linking patterns varying more across space, we will focus on modelling the outdegree of firms and target the outdegree distribution of the real-world network in our structural estimation. Lastly, in line with fact 4, we embed firm size as one of the three dimensions across which firm linking patterns can vary in our model, the other two being sector and geography.

4.1 Urban primacy in Kenya's firm network

Kenya's firm network is strongly concentrated around its metropolitan areas Nairobi and Mombasa.[10] 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[11] and the region's communication network (Memon, 1976; Obudho, 1997).[12] In 1960, 49% of the wholesale sector's turnover was generated by Nairobi-based firms, who in turn also employed 46% of this sector's workforce (MoF, 1963, as cited in Memon (1976)).[13] Today, as much as 68% of the sales volume within the network of formal firms is generated by Nairobi-headquartered firms. Notably today, as in the 1960s, the city's role in the firm (or trade) network is disproportionate relative to its population and even aggregate GDP.[14] In 2019 (1960), as little as 9% (3%) of Kenya's population

[10] 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.

[11] The railroad followed existing caravan routes. Mombasa and Nairobi then gradually replaced Zanzibar as the major trading hub of the region (Memon, 1976).

[12] As predicted by Memon (1976), Nairobi's disproportionate growth as an economic and urban centre would start to accelerate and the previously bi-polar system would effectively become a Nairobi-centric one.

[13] Mombasa accounted for 27% of employment and 35% of turnover in the wholesale sector.

[14] We map firm headquarter locations and population density in Appendix Figure A1.

lives in Nairobi County and the city contributes 37% of Kenya's GDP outside the agricultural sector (see Table 7).

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.

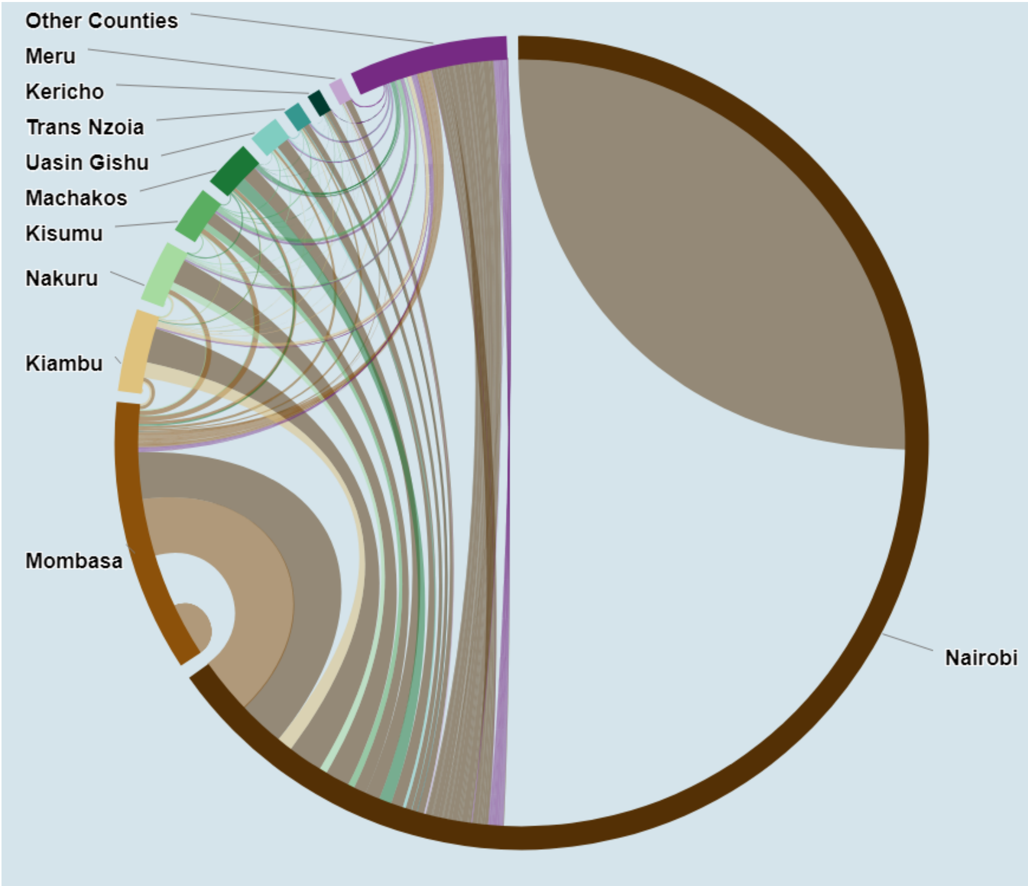
Table 1: Geographic concentration of economic activity in Kenya

	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	29	5	0.98	0.07
GDP w/o agriculture	37	6	0.95	0.06
GDP w/o non-market services	24	5	0.91	0.08
No. VAT firms	64	9	0.61	0.04
Employment in VAT firms	62	8	0.35	0.03
Value added of VAT firms	72	10	0.37	0.03
Network sales	68	13	0.24	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 of the pie is proportional to the respective county's sales within the network and the colouring of the trade flows aligns with the county of origin. 88% of the 21 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 45% of the total trade volume. The graph further reveals that trade flows out of Nairobi and Mombasa are larger than inflows into the metropolitan areas.

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 the ten counties with the largest aggregate amount of transactions within the domestic production network. We bundle the trade flows for the remaining 37 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. The α for each indicator is obtained via rank-size regressions (Gabaix and Ioannides, 2004).[15] A lower value indicates a flatter slope and hence more

[15] I.e. a county-level regression of each county's rank (log) on the respective measure x (log): $\log \text{rank} = \log A + \alpha \log x$.

inequality across counties. The Pareto exponents for both total county-level GDP and aggregate income generated outside the agricultural sector is close to unity - in line with Zipf's Law, which stipulates a power law distribution with an exponent of approximately one (Gabaix, 2009; Soo, 2005). At the same time, Kenya's population is more evenly distributed across counties than economic activity. Turning to the firm network, we find αs that are substantially lower than one, indicating a high degree of spatial inequality. An exponent of 0.61 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 (0.24), 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.

To quantify which margins of aggregate trade flows drive this spatial concentration, we decompose them into sub-components as a next step.

4.2 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 the granular 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:[16]

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

Table 2 summarises the share of the variance attributed to each term in both upstream (purchases) and downstream (sales) trade flows.[17] The number of firms operating in each county alone accounts for 67% of the variance in purchases across counties.[18] The number of relationships with suppliers of the county accounts for yet another 22%, leaving a little over 10% of the variance to be picked up by the intensive margins for trade, i.e. the number of transactions between firm pairs and the average transaction volume. Turning to downstream trade flows, i.e. the decomposition of the variance in sales across (sub-)counties, the location of firms plays a slightly less important role. Instead the number of firm-to-firm relationships now accounts

[16]The same is true for purchases.

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

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

for one third of the variance in network sales. The variance decomposition is a useful exercise to track the respective margins of trade. It, however, falls short of allowing us to 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.

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

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

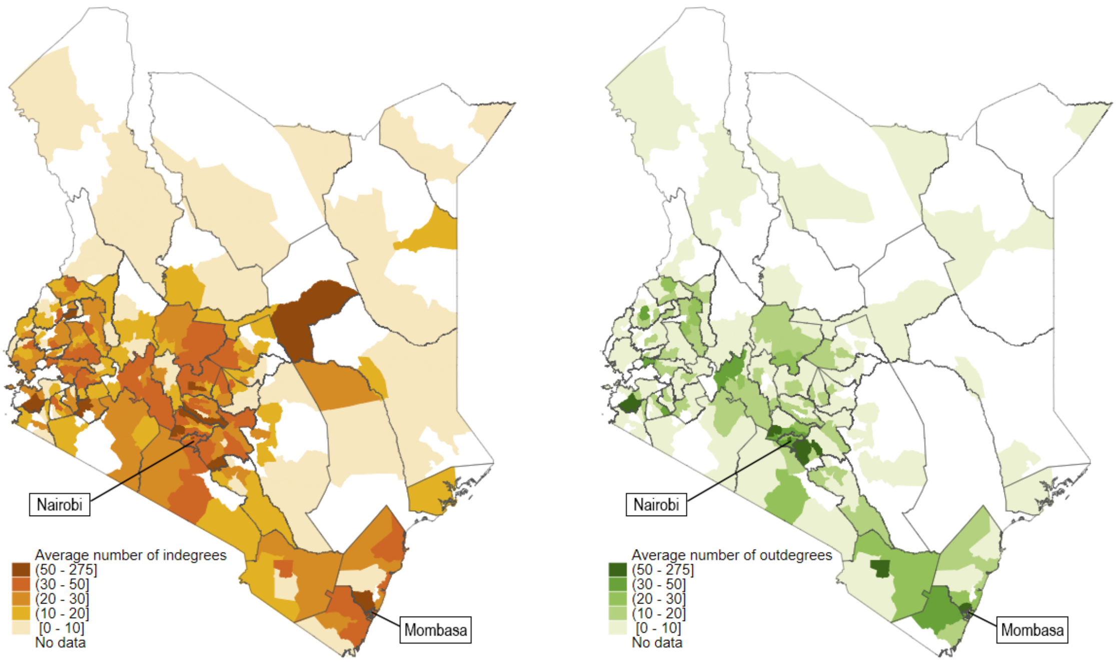
4.3 Upstream linkages are more equally distributed than downstream linkages

Our earlier 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. 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). Recent research on the origins of firm heterogeneity has highlighted the importance of the number of firm-to-firm relationships with buyers, the extensive margin, in determining firm size (Bernard et al., 2022). We show that this pattern replicates across space. Nairobi-based firms on average have 32 buyers with those based in the central business district reaching an average of 140 formal sector buyers. At the same time, firms in the median county (sub-county) have an average of 13 (5) buyers.[19] Turning to the average indegree (suppliers), Nairobi and other larger Kenya cities and towns no longer stand out as much. The left map in Figure 3 shows a much more equal distribution of the average indegree across space. Firms in the median county (sub-county) have 22 (20) suppliers,

[19]The contrast is even starker of course if we also exclude Mombasa and Machakos County from the average. Machakos is coloured in dark green in Figure 3 as it serves as an industrial hub with many large manufacturing firms.

Figure 3: Average in- and outdegrees across space

Average indegree | Average outdegree



The above map plots the average in- and outdegree of firms each sub-county. County borders are outlined in grey.

Nairobi-based firms have 30 suppliers on average. This pattern also aligns with the higher level of spatial inequality found in network sales relative to the location of firms in Table 1.[20]

4.4 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 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. Indeed in the

[20] 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.

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	✓	✓	✓	✓	✓	✓	✓	✓

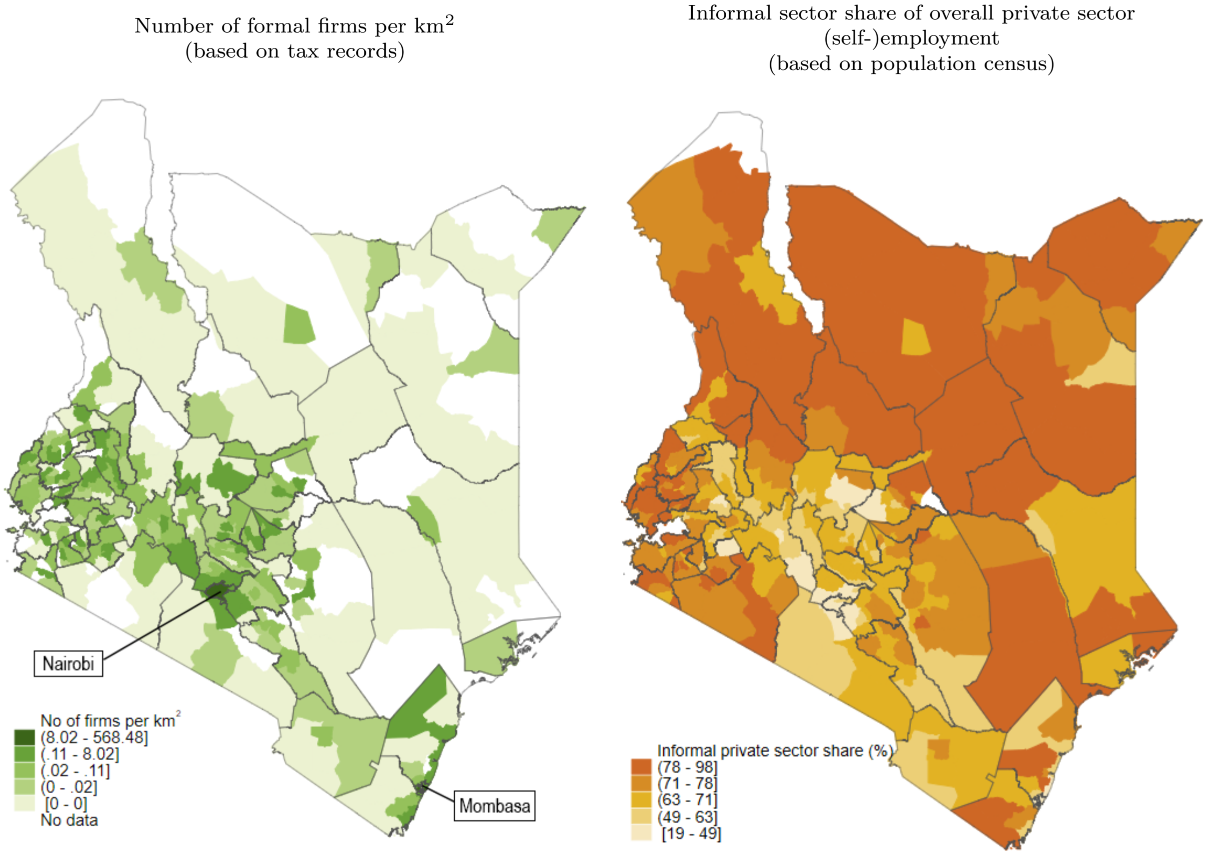
We aggregate linking probabilities with suppliers of specific characteristics by buyer type (sector-county-size) and regress the sum of these probabilities on whether or not the buyer type is a small buyer type. The column titles list the characteristics of the suppliers.

5 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 A1). However, while population is concentrated across space, it is not nearly as concentrated as formal sector trade flows (see Table 1). In Figure 4, we compare the number of formal firms per km^2 to the share of people being employed in the informal sector. 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.

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



The left map shows the density of firm headquarter locations at the sub-county level in 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 sub-county 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.[21] Hence, one might wonder whether links of firms in sectors and regions with high informality are captured in a complete manner. This, as we will show, can have substantial implications for the spatial pass-through of domestic and trade shocks.

[21] 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	
	buyers	suppliers	buyers	suppliers	buyers	suppliers	buyers	suppliers
Formal sector share (county, %)	0.120***	0.114***	0.046***	0.036***	0.014*	0.028***	0.050***	0.046***
	(0.028)	(0.022)	(0.012)	(0.006)	(0.007)	(0.006)	(0.013)	(0.009)
Population	1.279***	1.109***	0.355***	0.185**	-0.110	0.007	0.428***	0.316***
	(0.221)	(0.190)	(0.131)	(0.081)	(0.126)	(0.086)	(0.144)	(0.100)
Travel time to Nairobi	-0.073	0.009	-0.057	0.039	0.057	0.019	-0.042	0.052
	(0.369)	(0.262)	(0.181)	(0.094)	(0.109)	(0.077)	(0.182)	(0.117)
Travel time to Mombasa	-0.245	-0.146	-0.208*	-0.141	-0.194	-0.190**	-0.248*	-0.126
	(0.231)	(0.171)	(0.119)	(0.090)	(0.127)	(0.089)	(0.124)	(0.102)
No. observations	493	517	493	517	419	516	493	517
R2	0.576	0.642	0.451	0.377	0.257	0.336	0.455	0.367
Sector FE	✓	✓	✓	✓	✓	✓	✓	✓

In the forthcoming 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 establish a definition of informality within our context. Then, we detail and examine the additional data sources employed to study informality. Finally, armed with a definition and relevant data, we present three essential stylised facts that will inform how we account for unobserved informal firms in our subsequent section that presents the model. The purpose of the model-based counterfactual is to study the relevance of unobserved firms for the structure of Kenya's firm network and the spatial inequality in firm-to-firm links.

5.1 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[22] 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).*

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

Some customers will purchase with receipt and output VAT on it. Some customers 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.[23]

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.[24] 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.[25]

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. The Kenya Revenue Authority requires firms to record transactions of any size, conditional on both parties being VAT-registered.[26] Omission of transactions between two formal firms or the misreporting of trade volumes on the other hand remains a concern for us.[27] We are able to recover some of the omitted transactions

[23] 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.

[24] 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.

[25] Specific to Kenya's legal context, a lot of firm we miss out 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.

[26] The application of transaction thresholds varies substantially across contexts. Like the Kenyan Revenue Authority, the Dominican Republic and Uganda do not have a volume-dependent threshold either (Cardoza et al., 2023; Almunia et al., 2022). In Belgium, Costa Rica, and Turkey 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).

[27] 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

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.

5.2 Measuring informality

We now turn to the measurement of the informal sector outside the VAT system. Table B1 provides an overview of the data sources we draw on, while Table 6 summarises how we compute the various measures of 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. As discussed in the previous section, we will focus on a definition of informality that refers to economic activity by non-VAT paying firms. We use the administrative data as a measure for the size of the formal sector and a corresponding figure for the aggregate economy including both formal and the informal economic activity from KNBS. However, we will also rely on measures of informality that only rely on micro data and national accounts data published by KNBS. Why do we need this second group of measures? The measures based on the administrative data will tell us what part of the economy we miss out on by studying the Kenyan economy through the lens of VAT data. The comparison with measures that only rely on third party data then shapes our interpretation of those gaps between the administrative records and overall economic activity. Are these gaps between the VAT data and aggregate economic activity, for example, reflective of economy wide dynamics between the formal and informal sector?

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

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 interest.

added rely on estimates of the universe of businesses in KNBS (2016)[28] 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 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.[29]

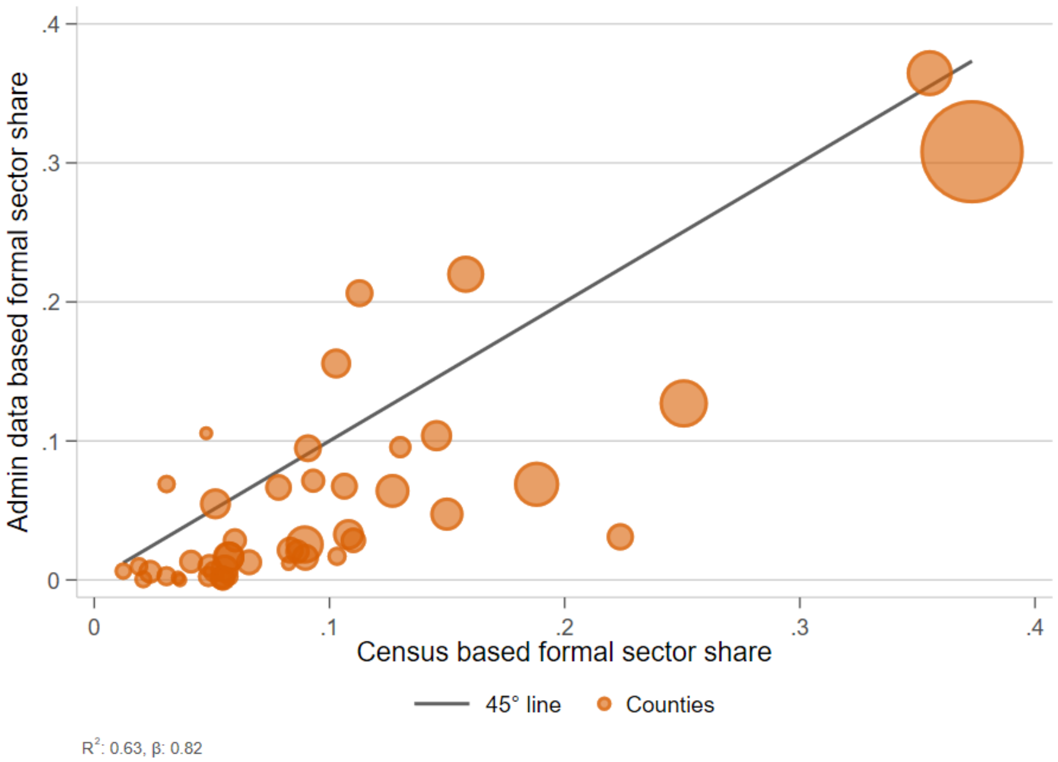
We exclude any employment or firms in the agricultural sector or non-market services when measuring informality at the regional level.[30] In both cases, the tax records cover a small and very specific sub-population of firms and employees only. In the case of agricultural firms, the administrative data cover large-scale commercial agriculture only. The vast majority of these firms are export oriented (Chacha et al., 2022). For non-market services, we 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 enjoys VAT exemptions and hence does not appear in the network.

[28] KNBS (2016) obtain information on the number of licensed businesses from county governments and estimate the number of unlicensed businesses based on household survey data.

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

[30] 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.

Figure 5: 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.

Our preferred measure of informality considers formal sector employment as per the 2019 population census. It correlates strongly ($\rho = 0.83$, Table B2[31]) with its counterpart based on the administrative records (also see Figure 5). In the final section of the paper, we will use the employment- and population census based measure of overall economic activity by sector and region to explore counterfactual network patterns that consider the presence of informal firms.

5.3 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.

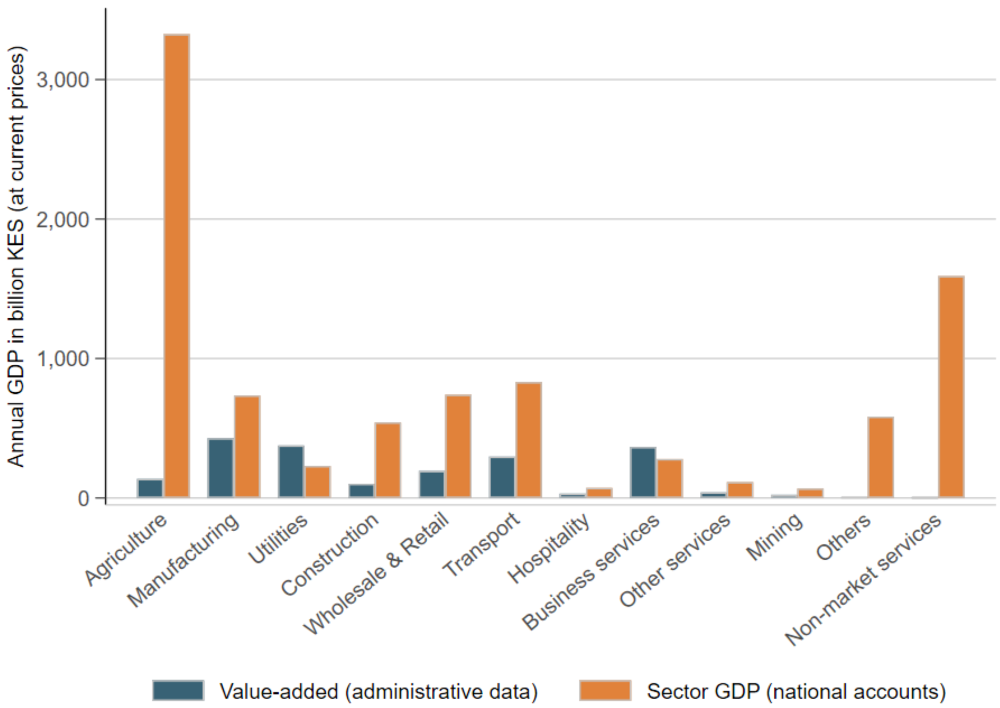
[31] Table B2 summarises the correlation coefficients of the alternative informality measures.

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.

5.3.1 Fact 1: The VAT-paying sector accounts for 34% of Kenya's GDP

The gross value added generated by VAT-paying firms on average corresponds to 34% of Kenya's annual GDP for the period 2015-2020 (see Appendix Table B3).[32] 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). However, not all of this gap can be attributed to informality. The differences arise for two reasons: The first reason is differences in the tax code, in particular the treatment of financial services, non-market services, and agriculture. The second reason is informality. We discuss the influence of sectors that are not well-represented in the data in explaining the gap in Appendix B.3.

Figure 6: 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.

The VAT sector accounts for 64% of residual economic activity, if we exclude sectors that are

[32]The value added recorded in the VAT data fluctuates between 40% (2016) and 24% (2020) of Kenya's GDP. Figure B1 plots the quarterly time series for both GDP and value added. We discuss related time trends in Appendix Section B.2.

to a large extent exempt from VAT (non-market services, agriculture) or have special reporting rules applied to them (financial services).[33] This implies an overall informal sector share of 36%. 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 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 data, while another part stems 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 36% of the Kenyan economy by looking at the formal sector only begs the question - where to find those informal firms?

5.3.2 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 6, shows that manufacturing and professional services are best represented in the administrative data. This pattern aligns with the fact that both sectors are well-connected in the firm network and mostly buy from and sell to other formal firms.[34]

The gap between value added and GDP is an aggregate of all margins of informality discussed in Section 5.1. We are unable to quantify the extent to which each margin contributes to aggregate informality. However, we get an idea of how the extensive margin of informality varies across sectors by comparing 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. In both instances, firms in the wholesale & retail sector show the highest incident of informality.

The first type of extensive margin informality arises due to non-reporting firms with revenues above the VAT threshold. Figure 7 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. Manufacturing,

[33] 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.

[34] For professional services, as much as 80% of their transaction volume is attributed to trade with other businesses (see Figure 1).

ICT, and professional services show the least deviation between the two data sources (see Table B4 for the same comparison in tabular form).[35] Wholesale & retail stands out as the sector with the largest number of non-compliant firms.[36] This is unsurprising given VAT self-enforcement is weakest for consumer facing firms (Naritomi, 2019), a theme we explore further below.

The vast majority of firms in Kenya, however, is simply too small to be VAT-registered. In 2016, KNBS (2016) estimated 7.4 million businesses to operate across Kenya. This includes licensed businesses registered with county governments, as well as unlicensed establishments, mostly street vendors and other forms of micro enterprises. Only 1 in 5 of the businesses is licensed. Of the licensed businesses the CoE and the VAT data capture 9% and 2.5% respectively. The bottom graph of Figure 7 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. As mentioned above, the definition of what constitutes a business and the inclusion of public sector establishments varies by data set. The comparison is thus primarily useful for a big-picture overview.

Turning to the intensive margin of informality, the prevalence and possibility of informal transactions involving registered firms becomes evident when considering the construction sector. As little as 20% of the construction sector's contribution to GDP is reflected in the VAT records. At the same time, the majority of construction sector firms is registered for VAT (see Figure 7). Therefore, most of the observed value added gap is a result of informality on the intensive margin, i.e. mis-reporting in VAT returns.

The VAT-sector share correlates positively with regional economic size and income.

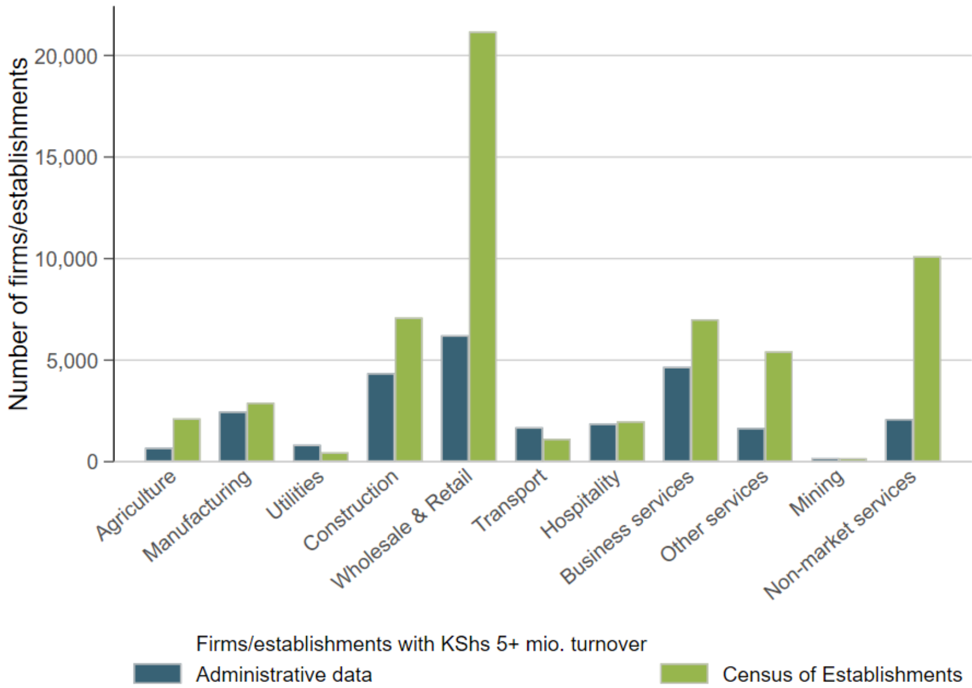
Moving on from sectors, where do we find informal firms in space? Here the answer is: in smaller markets. A county's overall economic size and income level (Gross County Product per capita) correlates strongly with its formal sector share. Economic size explains between 35% and 52% of the variation in 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

[35] The number of firms in the administrative records can be higher because the CoE tends to under-count firms without a highly visible establishment.

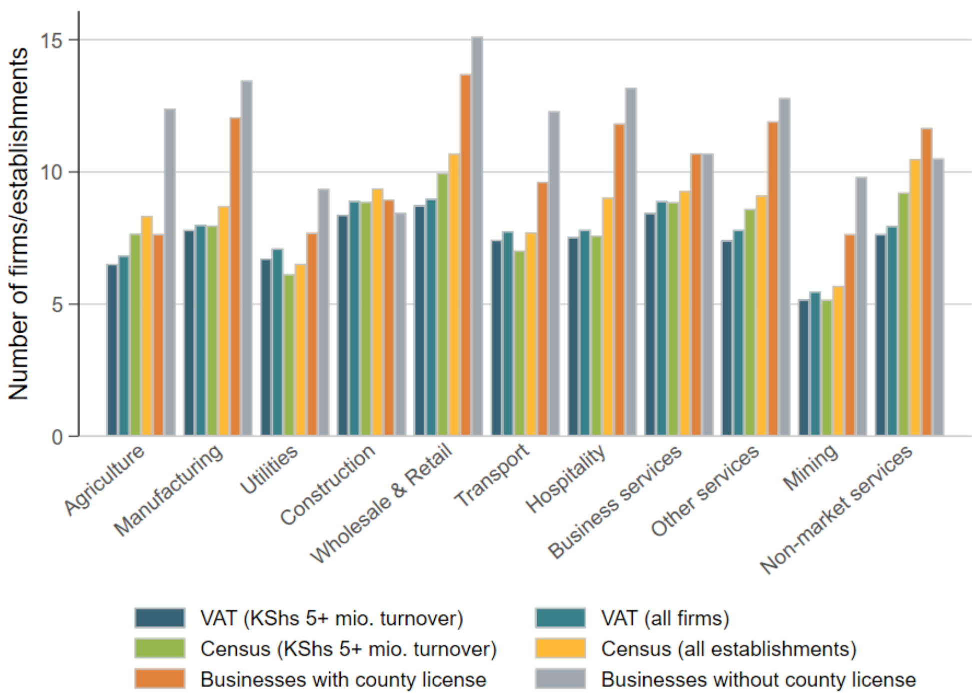
[36] In line with the discussion above, the gap for education & health is mainly a result of VAT exemptions that apply to most firms in this sector. The CoE, for example, includes public sector establishments such as public schools that offer educational services.

Figure 7: 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)

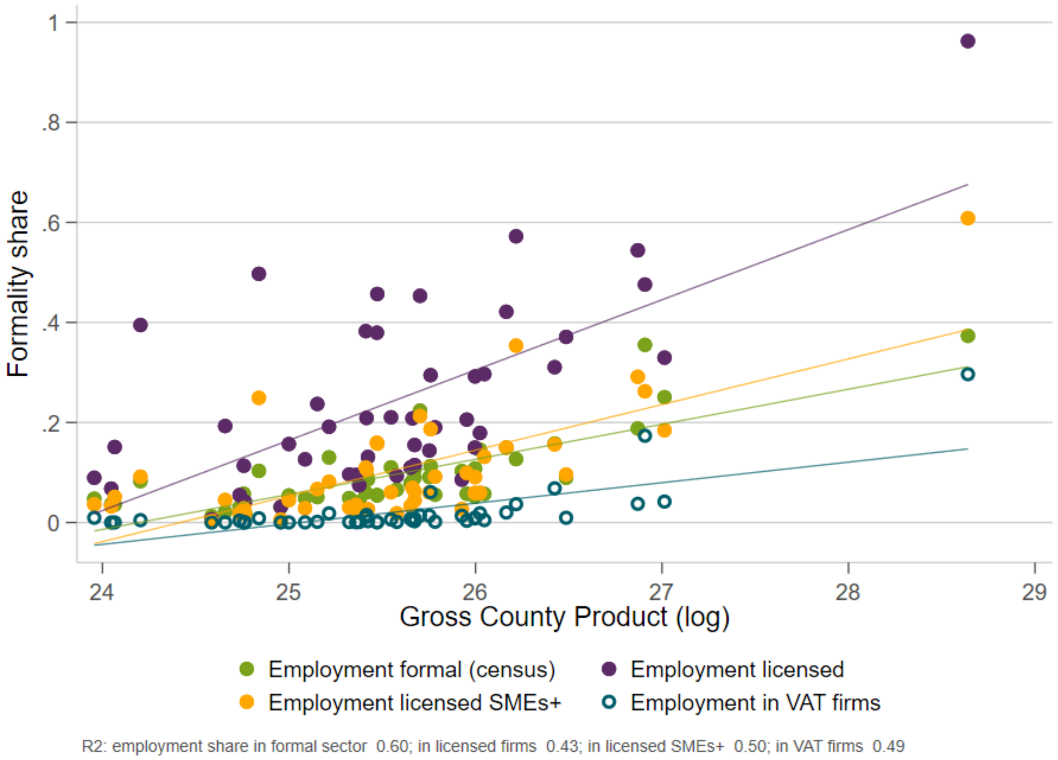


The top graph compares the number of firms covered in the administrative data and had a annual revenue of over KShs 5 million in 2016 to the number of firms with annual revenues above KShs 5 million in the 2016 Census of Establishments (CoE) (KNBS, 2017). The bottom graph compares the two groups of firms 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 KNBS in KNBS (2016).

informality of output declines less steeply with income levels than the share of informal workers (Kose et al., 2019).

To test whether the positive correlation of market size and the formal sector share is an artefact of the administrative records, we also correlate three other employment-based formality measures with the Gross County Product in Figure 8. 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 suggests that the variation in county-level informality that is explained by the economic size of the county is very similar for different measures of informality.

Figure 8: Share of formal sector employment and regional market size



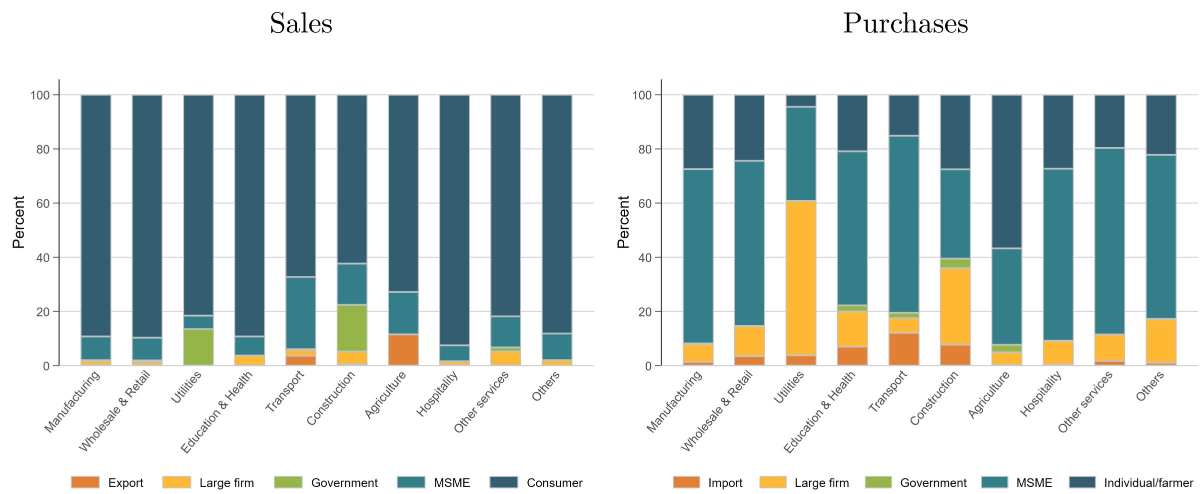
The first measures uses the formal sector employment share according to the 2019 population census, the second measure considers the number of employees in licensed businesses, the third uses the same measure but disregards mirco-enterprises, and the forth measure considers employment in the tax records.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

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 (Böhme and Thiele, 2014).[37]

While the high number of non-VAT paying firms in the wholesale & retail sector already is a good indicator for the relative downstream positioning of informal firms along the supply chain, we show that this pattern also emerges from survey data on trading partners of micro, small and mid-sized enterprises (MSMEs) by KNBS (2016). The survey asks firms for the type of entity that best describes their main source of inputs as well as their main customer. Only 2.3% of all MSMEs sell to large firms, while 14.5% purchase inputs from them.[38] Figure 9 shows that the pattern holds across sectors.[39] Our results confirm findings by Böhme and Thiele (2014); Zhou (2022) who document similar linking patterns for 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.

Figure 9: 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 two findings that the extensive margin of informality is particularly prevalent in the whole-

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

[38] KNBS (2016) defines non-MSMEs/large firms as entities with more than 99 employees. Other survey response options for inputs are: MSMEs, farmer, direct import, individual supplier, government. The corresponding options for customer types are: MSMEs, direct exports, individual consumer, government.

[39] The survey responses can be interpreted as a lower bound on the interaction between the VAT-registered and non-registered sector. The main trading partner of MSMEs are other MSMEs and the relevant survey question does not distinguish between the three types of enterprises - micro, small, and medium - summarised under MSME. Kenya's Micro and Small Enterprises Act No.55 of 2012 defines small enterprises as firms with up to 50 employees and up to KShs five million annual turnover (KNBS, 2016). Medium sized enterprises thus cross the VAT threshold of five million and any link with them would count as a link with a formal firm.

sale & retail sector, as well as informal firms being more likely to purchase from larger firms rather than vice versa, are well in line 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 a larger share of economic activity to take place outside the VAT system in more downstream sectors.

Consistent with this, we find that firms outside the metropolitan areas sell a larger proportion of their sales to non-registered entities, on average (Figure A2). This result could well be an outcome of differences in sales channels of firms in the metropolitan areas compared to those outside. At the same time, the steep decline in average outdegrees outside the metropolitan areas could be driven by informality along supply chains as sales to consumers are lumped together with sales to non-VAT paying (potentially informal) firms in the tax returns.

5.3.3 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 4 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	29	5	0.98	0.07
GDP w/o agriculture	37	6	0.95	0.06
GDP w/o non-market services	24	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	0.61	0.04
Employment in VAT firms	62	8	0.35	0.03
Value added of VAT firms	72	10	0.37	0.03
Network sales	68	13	0.24	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 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.

6 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 (a) predict the formal firm network as observed in the data and (b) estimate a counterfactual network that accounts for informal firms. Bramoullé et al. (2012) is particularly well-suited for our purposes for three reasons: First, 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.

Second, the model incorporates a flexible network formation process such that the emergent degree distribution can follow a power law. 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 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). Using a model that allows for preferential attachment to overcome information frictions also links back to theories on the emergence of urban centres as "communication nodes and loci of information exchange and accumulation" (Memon, 1976).

Finally, we opt for a tractable framework that allows us to estimate the spatial dispersion in firm-to-firm links in the steady state after a dynamic process of network formation. We abstract away from additional complexities like endogenous firm entry and exit or the decision of a firm to formalise. The underlying dynamic network formation process gives rise to the widely documented extreme heteroegeneity in outdegrees across firms (Bernard et al., 2019; Panigrahi, 2022; Bernard et al., 2022; Bernard and Zi, 2022; Demir et al., 2023; Arkolakis et al., 2023). We first present the dynamic network formation model proposed by Bramoullé et al. (2012) below and then discuss how we estimate it.

6.1 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 quartile 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 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 and regions it wants to link with. Having chosen the sector-region 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 further consistent with the third stylized fact from Section 4, 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" in that 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. This aspect of the model captures the core features of the context that we are interested in as 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 $p(\theta)$.

At the aggregate level, we are interested in the outdegree of each sector-county 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 following matrix 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.[40] 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.

6.2 Estimation strategy

Given the granular data on the empirical firm-to-firm network, we are able to obtain the majority of the model parameters from the data. 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. As a result, we only need to estimate the parameter r, i.e. the fraction of input links a firm obtains via undirected search independent of the network environment. 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.[41]

First, we classify firms into types defined by unique sector-location-size combinations. Sectors refer to aggregate 2-digit sectors, 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.[42] For example, all firms

[40] Note that we diverge from Bramoullé et al. (2012) here. In their setting newly entered nodes (scientific papers) form outlinks (citations) with the implication that everybody's inlinks evolve over time.

[41] 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.

[42] 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.

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 type θ's inlinks that it forms with type θ'.[43] 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. 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.

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 $\pi_t = \sum^t_{t_0} (\mathbf{p} \cdot \pi^t_{t0}{}')'$ where $\mathbf{p}$ 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 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

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

firms with all firms observed in the data. As a result, we compute $\pi_t = \sum_{t_0=1:100:56822}(\mathbf{p} \cdot \pi_{t0}^{t}{}')'$. 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.[44] In particular, r^* is defined as follows:

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

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 10).

6.3 Estimation results

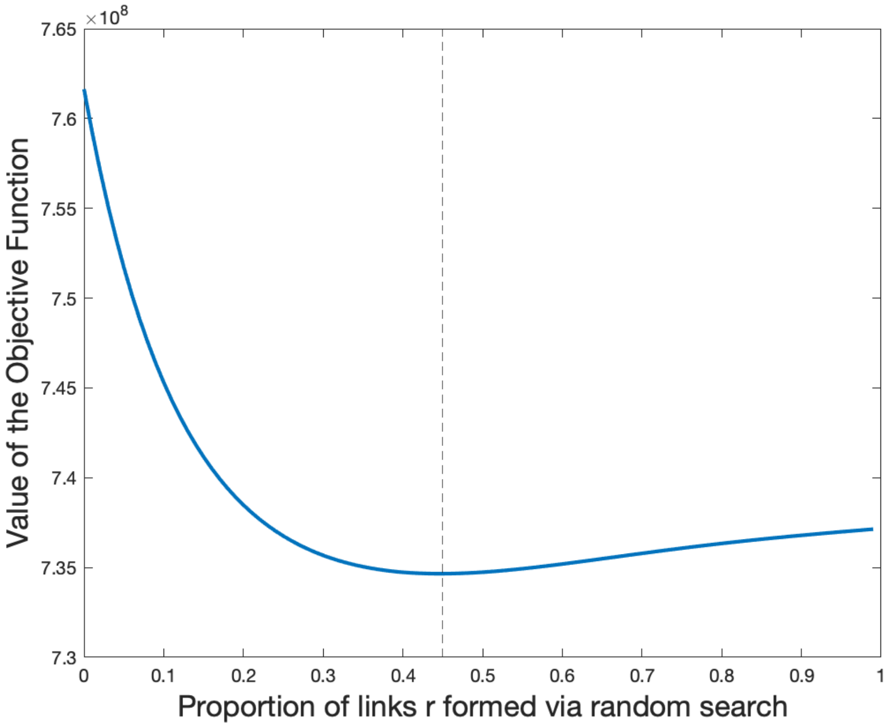
Our estimation strategy yields a result of $r^* = 0.45$. It suggests that a newly entered firm chooses 45% of its m suppliers randomly, and the remaining 55% among the suppliers of its existing suppliers. A network with 55% of all links being formed via preferential attachment suggests a prominent role for information frictions. It aligns with previous research documenting the importance of relational contracts in Kenya and neighbouring economies (Fafchamps, 2003). It is further backed by our own finding in Section 4 that social connectedness explains substantial variation in cross-regional trade flows, over and above distance. In the empirical section we found the elasticity of domestic trade flows with respect to social connectedness to be more than twice as high than for international trade flows (Bailey et al., 2021). In a variant of this model, Chaney (2014) estimates $r = 0.6$ for French exporters forming links with trade partners abroad, which

[44] In doing so, we assign greater weight to more common sector-regions types whose probabilities often tend to more stable over time.

also suggests a substantial, but not quite as prominent role for information asymmetries.[45]

In order to validate the estimated value of r, we plot the objective function in Figure 10 and show that it reaches a global minimum as $r = r^* = 0.45$.

Figure 10: Objective function for various values r



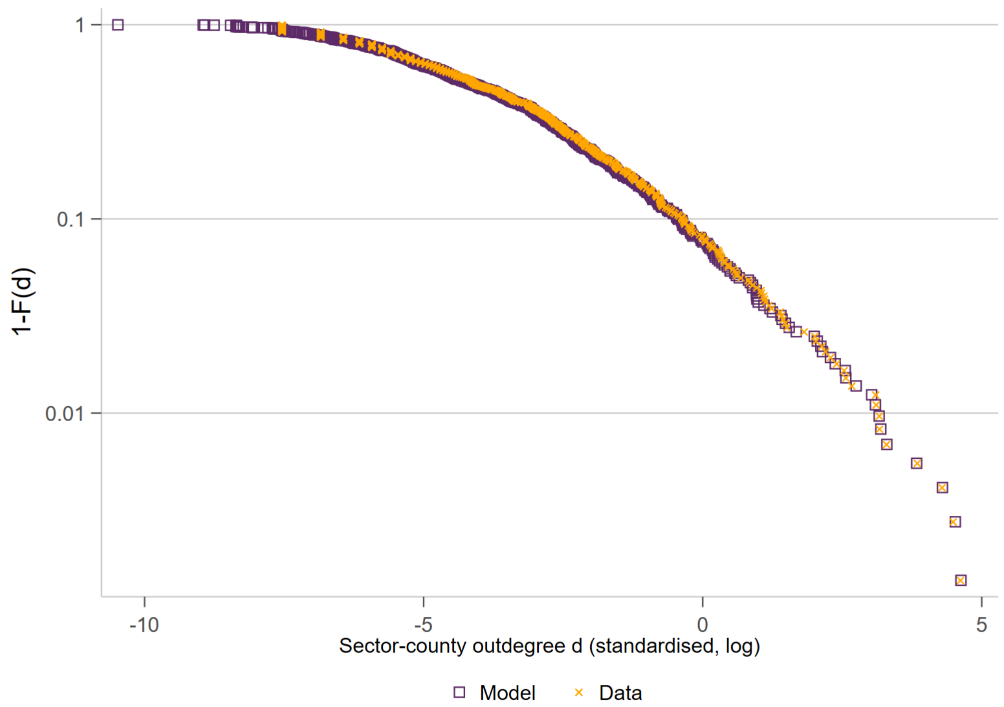
This figure plots the sum of the squared difference between each element of the model predicted interaction matrix π and the matrix π directly observed in the data, for various values of the parameter $r \in [0, 1]$. The figure shows that $r^* = 0.45$ obtained via simulated annealing minimises the objective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

Next, we plot the degree distribution (i.e. total number of outlinks) of each sector-county type as observed in the data and as predicted by the model. 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. Figure 11 shows that the key properties of the outdegree distribution are replicated by the model's predictions. 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 Pareto exponent for both degree distributions, we obtain an α of 0.3599 from the model and 0.3628 from the data. The standard errors are 0.0069 in both cases and we can hence not reject the null hypothesis

[45]From a welfare perspective, Chaney (2014) shows that in the case of highly substitutable goods, increasing r boosts aggregate welfare as it ensures more equal access to a variety of goods across buyers. However, aggregate welfare gains from overcoming preferential attachment dynamics are more ambiguous if goods are less substitutable. Boosting the average indegree m, i.e. network density, on the other hand is always beneficial.

that both coefficients are equal.

Figure 11: 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.

7 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 5 to update our estimates for all $p(\theta) \in \Theta$. Specifically, we attribute all informal economic activity to the types with small firms.[46] Knowing r^* and our updated $p(\theta)$s, we can then once again predict the matrix π, keeping everything else constant.

[46]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. We rely on formal sector wage-employment to compute the entry probabilities of firms in the bins with larger firms. The remaining three categories pin down the probabilities for any type related to small formal and informal firms.

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 an endogenous relationship between $p(\theta)$ and $p(\theta, \theta')$. We further assume that conditional on their sector of operation and their geography, informal firms have similar preferences about which other sectors and geographies they source from. We showed that small formal firms, for example, are more likely to purchase from wholesalers - in line with multi-intermediary supply chains.

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

To incorporate informal firms into the network, we need to update the firm-type probabilities $p(\theta)$ for each sector-county-size cell. We will refer to the updated, alternative versions of $p(\theta)$ that incorporates the sector-county characteristics of entire firm size distribution as $p(\theta_a)$. Given informality correlates strongly with firm size, we allocate all informal firms to the sector-county-size cell of small firms. 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 the private sector in each sector-county-size cell to compute $p(\theta_a)$:

$$p(\theta_{a,large}) = \frac{\text{No. formal employees who work in sector s in county c}}{\sum_c^{47}\sum_s^{12} \text{No. people who work in sector s in county c}}$$

and

$$p(\theta_{a,informal}) = \frac{\text{No. formal and informal self-employed + informally employed people}}{\sum_c^{47}\sum_s^{12} \text{No. people who work in sector s in county c}}$$

The denominator is equivalent to the total number of people who work in the private sector (employed and self-employed, formal and informal) across Kenya. For both $p(\theta)$ and $p(\theta_a)$ the sector-region-size probabilities sum to one.[47] 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.

We apply an exception for agriculture and non-market services: We estimate their $p(\theta_a)$ drawing only on formal private sector employment as formal VAT-paying firms occupy a very specific

[47] 12 and 47 refer to the 12 aggregate sectors and 47 counties respectively.

niche in both cases (see discussion in Section 5.2).[48] 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 cell versus the alternative probability $p(\theta_a)$ that takes informal ones into account? Sector-county types with a high degree of formality will have a $p(\theta)$ that is larger than the new $p(\theta_a)$. Put differently, their importance in the economy is overstated by the administrative data. For sector-regions with high levels of informality we thus expect $p(\theta_a)$ to be larger than $p(\theta)$. In Figure 12, we plot the share of total private sector employees who have a formal sector job[49] against the difference between the baseline $p(\theta)$ and the augmented $p(\theta_a)$s.[50] A 10 percentage point increase in formality leads to an increase of $p(\theta)$-$p(\theta_a)$ by half a percentage point (0.36 standard deviations).[51]

By using simple employment shares to compute $p(\theta_a)$, we rely on the assumption that the mapping of employees to firms is the same across all sectors and regions. This assumption does not hold 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 propose an alternative approache to compute $p(\theta_a)$, which entail re-scaling the number of employees using the average firm size in each sector-county-size cell. We use the administrative data to compute the average number of employees of bigger formal sector firms. 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).[52] We plot the plain vanilla measure and the alternative that accounts for firm size against the respective sector-region formality shares in Figure C1. Plotting them as the results of a local polynomial with confidence intervals reveals that the three measures behave similarly.[53]

Our suggested counterfactual accounts for informal firms being born into the network based on their sector-region profile. At the same time, our exercise assumes that linking probabilites

[48] Alternatively, we could have excluded them entirely from both the counterfactual and the model estimation. However, this would entail disregarding valid participants in the network. Especially commercial agriculture businesses are highly interlinked with local firms for non-agricultural inputs and services. Firms in these sectors participate in 10% of the observed firm-to-firm relationships.

[49] We use total private sector employment rather than the overall working population as the denominator to abstract from employment in public sector entities.

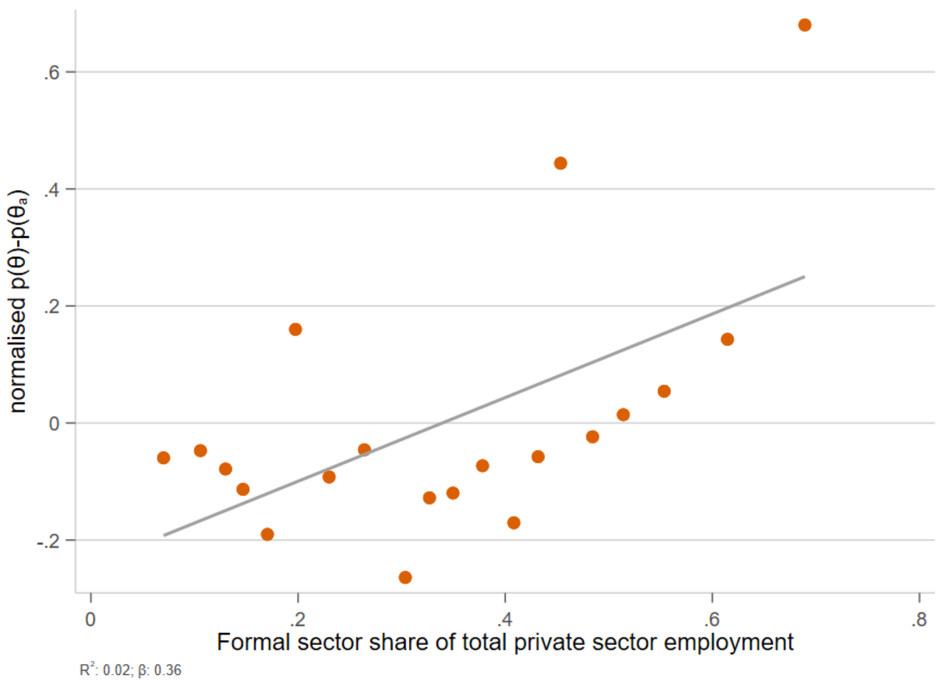
[50] We normalise the difference such that the y-axis can be interpreted in terms of standard deviations.

[51] To estimate the slope in Figure 12, we exclude five sector-county pairs which are adjusted by more than two standard deviations. All of the five sectors are Nairobi-based. The slope becomes twice as steep if the five sectors-county pairs are included.

[52] Medium-sized enterprises in the survey can have up to 100 employees. Using the enterprise survey comes with the caveat that the data is only representative at the county, but not necessarily sector-county level.

[53] The only notable difference is that correcting for firm size yields higher (but not significantly high) entry probabilities in sector-county cells with the lowest formal sector share.

Figure 12: Sector-region-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. For a version of the same graph as a local polynomial with confidence intervals and including two alternative measures see Figure C1.

$p(\theta, \theta')$ for small and informal firms are the same. We thereby treat $p(\theta, \theta')$ as a fundamental production technology. Rather than thinking of the counterfactual as adding new firms, we adjust the weights of each sector-region-size type.

7.2 Counterfactual results

7.2.1 Reduction in inequality in network links

First, we find that accounting for informal firms reduces the variation in outdegrees across counties by 5.6% (see Table 8). Meanwhile, the standard deviation to mean ratio in outdegrees across all sector-county pairs decreases by 4.4%. The dispersion of outdegrees across sectors on the other hand rises by 3.7%. We look at the coefficient of variation as the key metric. Adjusting for the mean accounts for the fact that the number of outlinks predicted by the model needs to be looked at in relative rather than absolute terms (see Section 7.1).

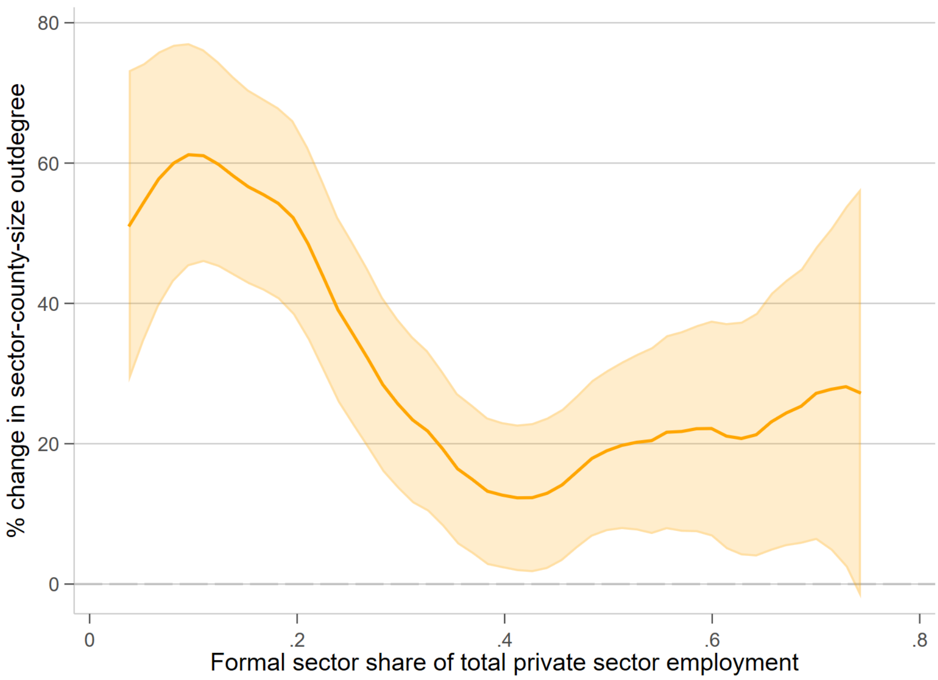
The reduction in variance in outdegrees across counties is also seen in Figure 14 where we plot the adjacency matrices, before and after accounting for informal firms, at the county and sector level respectively. For example, we find that Nairobi (the column in yellow-green) becomes less

Table 8: Changes in the dispersion of outdegrees

Aggregation	Δ sd/mean (in %)
County outdegree	-5.6
Sector outdegree	3.7
Sector-county outdegree	-4.4

The above table reports the result from our baseline counterfactual that adjusts entry probabilities of nodes in sector-county-size cells using employment figures from the 2019 population census. Results excluding the degrees of Nariobi and Mombasa or adjusting our estimates for the entry probabilities by proxies for the firm size distribution are reported in Table C1.

Figure 13: Predicted change in type-level outdegree and formal sector shares



connected once we update the probabilities. Counties are now less likely to buy from Nairobi than previously estimated. This result is driven by the fact that the updated network now includes a larger proportion of small firms, which we showed in the stylized facts section are less likely to source inputs from Nairobi. Additionally, Figure 14 shows that while inter-county variance in outdegrees reduces, this reduction could be driven by links within the counties rather than across. To this effect, we find that the diagonal entries in the adjacency matrix are larger when we account for informal firms than when we don't. This suggests that firms in these counties are (a) not as unconnected as we would predict if we did not account for informality and (b) the increase in their connectedness is mainly driven by intra-county links rather than inter-county links. This is again in line with the stylized facts in the previous section showing that smaller firms are more likely to source from the same county. We will discuss this result in more detail in the next subsection.

At the sector level, we find that other services, construction, manufacturing, and wholesale sectors gain the most links in relative terms. This is also seen in Figure 14. This is expected as these sectors have high levels of informality and are located downstream, where informality is concentrated.

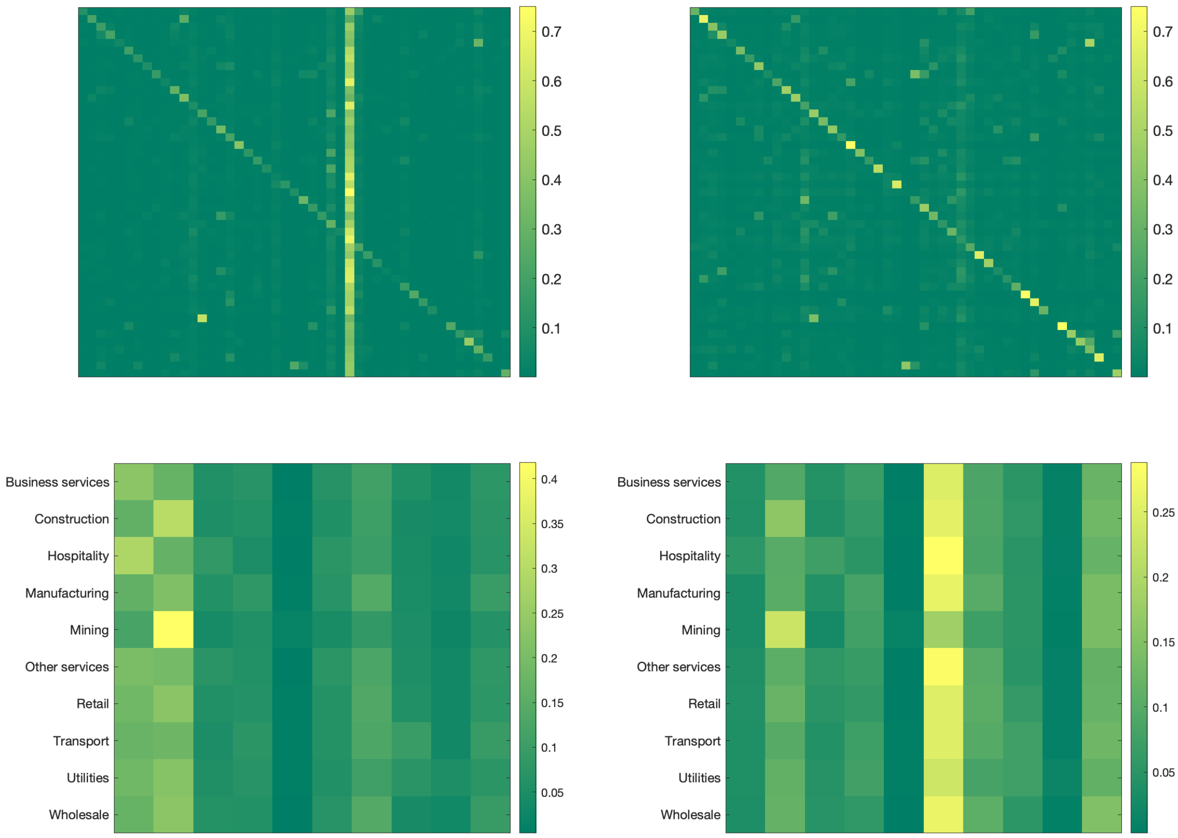


Figure 14: The above figures show heatmaps of the predicted adjacency matrix of the network (where row buys from column) as per the baseline $p(\theta)$ on the left and augmented $p(\theta_a)$ on the right at the county level (top) and sector level (bottom).

7.2.2 Reduction in inequality is driven by links within the same sectors and counties

As discussed in the previous section, we find that while counties previously considered to be less connected are more connected (once we consider informal firms), the increase in their connectedness is driven by trade within these counties. Consistent with this prediction, we find that the network is more partitioned after we account for informality i.e. is more likely to contain types that interact with the same types as opposed to other types. We quantify the extent to which the network is partitioned by looking at the network's modularity (Newman, 2006). The modularity of a network is higher if it contains groups that have more links among each other than one would predict in a network where links were formed at random. We compute the modularity of the weighted adjacency matrix at the sector-county level. We find that modulatiy in the counterfactual network with informal firms increases by 40%, which suggests that the counterfactual network exhibits a more pronounced group structure.

8 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 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 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.

8.1 Domestic shocks

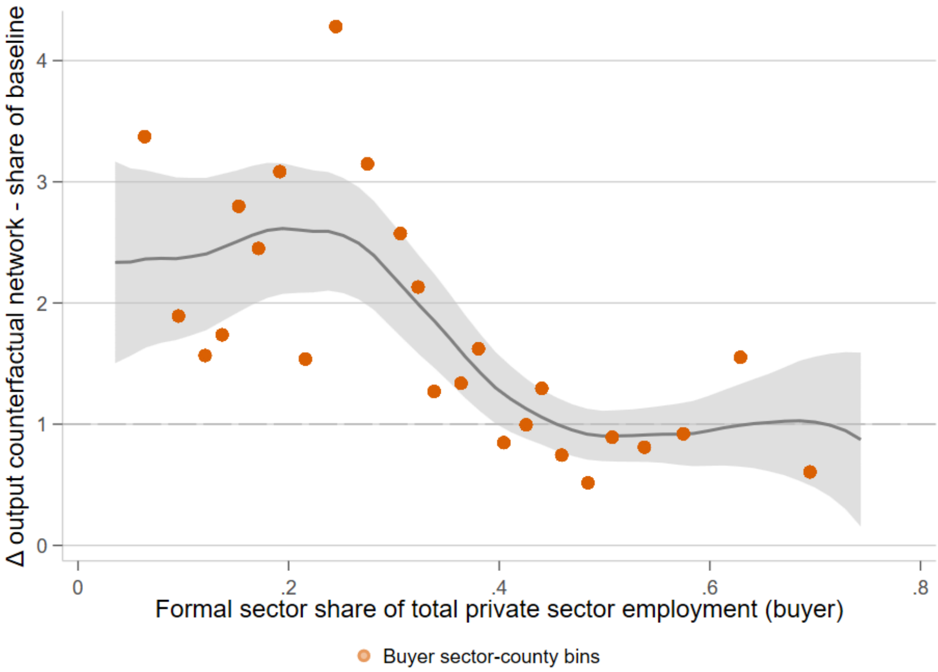
Following the supply side version of classic input-output models (Sargent and Stachurski, 2022), firm type j's output y_j in period t as a function of the sum of intermediate inputs it purchases from other types plus payments to other factors of production (value added) υ_{it}:[54]

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

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 υ_{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

[54]Alternatively, υ_{it} 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.

Figure 15: 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.

and then simulate the output process after a negative output shock to sector-region-size type i in the first time period.[55] We repeat this exercise for all types i.

To study the relevance of unobserved informal firms, we implement the entire simulation exercise twice. For the first case, we rely on the matrix π computed from the administrative records. For the second case, we use our alternative version of π, which incorporates the predicted number type-by-type linkages accounting for the presence of informal firms.[56] We ask, how does the shock impact a type j if we take or do not take informal firms into account? Given a fixed type i that is shocked in the first period, we compute the following measures for each $i \neq j$: We compute (i) the absolute value of the reduction in output of i, computed using the old adjacency matrix that does not account for informal firms, (ii) the absolute value of the reduction in output of i, computed using the new adjacency matrix that accounts for informal firms, and (iii) the percentage change in the absolute effect of the shock with the new and old matrix respectively.[57]

[55] 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. We further abstract from any endogenous network response to the shock at the type level.

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

[57] For both of the first two metrics we consider the average output reduction across all time periods.

In Figure 15, we plot the drop a sector and county's drop in output if we account for informal firms, as a share of the baseline scenario without informal firms in the network, against the formal sector share at the sector-region level. Relying on formal sector data only, we might underestimate the exposure of sectors and regions with a high incidence of informality to shocks. Put differently, the higher the incidence of informality in a sector and region, the more we underestimat the adverse impact of a domestic output shock. Specifically, a one percentage point decrease of the formal sector share corresponds with a 3.8 percentage point larger drop in the sector-region's output (see Figure 9).

Table 9 shows how the ratio of the impact of the shock using the counterfactual versus baseline network depends on the formality share of the sector and county. We find that the ratio is greater than 1 for 42% of the types implying that we underestimate the impact of the shocks in these cases. As the table shows, we find that this underestimation is larger for sector-county with a higher share of informality and this is not dependent on the type of specification used. Controlling for sector or county fixed effects also does not lead to significantly different coefficients. Moreover, out of the 41% of types for whom we underestimate the impact of the shock if we do not account for informality, 73% are types with small firms. This suggests that the underestimation is largely driven by not accounting for the prominence of small, informal firms in the domestic economy.

Table 9: % difference in output reduction depending on whether or not informal firms are accounted for

	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.

This simple exercise illustrates that not accounting for informality can lead us to underestimate how connected the network is and thereby under-predict how vulnerable different sectors and locations are to adverse economic shocks.

8.2 Import shocks

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

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

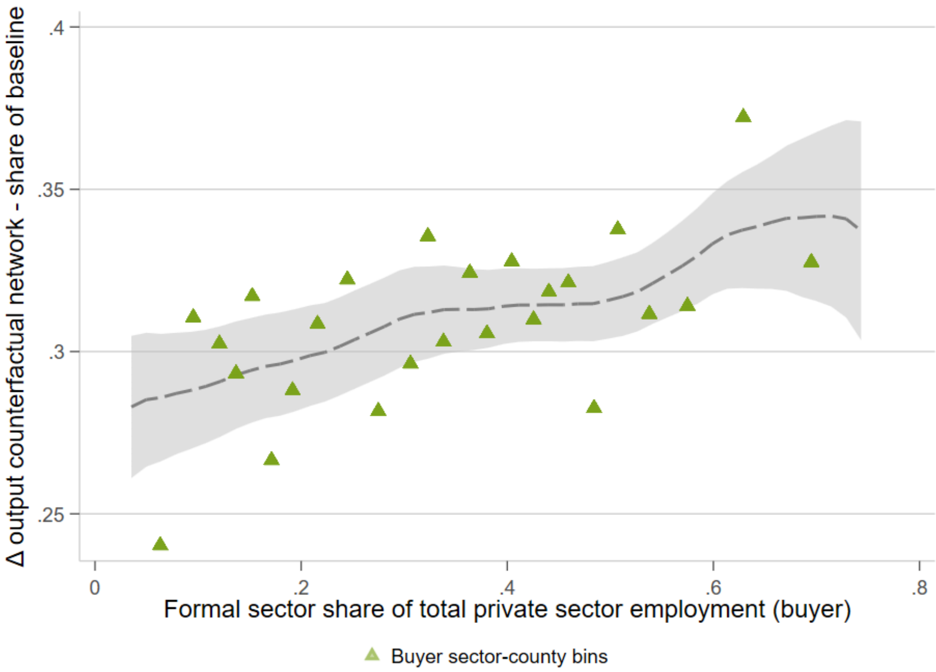
This is the same as before except that firm j's output now additionally depends on world output y_w in line with its import share m_{iw} that we obtain from the data. We re-normalise the rows of the adjacency matrix such that $\sum_j g_{ij} + m_{jw} = 1$. Next, we simulate a series of negative shocks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

Figure C2 shows the results by plotting the impact of the shock with the counterfactual network as a proportion of the impact using the original network. Unlike domestic shocks, we find that we would overestimate the reduction in output due to an import shock if we were to only take the economy's formal sector network into consideration. In fact, the impact of the shock is always lower (i.e. less negative) when we consider the counterfactual network as opposed to the original network.

Next, we show that this is particularly the case for sectors and regions where informal economic activity contributes a relatively larger share. In particular, we show in Table 9 that a 10 percentage point larger informal sector share corresponds to a 1 percentage point overestimation of the output reduction. Interestingly, we find that this effect is driven by formal firms in largely informal markets now being less exposed given we correctly account for their local connections.

Why do the predictions differ for domestic and import shocks? By accounting for informality, we make the types with a high informality share more prominent in the network and thereby more vulnerable to economic shocks. On the other hand, by accounting for informality we also reduce the prominence of types with a high share of formality. These types are very likely to have a higher import share and are more exposed to the international markets. By reducing their prominence (i.e. by adjusting their entry probabilities and accounting for informality) we find that the economy is more resilient to trade shocks and more vulnerable to domestic shocks. This explains why not accounting for informality leads us to underestimate the impact of domestic shocks and overestimate the impact of trade shocks.

Figure 16: 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.

9 Conclusion

We use granular transaction-level records to document a high degree of spatial concentration of trade amongst formal private sector firms in Kenya. This concentration can be traced back to the fact that the share of the formal sector increases with regional market size. We use a network formation model with preferential attachment (Bramoullé et al., 2012) to show that accounting for informal firms in the network increases the relative number of outlinks in counties with the highest level of informality. The overall spatial inequality in network links declines as a result. We further simulate how domestic output shocks propagate through the network using our counterfactual version with informal firms and find that using data on formal firms only, we are likely to underestimate the connectivity and thereby also vulnerability of smaller regions within the country to shocks.

An important question that lies beyond the scope of this paper concerns the optimal spatial concentration of economic activity from a welfare perspective. This is closely linked to further explorations into the underlying drivers of this concentration. Further, we are unable to speak to the question of whether the observed spatial concentration of formal sector firm networks is the result of market frictions or a feature of structural transformation (Gollin, 2008).

References

Acemoglu, D., Carvalho, V. M., Ozdaglar, A. and Tahbaz-Salehi, A. (2012), 'The network origins of aggregate fluctuations', *Econometrica* **80**(5), 1977–2016.

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

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-Urena, 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., 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.

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.

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*

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.

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* .

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. (2019), 'For the first time, the relative economic size of kenya's counties is clear', *World Bank Blogs* .
URL: *https://blogs.worldbank.org/africacan/for-the-first-time-the-relative-economic-size-of-kenyas-counties-is-clear*

Chacha, P. W., Kirui, B. K. and Wiedemann, V. (2022), 'Supply chains in times of crisis: Evidence from kenya's production networks', *CSAE Working Paper Series* **2022**(09.2).

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., Fieler, A., Xu, D. and Yang, K. (2023), 'O-ring production networks', *Journal of Political Economy* .

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*

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

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

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.

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.

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', *Job Market Paper* .

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

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

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

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

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

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

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

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

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.

Liu, E. (2019), 'Industrial policies in production networks', *The Quarterly Journal of Economics* **134**(4), 1883–1948.

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

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.

Spray, J. (2019), 'Search externalities in firm-to-firm trade', Job market paper.

Spray, J. and Wolf, S. (2018), 'Industries without smokestacks in uganda and rwanda', *Industries without Smokestacks: Industrialization in Africa Reconsidered* pp. 341–363.

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.

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.

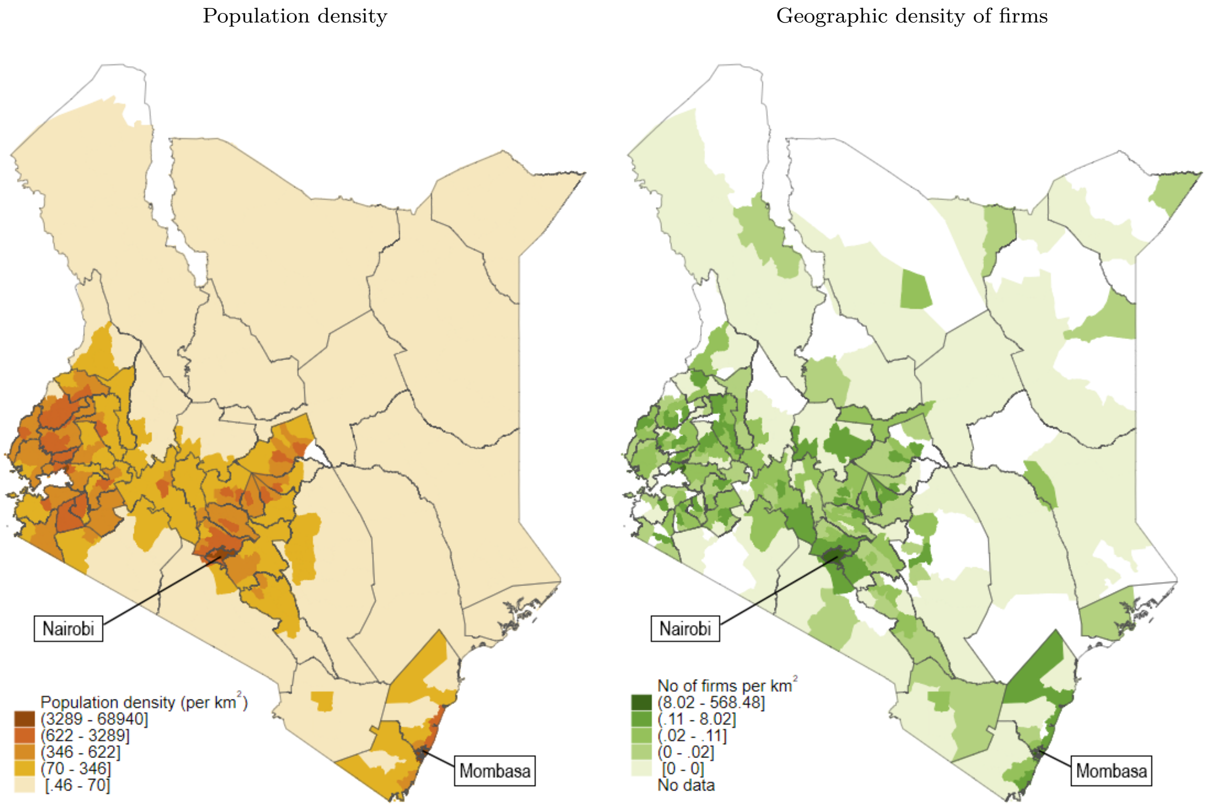
Zhou, Y. (2022), The value added tax, cascading sales tax, and informality, *in* M. Bussolo and S. Sharma, eds, 'Hidden Potential: Rethinking Informality in South Asia', World Bank Publications, chapter The Value Added Tax, Cascading Sales Tax, and Informality, pp. p. 61–90.

Appendices

Appendix A Complementary material for Section 4 on spatial trade patterns

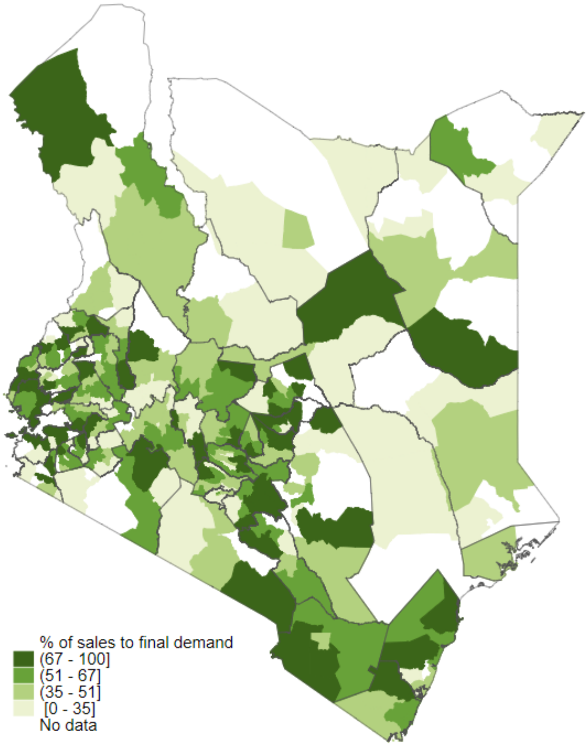
A.1 Additional graphs and tables

Figure A1: Firm headquarter locations and population density



The right map shows the density of firm headquarter locations at the sub-county level in 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 A2: Share of sales to final demand (non-registered entities)



The above map plots the average share of sales firms in each sub-county sell to non-registered entities. These are mostly consumers, but can also be non-VAT paying firms. County borders are outlined in grey.

Table A1: The intensive and extensive margin of domestic trade

All trading county pairs with origin and destination FE

	Total trade volume		# of links		Avg # of transactions per link		Avg volume per transaction	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Travel time	-2.353***	-1.137***	-1.906***	-1.012***	-0.448***	-0.176***	0.001	0.051
	(0.101)	(0.138)	(0.050)	(0.056)	(0.035)	(0.051)	(0.053)	(0.082)
Social connectedness		1.042***		0.767***		0.233***		0.042
		(0.086)		(0.040)		(0.032)		(0.055)
No. observations	1,652	1,652	1,652	1,652	1,652	1,652	1,652	1,652
R2	0.739	0.759	0.890	0.914	0.423	0.439	0.242	0.242
Origin FE	✓	✓	✓	✓	✓	✓	✓	✓
Destination FE	✓	✓	✓	✓	✓	✓	✓	✓

All trading county pairs with origin and destination characteristics

	Total trade volume		# of links		Avg # of transactions per link		Avg volume per transaction	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Travel time	-0.878***	-1.317***	-0.702***	-1.055***	-0.308***	-0.356***	0.132*	0.094
	(0.162)	(0.125)	(0.111)	(0.053)	(0.048)	(0.049)	(0.077)	(0.078)
Social connectedness	0.876***	0.823***	0.735***	0.691***	0.106***	0.102***	0.035	0.031
	(0.111)	(0.075)	(0.077)	(0.036)	(0.031)	(0.031)	(0.053)	(0.052)
Pop density (destination)	0.643***	0.222***	0.375***	0.023	0.138***	0.101***	0.130***	0.098***
	(0.056)	(0.047)	(0.034)	(0.019)	(0.016)	(0.017)	(0.027)	(0.030)
Pop density (origin)	0.667***	-0.034	0.467***	-0.077***	0.075***	-0.015	0.124***	0.059*
	(0.053)	(0.050)	(0.032)	(0.019)	(0.015)	(0.017)	(0.027)	(0.033)
Distance to NBO (destination)	-0.505***	0.248***	-0.537***	0.105***	-0.016	0.041	0.048	0.102*
	(0.108)	(0.094)	(0.074)	(0.038)	(0.030)	(0.036)	(0.051)	(0.062)
Distance to NBO (origin)	-1.096***	0.297***	-0.926***	0.153***	-0.212***	-0.029	0.041	0.174***
	(0.103)	(0.092)	(0.071)	(0.038)	(0.030)	(0.036)	(0.047)	(0.059)
Distance to MSA (origin)	-1.005***	-0.174***	-0.679***	-0.038	-0.210***	-0.100***	-0.115***	-0.035
	(0.080)	(0.063)	(0.053)	(0.026)	(0.024)	(0.026)	(0.038)	(0.042)
Distance to MSA (destination)	-0.184*	0.241***	-0.308***	0.055*	0.005	0.038	0.118**	0.149***
	(0.098)	(0.078)	(0.061)	(0.033)	(0.026)	(0.027)	(0.053)	(0.056)
No. buyers (destination)		1.195***		1.000***		0.103***		0.092**
		(0.056)		(0.023)		(0.022)		(0.036)
No. suppliers (origin)		1.709***		1.328***		0.220***		0.161***
		(0.056)		(0.022)		(0.021)		(0.038)
No. observations	1,652	1,652	1,652	1,652	1,652	1,652	1,652	1,652
R2	0.396	0.674	0.490	0.880	0.226	0.286	0.037	0.053
Origin FE	✗	✗	✗	✗	✗	✗	✗	✗
Destination FE	✗	✗	✗	✗	✗	✗	✗	✗

All variables are aggregated to the county level. The outcome variables for the OLS regressions are in log terms. Travel time is measured in log minutes. Social connectedness is measured using the Facebook Social Connectedness Index, which captures the total number of Facebook friendship connections between two counties, divided by the product of the number of Facebook users in each county (Bailey et al., 2021). The total number of possible county pairs is 2,209. We pool data from all available years to minimise the incident of zero trade flows. *; **; and *** denote significance at the 10; 5; and 1 percent levels respectively.

A.2 Exploring the robustness of spatial concentration with respect to multi-establishment firms

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. We use micro-data from the 2010 Census of Industrial Production (KNBS, 2010) to compare the spatial concentration of sales and firm location for all firms, including those with multiple establishments, to the spatial concentration in single establishment firms in Table A2. 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 anonyomised 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 observe 3,960 VAT-paying firms[58] 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 much as 61% of total sales in 2010. These figures are very similar to concentration of formal manufacturing firms reported by Storeygard (2016) for Tanzania.[59] Restricting the census data for Kenya to single establishments only, 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, but we aggregate their sales geographically based on headquarter information only. Nevertheless, the difference is not sufficient to explain the relatively higher spatial concentration in the VAT-paying sector relative to overall economic activity.

Table A2: Geographic concentration of industrial activity

	All firms		**Single est. firms**	
	in %	α	in %	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 = 3960
No. firms	64	0.51	-	-
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 are the estimated coefficients from a county-level regression of each county's rank (log) on the respective measure (log): $logrank = logA + log\alpha$. The Census of Industrial Production was carried out by KNBS (2010).

[58] 2015 is the earliest year for which the VAT records have been fully digitised.

[59] Dar es Salaam, Tanzania's primate city, accounts for 8% of its population (Storeygard, 2016), as very similar figure to Nairobi's population share in Kenya (KNBS, 2019).

Appendix B Complementary material for Section 5 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).

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	0.80
Employment (licensed MSMEs)	0.58	0.69	0.65
No. firms (licensed)	0.20	0.16	0.13

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 Fluctuations of the VAT-paying sector as a share of GDP over time

Figure B1 and Table B3 illustrate 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.[60] Secondly, certain 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. This is highlighted by the rapidly rising GDP for non-market services in Figure B2 (orange line), where we present the trends for each sector.

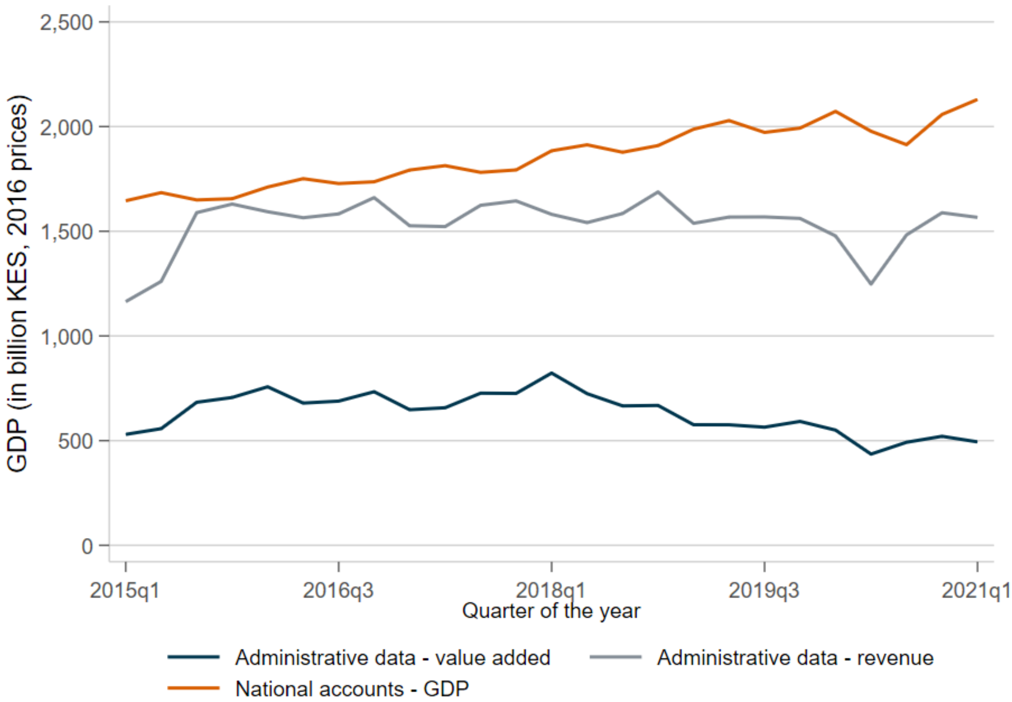
Table B3: Share of GDP covered in the administrative records

	Share of GDP (%)						
Year	All	ex Fin.	ex NMS+Fin.	ex Agri.	ex NMS+Fin.+Agri.	NMS	Agri.
2015	36	39	51	43	69	22	22
2016	40	43	56	46	74	22	21
2017	37	40	53	44	71	23	20
2018	37	40	53	43	70	23	20
2019	28	30	40	33	53	24	20
2020	24	26	35	28	46	24	21

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 capture 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 published by the Central Bank of Kenya: https://www.centralbank.go.ke/statistics/national-accounts-statistics/.

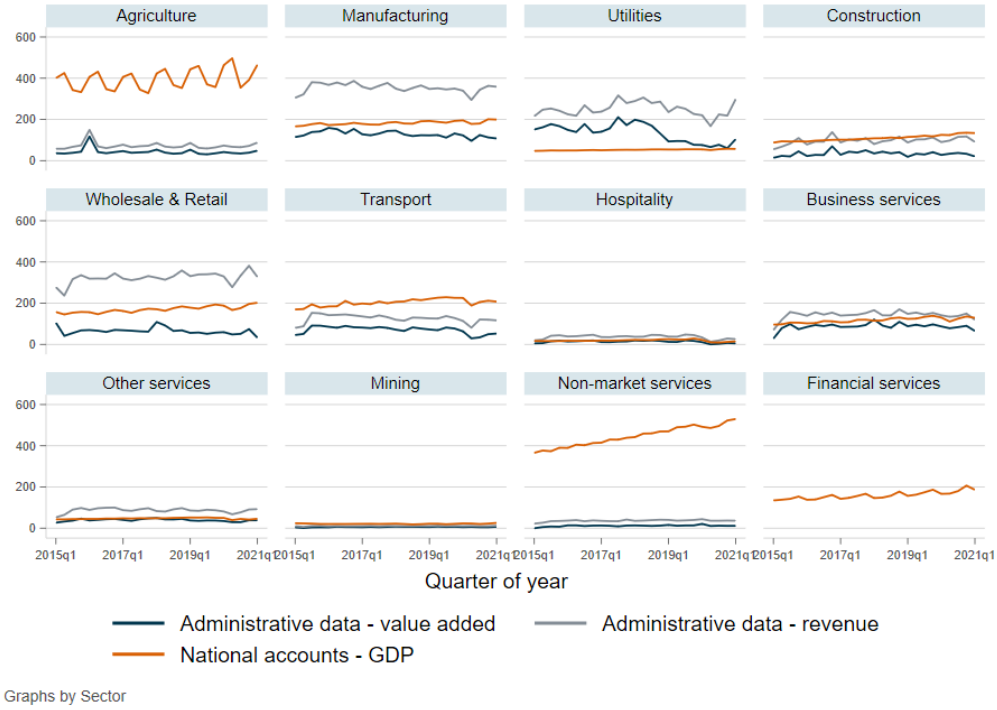
[60] The impact of this tax is particularly evident in the plot for utilities (electricity, gas, and water) in Figure B2. However, this sector alone cannot fully explain the overall downward trend and kink in the data.

Figure B1: Quarterly GDP in billion KES (2015-2020)



This figure compares the quarterly GDP in 2016 prices to the quarterly value added (domestic sales + exports - domestic purchases - imports - salaries) from the administrative data (deflated). The quarterly GDP time series was downloaded from the Central Bank of Kenya's database: https://www.centralbank.go.ke/statistics/national-accounts-statistics/

Figure B2: Quarterly GDP in billion KES (2015-2020)



This figure compares the quarterly sector-level GDP in 2016 prices to the quarterly value added (domestic sales + exports - domestic purchases - imports - salaries) from the administrative data (deflated).

B.3 Agriculture and non-market services

The most relevant excluded sector is agriculture, which generates 20-22% 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 6 shows that only a fraction of the sector's GDP is captured in the VAT data.

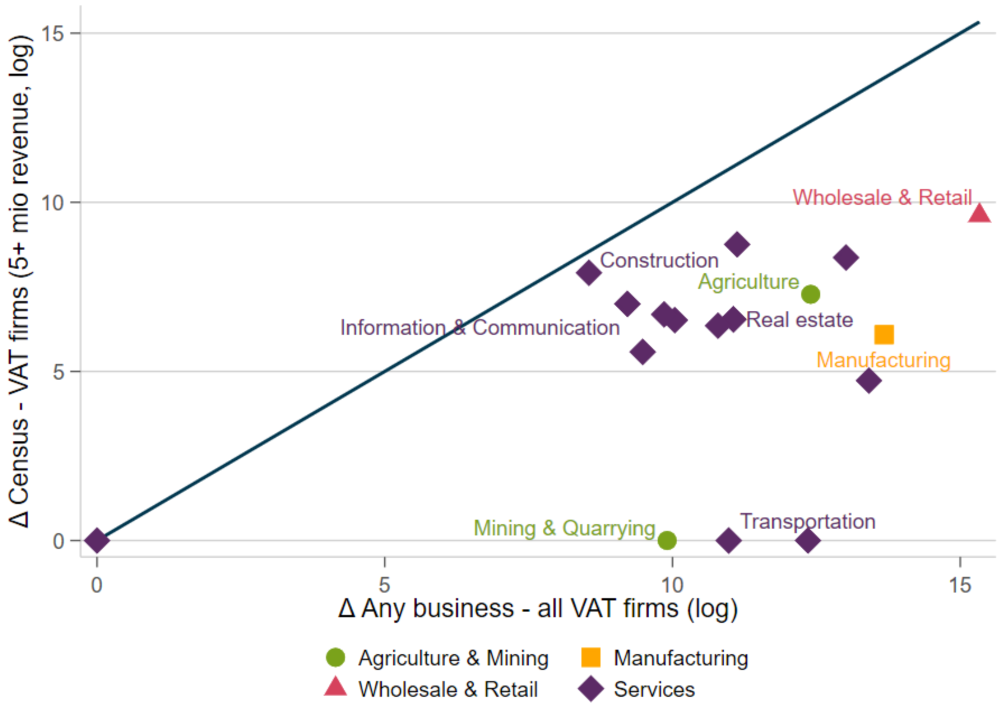
Non-market services include education, health, public administration and real estate (Herrendorf et al., 2022). They contribute 22-24% to Kenya's GDP, but are barely represented in the VAT data as most of the entities operating in these sectors are VAT exempt, not-for-profit, or the underlying sector's size in the national accounts being estimated using non-market prices (see penultimate column of Table B3). Figure 6 highlights another sizeable gap for "others", which includes international organisations, unclassified firms, and financial services.

Table B4: Sector-level comparison with the 2016 Census of Establishments

Sector	**Admin data**	**Census of Establishments**		Δ in %
		all	5+ mio. KES turnover	
Agriculture, Forestry, & Fishing	931	4,189	2,128	-56
Mining & Quarrying	239	297	176	36
Manufacturing	2,965	6,038	2,904	2
Electricity & Gas	1,031	84	67	1,438
Water supply	189	598	391	-52
Construction	7,351	11,846	7,096	4
Wholesale & Retail	7,991	44,023	21,175	-62
Transportation & Storage	2,333	2,241	1,123	108
Hospitality	2,474	8,364	1,982	25
Information & Communication	2,349	3,480	2,620	-10
Real estate	2,122	3,680	2,322	-9
Professional, Scientific & Technical	2,579	3,669	2,256	14
Administrative & Support Services	2,386	3,622	2,130	12
Public Administration	55	307	201	-73
Education	203	29,582	6,478	-97
Human Health & Social Work	470	2,427	1,109	-58
Arts, entertainment & recreation	433	575	215	101
Other services	1,339	8,440	5,174	-74
Households as employers	698	46	39	1,689
Total	38,138	133,508	59,587	-36

The last column captures the percentage deviation of the firm count in the administrative data (column 1) from the number of establishments with a turnover of more then five million KES (column 3) as per Census of Establishments.

Figure B3: Sector-level correlation of the extensive margin of informality



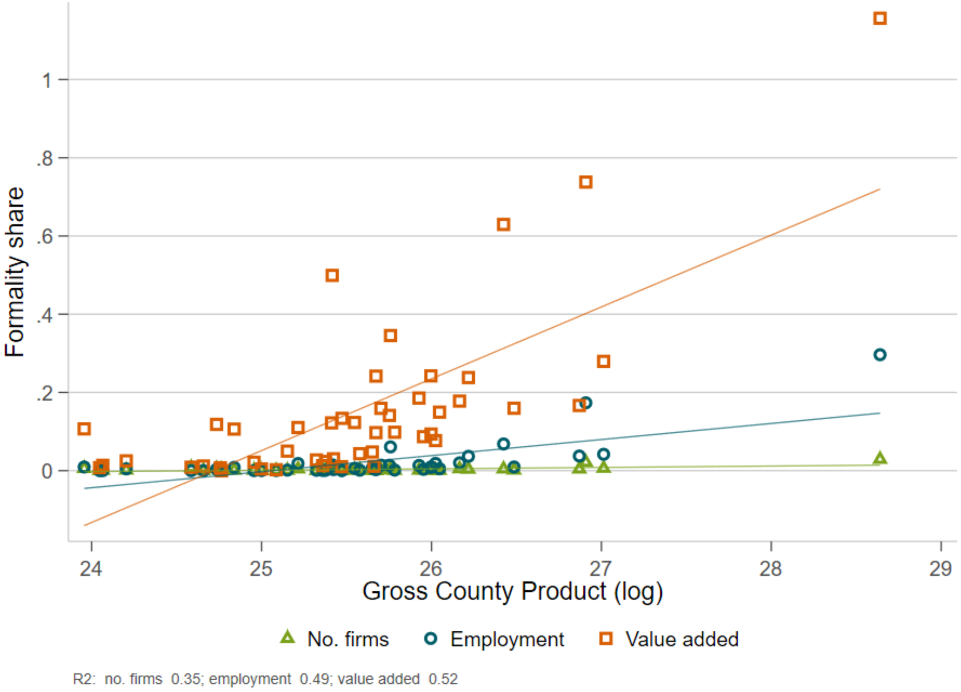
The above graph correlates the two types of extensive margin informality at the sector level. On the y-axis we put the gap between the number of firms with more than KShs 5 million of revenue from the Census of Establishments (KNBS, 2017) and the administrative data. The x-axis shows the gap between the total number of businesses (KNBS, 2016) and the number of VAT firms in each sector.

B.4 Informality, market size, and income levels

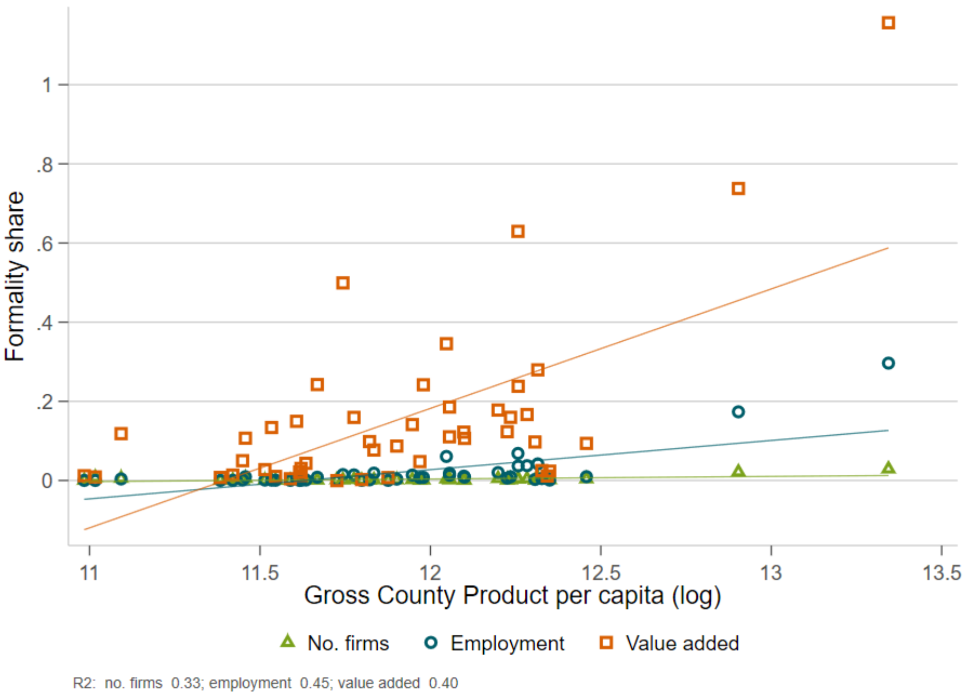
Figure B4: Informality, market size, and income levels

Correlation of the formal sector share and ...

... Gross County Product



... Gross County Product per capita



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.

B.4.1 Within-sector informality variation drives spatial disparities in aggregate informality

Having considered both the difference in informality levels across sectors as well as regions within Kenya, we ask which of the two margins matters more for spatial inequality in formal sector activity. In other words, to what extent does inequality in formal sector activity arise solely due to differences in sector compositions across space? Or does it arise because the share of formal activity within a sector varies across counties?

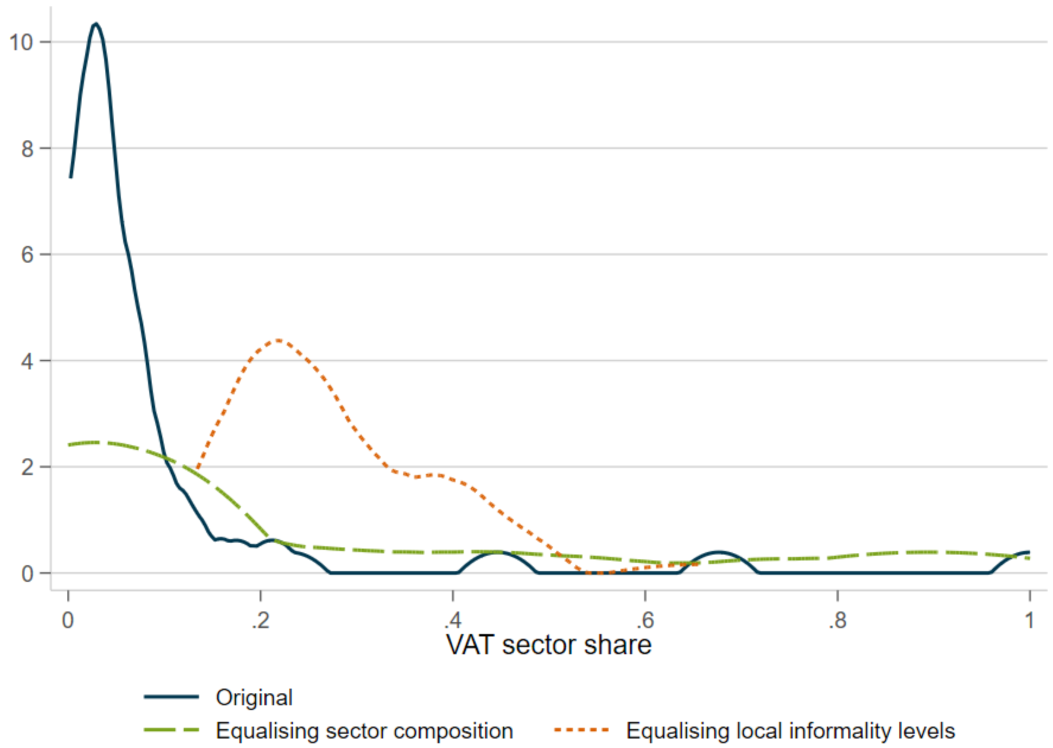
We rely on an accounting exercise with two counterfactual scenarios to study this question. We use the value-added based measure of informality for the purpose of this exercise. We start by noting that the aggregate formal sector share f_c is each county c is the sum of each sector s's degree of formality f_{sc} weighted by the respective sector's contribution to the county economy s_{sc}:

$$f_c = \sum s_{sc} \times f_{sc} \tag{5}$$

In the first counterfactual scenario, we alter the sector composition of each county's economy to align with Kenya's sector composition on a national level, i.e. we equalise s_{sc} for each sector to align with its national share s_{sKenya}. In the second counterfactual scenario, we keep the sector composition unchanged but harmonise the degree of formality of each sector f_{sc} to align with the national degree of formality f_{sKenya}. We plot the baseline dispersion in county-level formal sector shares in Figure B5 alongside the results from the two counterfactuals. First, note that the majority of counties have a formal sector share f_c of less than 20%. Only three counties have an above average degree of formality. The unweighted average formal sector share is 9%. By design, neither of the counterfactuals notably change weighted aggregate formality levels.[61] They do, however, affect the spatial dispersion. Altering the sector composition s_{sc} to s_{sKenya} increases the formal sector share in counties with the lowest levels of informality. As a result the average formal sector share increases from 9% to 11%. This moderate impact pales in comparison with the counterfactual that alters the formal sector shares f_{sc} to align with the national level of formality f_{sKenya}, but keeps the sector composition s_{sc} of each county constant. In this scenario, all counties' construction sectors, for example, get assigned the same degree of formalisation. As a result the (unweighted) average formal sector share f_c jumps up to 28%. As the dashed orange density plot in Figure B5 shows, the counterfactual pushes the formal sector share f_c above 20% in virtually all counties. The variance in outcomes shrinks drastically.

[61] While counties with low levels of informality will improve if national standards are applied, the top counties are worse off.

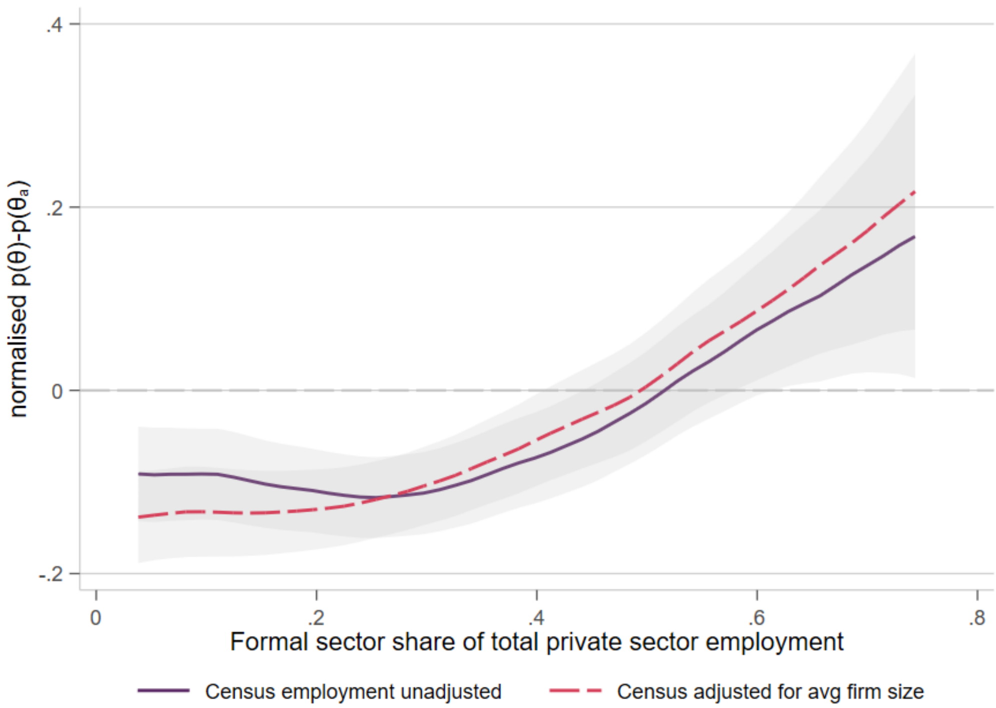
Figure B5: Spatial variation in informality - a story of sectors or geographies?



The above graph plots the pdf for the dispersion of formal sector shares across Kenya's 47 counties. The solid blue line captures the original dispersion in informality levels across counties. The long-dashed green line keeps local informality levels constant, but equalises the sectoral composition of the counties' economies by weighting each sector based on its share of the national economy. The short-dashed orange line instead equalises the local level of informality to be the same across all counties and to correspond with the respective sector's national formal sector share.

Appendix C Complementary material for Section 7 on the counterfactual

Figure C1: Sector-region probabilities and formal sector shares



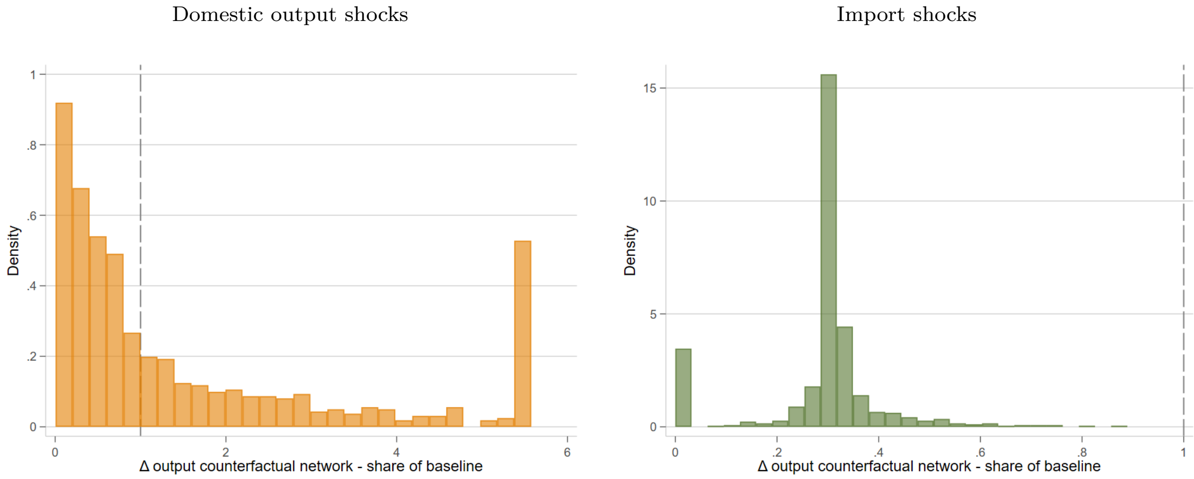
The graph plots each sector-regions formality share against the normalised difference between the baseline $p(\theta)$ and the augmented version $p(\theta_a)$ that takes into account informal firms. $p(\theta)$-$p(\theta_a)$ is reported in terms of standard deviations.

Table C1: Robustness checks for changes in the dispersion of outdegrees

County outdegree	Δ sd/mean (in %)
Baseline	-5.6
Baseline without Nairobi & Mombasa	-11.8
Adjust for firm size	-7.5

The above table reports robustness checks for the baseline findings reported in Table 8. The first two lines capture the results from the baseline counterfactual except for in the second row we exclude the degrees of Nairobi and Mombasa when we compute the coefficient of variation. Row three reports the results for alternative distributions of $p(\theta)$ that correct for differences in the firm size distribution across sector-county types.

Figure C2: 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 1 if we underestimate the impact of the shock and smaller than 1 if we overestimate it, if we do not account for informality.