Spatial Inequality and Informality in Kenya's Firm Network *

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

16th October 2023

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 patterns 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 network, over and above the concentration of aggregate economic activity. 90% of the cross-regional variation in trade volumes can be attributed to the extensive margin of trade, the location of firms and the number of firm-to-firm relationships they form. Using data from the population census and national accounts, we further show that informality is particularly prevalent in downstream economic activities and smaller regional markets. We link our insights about the sectoral and spatial composition with a network formation model to investigate how accounting for informal firms affects spatial inequality in firm-to-firm trade. We find that including informal firms increases the outdegree of firms in regions with the highest level of informal activity. Overall, our results suggest that not accounting for informal firms underestimates the connectivity and vulnerability of smaller regions to shocks, especially those that pass through hubs such as Nairobi.

JEL classification: F15, D85 , R11, R12, E26, E17.

*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, 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 - E-mail: verena.wiedemann@economics.ox.ac.uk

‡Kenya Revenue Authority - E-mail: bkkirui@gmail.com

§Department of Economics, University of Oxford - E-mail: vatsal.khandelwal@economics.ox.ac.uk

¶International Monetary Fund - E-mail: PWankuru@imf.org

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). A more limited scope for export-led growth models has led to an increased focus on domestic supply chains (Atkin and Donaldson, 2015; Grant and Startz, 2022). The structure of these networks can influence economic development by affecting the dispersion of welfare gains from trade shocks, infrastructure investments (Atkin and Donaldson, 2015; Arkolakis et al., 2023), the impact of industrial policies (King et al., 2019; Liu, 2019; Lane, 2023), the propagation of adverse shocks (Acemoglu et al., 2012; Grassi et al., 2017; Baqaee, 2018; Baqaee and Farhi, 2019), and the exposure of the network to systemic risk (Erol and Vohra, 2022). However, studying domestic firm networks is empirically challenging due to lack of granular and comprehensive data.

A rapidly growing literature has turned to using value-added tax (VAT) returns to track firm networks (Dhyne et al., 2021; Bernard et al., 2022; Coşar et al., 2022; Alfaro-Urena et al., 2022; Adão et al., 2022; Demir et al., 2023), 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 across the entire economy or how welfare gains of government projects are dispersed across space. We address this concern by drawing on transaction-level data from close to 57,500 formal firms and over four million supplier-buyer relationships in Kenya and complementing them with survey and census data on the regional and sectoral composition of overall economic activity. As the use of administrative data to map firm networks is becoming more and more popular, this paper provides insights on how including informal firms might alter the observed structure of the firm network across space.

We ask the following questions: What are the spatial patterns of trade flows among formal firms in Kenya? How are these patterns affected when informal firms are included? What does this mean for spatial inequality and the vulnerability of different regions to shocks? The Kenyan context is particularly well-suited to answer these questions. First, informal economic activity accounts for 25-35% of Kenya's GDP (Schneider and Enste, 2000; Elgin et al., 2021). Second, as East Africa's largest economy, Kenya boasts a substantial 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

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

can otherwise be difficult to quantify in many contexts as statistical bureaus tend to focus on national aggregates (Chacha, 2019).

We proceed as follows. Firstly, we establish three stylized facts concerning the spatial patterns of domestic trade among formal firms in Kenya. Subsequently, we utilize census and national accounts data to document five facts related to the informal sector. Based on these insights, we estimate a network formation model with preferential link attachment to investigate whether taking into consideration informal firms alters the spatial inequality in firm-to-firm links documented in the tax records.

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. 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). That is, regions are more similar in their purchase behaviour than in the composition of their downstream sales channels.

Next, we ask whether the observed patterns are a result of our limited view due to the spatial selection of firms into the administrative data or whether it is reflective of the economy's underlying structure. We document five stylised facts about the informal sector. First, the VAT-paying sector accounts for 34% of Kenya's GDP. Second, informal firms are usually located downstream of large (formal) firms. Third, informality decreases as regional market size and income levels increase. Fourth, the spatial concentration in economic activity is largely a result of the concentration of formal sector activity. Fifth, we use a data-driven accounting exercise to show that the sectoral composition of regions explains little of the spatial inequality in the level of informality. Instead, it results from differences in the level of formalisation within the same sector and across locations.

In summary, we document a higher share of informal activity in smaller regional markets that are dominated by downstream economic activities such as wholesale, retail, personal services, and

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

construction. VAT self-enforcement mechanisms are the lowest in more consumer-facing activities (Naritomi, 2019). We might therefore systematically undercount the number of outlinks of formal firms in these smaller markets and downstream sectors, which in turn could explain some of the observed divergence in patterns between in- and outlinks in the formal firm data.

To more formally, and quantitatively study the role of informal firms, we apply 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 and location. A new 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 links with firms of this type either via undirected search[3] or preferentially 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. This search behaviour can be rationalised by underlying information frictions about the quality of potential suppliers (Chaney, 2014). The model provides predictions for the number of links between firms of different sector-location types. We first estimate this network formation model to predict the Kenyan firm network as it is. We find that new firms choose 42% of their suppliers through undirected search, conditional on their bias, and the remaining 58% 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 58% of links being formed as a result of preferential attachment hence suggests that information frictions are potentially even more binding for firms in Kenya's domestic firm network.

We then simulate 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 to answer the question of interest: keeping constant the probability with which firms interact across sectors and locations, how do spatial patterns of trade change when informal firms are accounted for? The counterfactual results confirm 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. At the same time, these firms source disproportionately with the largest trading hubs in the country, including Nairobi and Mombasa. This in turn boosts the expected outdegree of these hubs. Therefore, accounting for the spatial and regional incident of unobserved informal firms increases spatial inequality across counties. This has implications, for example, when we think about the pass-through of shocks across space. By considering the formal network only, smaller regions with a higher degree of informality

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

appear less exposed to shocks to network hubs like Nairobi. A limitation of our setting is that we implement the counterfactual under the assumption that conditional on their sector and region, informal and formal firms link with other types in a similar fashion.

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 8 concludes.

2 Contributions

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

Our paper's first contribution is to describe the spatial mapping of the formal firm network in a context with a sizable informal sector, to show that the informal sector is relatively more important in smaller regional markets and downstream sectors, and to highlight the spatial patterns that we can miss if we do not account for these firms. Despite the large size of the informal sector in various contexts, informality along supply chains remains understudied to date. 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 are able to observe firm-to-firm linkages between VAT-paying and non-VAT paying, but registered, firms. We complement their 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.

Thereby, 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). Reyling 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 the 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 a 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).

Second, 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 the 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 on the potential bias that arises due to informality.

Third, we also contribute to a small but growing literature on domestic trade patterns and the role of supply chains in shaping cross-regional welfare outcomes in low- and middle-income countries (Allen, 2014; Atkin and Donaldson, 2015; Bergquist et al., 2021, 2022; Grant and Startz, 2022). Due to the granular transaction-level data, we are able to decompose the so-called intensive margin of trade into transactions and average trade volume per transactions.[4] The spatial decay in cross-regional trade is fully accounted for by the decline in the number of firm-to-firm relationships and transactions with distance. This concurs with findings on the key role of the extensive margin of firm sales, its outdegree, for firm size (Bernard et al., 2022) and subsequently how it governs international trade flows (Chaney, 2014; Eaton et al., 2022). Our finding that the average trade volume mildly rises with distance underscores the relevance of economies of scale in shaping (domestic) trade patterns (Coşar and Demir, 2018; Grant and Startz, 2022; Bergquist et al., 2021).

Finally, urban primacy, disproportionate agglomeration of economic activity in one or two primate cities, is a long-studied phenomenon (as published in Jefferson (1989), Jefferson, 1939; Ades and Glaeser, 1995; Henderson, 2002; Soo, 2005).[5] While the role of cities as hubs for the exchange of information (Pred, 1980) is a global phenomenon, Memon (1976); Alonso (1968) argue that considerable concentration of economic activity is expected in settings with fast changing economic and political environments where market participants attach a considerable price to

[4]The result adds nuance to the more general literature on domestic firm-to-firm trade, which has thus far mainly estimated the gravity equation for domestic trade at the relationship rather than transaction level and thus found the average trade volume per relationship to decline with distance (Arkolakis et al., 2023).

[5]Berry (1961) notes that how and why primate cities emerge is very much a function of a countries economic and political history. Given the vast heterogeneity in the history of urbanisation across the globe it is not possible to link it directly to a country's economic development.

up-to-date information. Locating in a country's primate city allows market participants to overcome information frictions. Thanks to novel data sources that have not been available to geographers and economists in previous decades, we are able to describe the individual margins of economic activity that contribute to the documented spatial concentration. While firm networks and the role of backwards and forwards linkages have long been at the core of theories around agglomeration (Venables, 1996), only the recent emergence of granular firm-to-firm trade data has allowed researchers to study them empirically (Miyauchi, 2023). Our finding on the relative importance of firm location and firm-to-firm links compared to within-relationship trade volumes is highly suggestive of a crucial role for search and matching frictions (Miyauchi, 2023). We further add to the literature on urbanisation, agglomeration and city sizes by documenting that the concentration of economic activity in Nairobi is even more a formal sector phenomenon than agglomeration of overall economic activity.

In Kenya, the economic dominance of Nairobi and Mombasa has been a concern for geographers, economists and policy makers alike since its emergence as a result of the colonial economic system and transportation infrastructure (Memon, 1976; Obudho, 1997; Otiso, 2005). Nairobi continues to dominate despite government efforts to change this (Obudho, 1997). Moreover, Obudho (1997) notes how concentrated formal sector employment was in the Nairobi of 1992, i.e. how much Nairobi dominated the formal sector relative to other urban centres. We are able to confirm Obudho (1997)'s result using more granular and comprehensive geospatial data that covers Kenya's entire economy.

The role of information frictions in rationalising the spatial dispersion of economic activity more generally, and firm-to-firm trade more specifically, also informed our choice for a network formation model. We provide an empirical stage for Bramoullé et al. (2012)'s theoretical framework and show how it can be translated from a context of academic citation networks to firm networks. Bramoullé et al. (2012) build on the seminal paper by Jackson and Rogers (2007). Chaney (2014)'s paper on the geography of trade networks, which has become a seminal paper in its own right for the trade literature, builds on the very same framework by Jackson and Rogers (2007) and introduces geography into it. While this class of models does not micro-found firm entry decisions and is very simplistic in its approach to modelling linking decisions of network agents, it emphasises the role of preferential attachment for rationalising the power law distribution common in many networks (Newman, 2018), including firm networks (Bernard and Moxnes, 2018). Preferential attachment summarises the notion that network agents link with others (make friends, cite papers, exchange goods) through their existing links (friends, cited papers, firms). In the context of firm networks the process of searching through existing suppliers or buyers can be rationalised by information asymmetries about the quality of potential future suppliers (Chaney, 2014).[6]

3 Administrative data

3.1 Description of data sources

Our analysis draws on mirco 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 not only contain information on the firms' aggregate monthly sales and purchases but further include two sections in which firms report detailed information on sales and purchase transactions with other businesses and VAT-registered entities. These records allow us to identify the supplier and the buyer for each firm-to-firm transaction, the transaction date, an unstructured product description, and the transaction volume. Sales to and purchases from non-registered parties (e.g., exempt parties, non-registered businesses, final consumers) are recorded as an aggregate monthly figure and are hence not tracked at the transaction-level. We deflate all variables denoted in monetary terms using the monthly Consumer Price Index (CPI) published by the Kenya National Bureau of Statistics (KNBS).

VAT applies to individuals and firms with an annual turnover of KShs five million and above ($40,500 as of December 2022). 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.

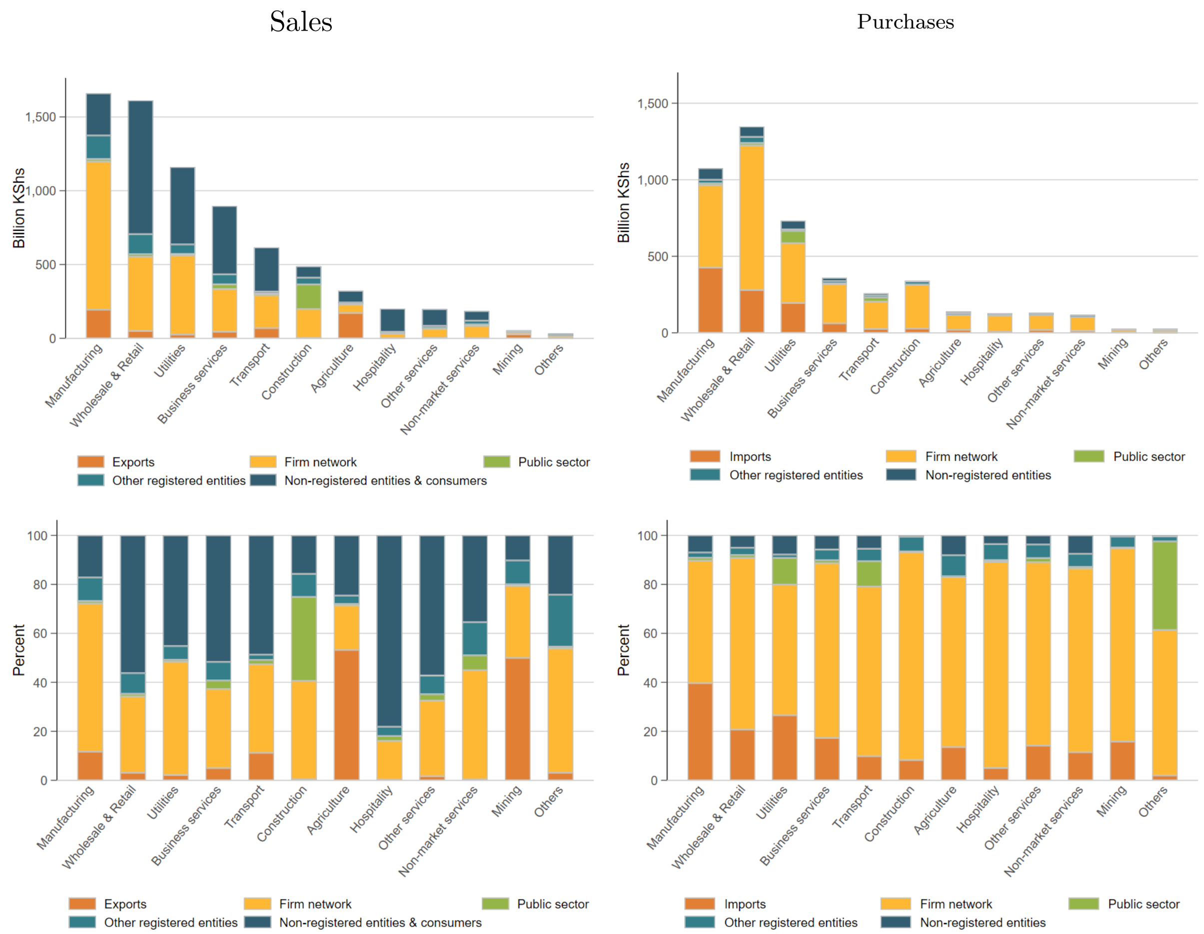
We filter the data set for entities that identify as private companies or partnerships in their tax-registration form. In doing so, we exclude all government-owned firms, government agencies, international organisations, NGOs, trusts, and clubs. We exclude firms operating in the financial sector (most 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. In 2019, we observe information on over 236,000 tax-paying entities in the VAT data. Our sample restriction reduces the number to 51,749 firms. However, the sales of those firms account for over 90% of the total sales observed in 2019.

Figure 1 plots the sector composition and the respective sales and input channels of firms covered

[6] An additional rational could be contracting and enforcement frictions (McMillan and Woodruff, 1999).

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.

3.2 Trade with informal firms in tax records

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.

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.

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

We begin by examining the geography of Kenya's firm network. Firstly, we quantify the level of spatial concentration within the network. Next, we break down trade within the network into its individual components, identifying firm locations and firm-to-firm relations as the primary margins driving spatial concentration. Lastly, we demonstrate that spatial inequalities are particularly pronounced in downstream trade flows. When discussing spatial patterns of economic activities, our primary focus lies in measures disaggregated at the county level, which represents Kenya's first administrative layer. In some instances, we also consider the second layer, sub-counties.

4.1 Urban primacy in Kenya's firm network

Kenya's firm network is strongly concentrated around its metropolitan areas Nairobi and Mombasa.[7] 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[8] and the region's communication network (Memon, 1976; Obudho, 1997).[9] 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)).[10] 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.[11] In 2019 (1960), as little as 9% (3%) of Kenya's population lives in Nairobi County and the city contributes 37% of Kenya's GDP outside the agricultural sector (see Table 7).

[7] 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.

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

[9] 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.

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

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

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.

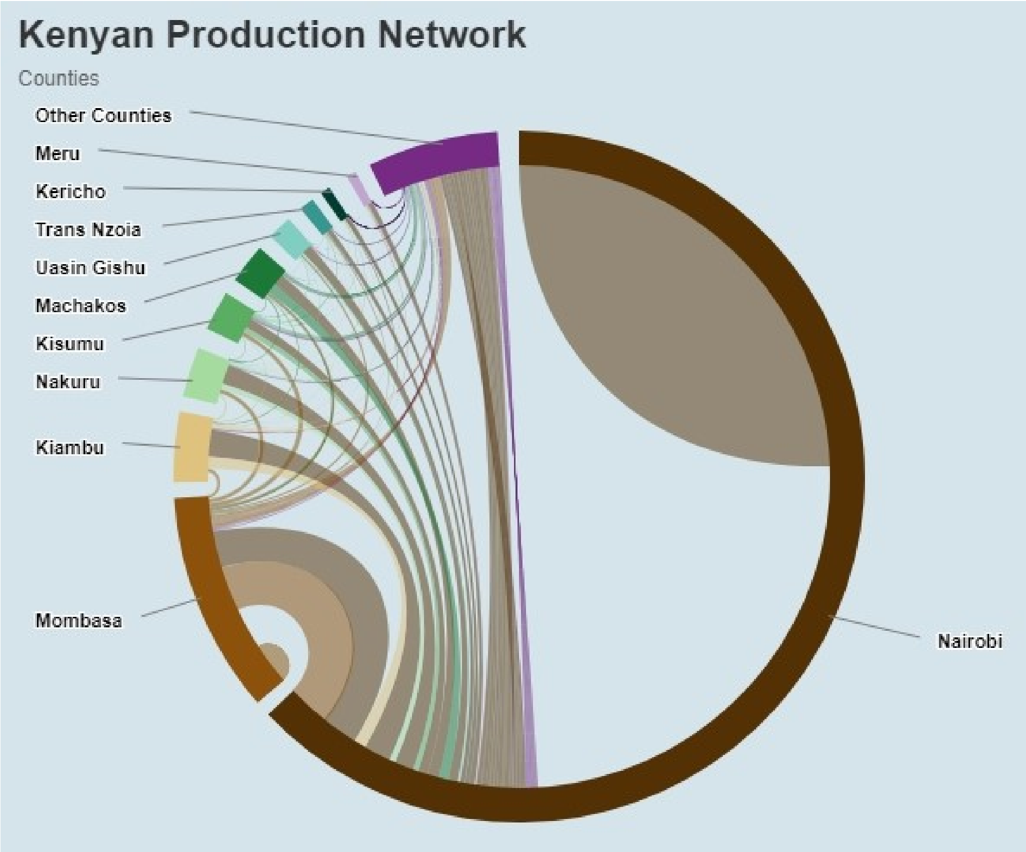
Table 1: Geographic concentration of economic activity in Kenya

	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. Such asymmetries in trade flows can exacerbate spatial frictions as logistics companies need to haul back empty containers along trading routes with a stark directional imbalance in trade flows (Ishikawa and Tarui, 2018; CAK, 2019; Wong, 2022).

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 indiacator is obtained via rank-size regressions (Gabaix and Ioannides, 2004).[12] A lower value indicates a flatter slope and hence more 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

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

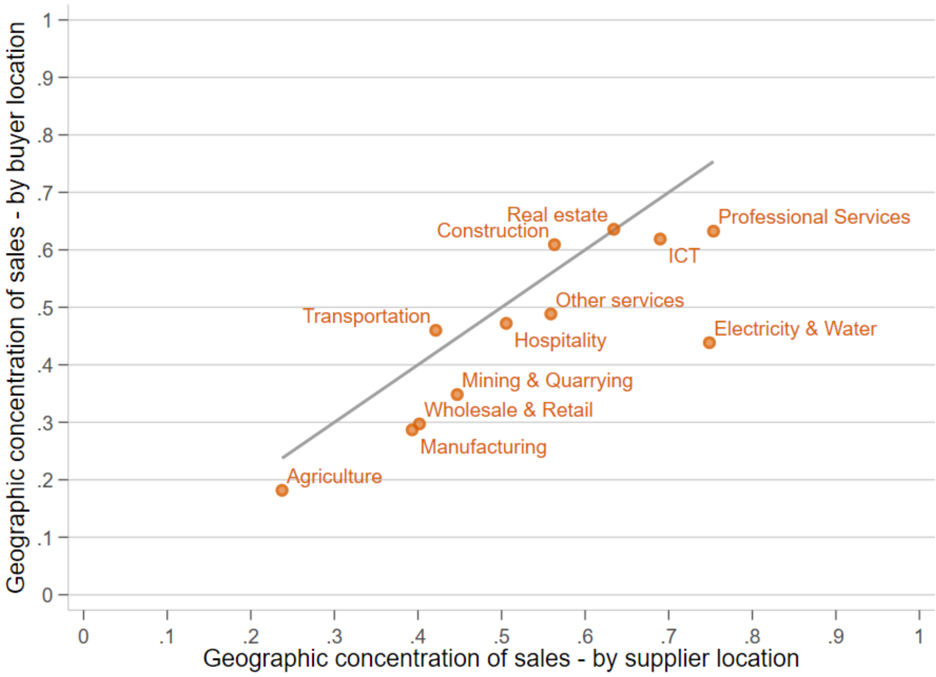
stipulates a power law distribution with an exponent of approximately one (Gabaix, 2009; Soo, 2005). At the same time, Kenya's population is more evenly distributed across counties than economic activity. Turning to the firm network, we find αs that are substantially lower than one, indicating a high 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 a smaller number of counties supply disproportionate amounts of inputs to the rest of the country.

This pattern aligns with our previous finding that trade flows out of Nairobi and Mombasa are larger than inflows. In Figure 3, we directly compare the spatial dispersion of sales relative to purchases.[13] If the geographic concentration of sales origins and destinations are the same, sector markers will lie directly on the 45° line added to the graph. As expected, the marker for most sectors lies well below the line, indicating that buyers of network inputs are more geographically dispersed than suppliers. This pattern is particularly pronounced for business services and the utilities sector, which are largely concentrated in Nairobi.[14]

[13] Our measure of dispersion is a Herfindahl-Hirschman Index computed from county-level trade flows and takes the value zero if sales are highly dispersed and one if they are highly concentrated.

[14] A caveat of this approach is that we draw on administrative boundaries to measure the geographic dispersion instead of actual geographic proximity. This becomes evident in the manufacturing sector, whose sales appear to be equally distributed. However, most manufacturing firms are based in counties bordering Nairobi, including Machakos and Kiambu.

Figure 3: Geographic concentration of sales based on supplier vs buyer location



The above graph plots the geographic dispersion of sales within the network based on the supplier's location against the dispersion based on the buyer's location. The dispersion measure is a Herfindahl-Hirschman Index and takes the value zero if sales are highly dispersed and one if they are highly concentrated. The solid blue line indicates the 45° line. To compute the geographic concentration of sales, we aggregate sector-level sales at the county level based on firm locations.

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:[15]

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

Table 2 summarises the share of the variance attributed to each term in both upstream (purchases) and downstream (sales) trade flows.[16] The number of firms operating in each county

[15] The same is true for purchases.

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

alone accounts for 67% of the variance in purchases across counties.[17] 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 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

We conclude this section by studying how each of the trade margins behaves in relationship to distance. Instead of considering trade flows in and out of a county, we turn to bilateral trade between two counties. In line with the classic gravity model of trade (Head and Mayer, 2014), trade flows within Kenya decline with distance and travel time. Consistent with the variance decomposition, this spatial decay in trade can be traced back to the decline in firm-to-firm relationships, the extensive margin of trade.

Total trade flows between two locations can be decomposed into the number of firm-to-firm relationships R_{od}, the average number of transactions per relationship $\bar{c}_{od}$, and the average trade volume per transaction $\bar{v}_{od}$:

$$\tau_{od} = R_{od} \times \bar{c}_{od} \times \bar{v}_{od} \tag{2}$$

Essentially, the above Equation 2 is the bilateral equivalent of Equation 1. In Figure 4, we aggregate trade flows to the county level and regress the total trade volume and each of its sub-

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

components on travel time between locations. Conditional on two pairs of counties trading with each other,[18] as much as 81% of the elasticity of overall trade flows with respect to travel time can be explained by differences in link formation between firms.[19] Put differently, the decline in trade flows over space is almost entirely driven by the number of firm-to-firm relationships declining. The elasticity result for the extensive margin of trade closely aligns with findings by Arkolakis et al. (2023) for Chile. Our granular transaction-level data allow us to go a step further and unpack what constitutes the intensive margin of trade in Arkolakis et al. (2023) - the average trade volume per relationship. We find that the spatial decay in the intensive margin is entirely driven by the decline in the number of transactions. The average trade volume per transaction on the other hand is zero or if anything increases with travel time between locations.[20] Rising average trade volume results are consistent with the role of economies of scale in governing trade flows and spatial price gaps (Grant and Startz, 2022).

We further document that social connectedness (Bailey et al., 2021) between counties is an important determinant of spatial trade flows over and above distance. The trade volume between any two counties increases with the strength of their social links. Social connectedness is measured at the county level and captures the probability of two random people drawn from two different locations being friends on Facebook.[21] Social connectedness can be a proxy for a number of different factors such as the number of transportation options between two localities or the strength of the trust network between potential trading partners. The finding backs up our result on the importance of the extensive marging. For the subset of trading counties, we find domestic trade flows in Kenya to be more than twice as sensitive to social connectedness than the international trade flows on a global level considered in Bailey et al. (2021).[22]

To study the importance of the number of buyers and suppliers in the origin and destination county, we drop the county fixed effects in an alternative specification (see Appendix Table A1). Instead, we include a series of other origin and destination characteristics like distance to Nairobi or Mombasa and population density alongside bilateral travel time and social connectedness. We find that including the number of suppliers and buyers in the origin and destination county plays

[18] The results can be found in table format in Appendix Table A1. Our results are robust to using a Poisson pseudo-maximum-likelihood approach to estimate the gravity equation, which allows us to account for zero trade flows between two pairs of counties. During the time period captured in our data, 75% of all possible county combinations engage in bilateral trade (1,652 out of 2,209).

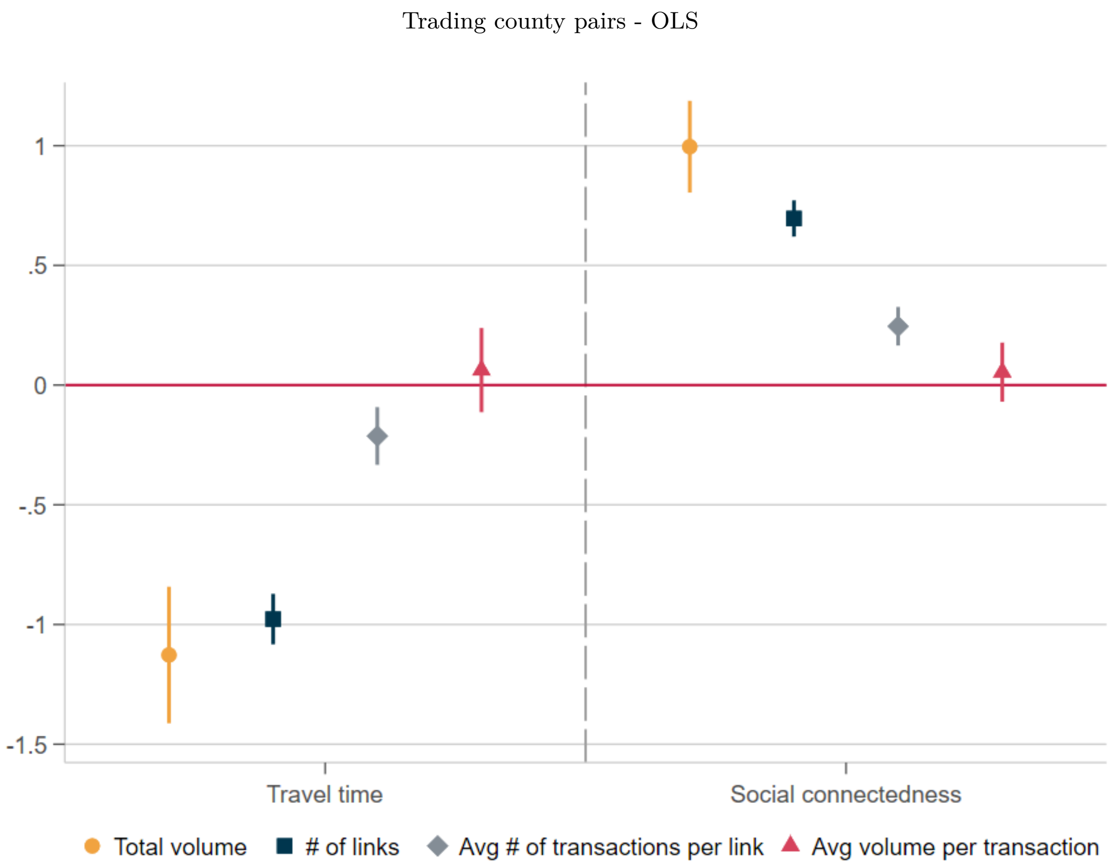
[19] Using distance rather than travel time does not change the result. Using the cross-section for 2019 rather than pooling firm-to-firm data across all years yields a slightly lower percentage of the overall trade elasticity that is explained by the number of firm-to-firm relationships (78%)

[20] Depending on the specification the coefficients are small, but positive, albeit often not significant.

[21] The measure is called the Social Connectedness Index and 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).

[22] Bailey et al. (2021) report a coefficient of 0.454 for social connectedness and -1.027 for distance, if both are included in the same regression with origin and destination fixed effects. Similar to their paper, we find that controlling for both social connectedness and travel time reduces the elasticity of each, indicating that they capture overlapping channels (see Table A1). At the same time each independently contributes to explaining variation in trade flows between counties.

Figure 4: The intensive and extensive margin of domestic trade



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 and travel time are included in the same regression. 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. The regression results can be found in table format in Appendix Table A1.

an important role in governing overall trade flows. The R^2 increases by 70% if we include the two regressors (column 1 versus 2).

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.[23] In other words, there is greater

[23]Under the premise that the in- and outdegree distribution follows a power law, we can again estimate the Pareto exponent, which is 1.85 for the indegree and 1.24 for the outdegree distribution. The outdegree distribution hence features more inequality in the number of links observed in the top end of the outdegree distribution. The point estimates are very close to what has been found in the Japanese (Bernard et al., 2019) and the Costa Rican (Alfaro-Urena et al., 2018) firm network (see Figure A3). The firm network in Chile (Grigoli et al., 2023) and the Dominican Republic (Cardoza et al., 2023) are also within the same range, but feature less inequality in their

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.[24] 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 5 shows a much more equal distribution of the average indegree across space. Firms in the median county (sub-county) have 22 (20) suppliers, 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.[25]

Understanding why the number of outdegrees varies substantially across locations is particularly important in a context with high informality. Figure A2, for example shows that firms outside the metropolitan areas on average sell a larger proportion of their sales to non-registered entities. 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. Recall that sales to consumers are lumped together with sales to non-VAT paying firms in the monthly tax returns. Similarly, one might wonder whether links between the metropolitan areas and the rest of the country are captured in a complete manner. This question matters for the spatial pass-through of shocks that are more likely to affect firms in Nairobi, e.g., international trade shocks. Erol and Vohra (2022) show that in networks with a core-periphery structure, systemic risk is closely linked with the size of the core.[26] While we cannot directly recover the importance of the informal firms for the supply chains of VAT-paying firms, we use third party data and a network formation model to narrow in on the informal sector and its position within the firm

upper tail. Note that there are some discrepancies between contexts that can be attributed to differences in the way the Pareto exponents are estimated and reported. Bernard et al. (2019) follow the broader literature on power laws in economics (Gabaix, 2009) and estimates the Pareto exponent α by regressing the inverse CDF (log) on the degree (log). Grigoli et al. (2023); Cardoza et al. (2023); Alfaro-Urena et al. (2018) estimate the Pareto exponent by regressing the log degree on the inverse CDF (log), which yields an estimate that is close to $\frac{1}{\alpha}$ and report the resulting regression coefficient as the Pareto exponent. Looking at the inverse, all estimated exponents fall within the range 1.24 and 3.5.

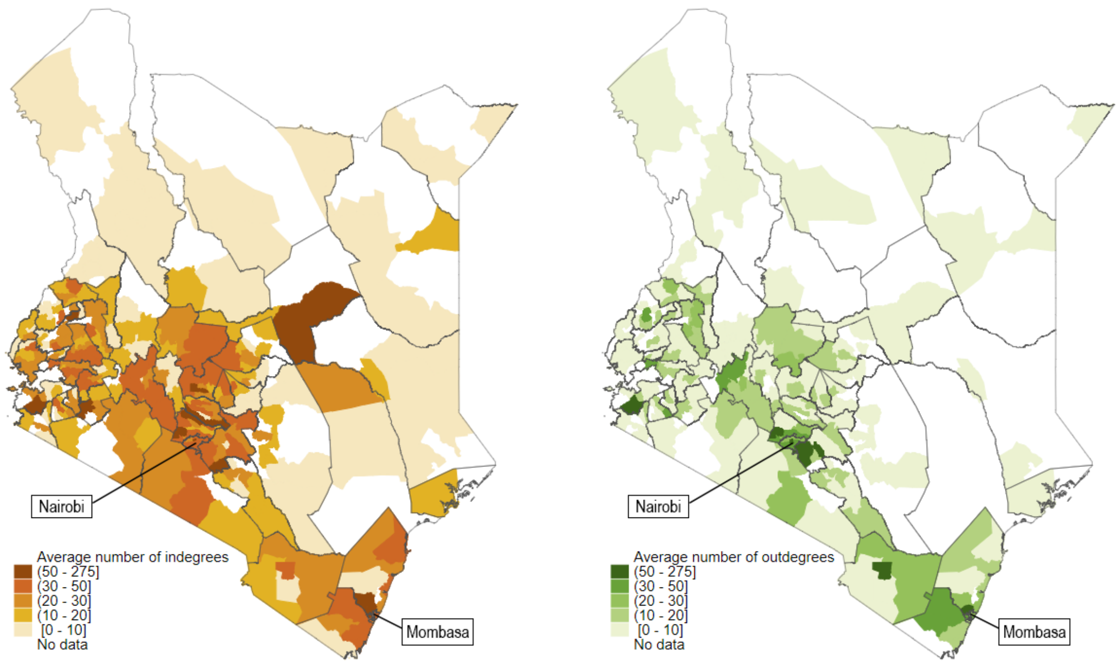
[24]The contrast is even starker of course if we also exclude Mombasa and Machakos County from the average. Machakos is coloured in dark green in Figure 5 as it serves as an industrial hub with many large manufacturing firms.

[25]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.

[26]Core-periphery networks are characterized by a core subset of nodes with strong links among each other and a periphery subset of nodes that maintain connections with the core but demonstrate relatively weaker connections with each other. Periphery nodes can also form clusters among themselves. These clusters are, however, then only loosely linked with other periphery clusters.

Figure 5: 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.

network.

5 The role and position of the informal sector

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 clear definition of informality within our context. Then, we detail and examine the additional data sources employed to study informality. Finally, armed with a precise definition and relevant data, we present five 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). Ulyssea (2018) documents an intensive margin of informality for employment i.e. informal employees in otherwise formal, registered firms. A wholesaler we interviewed in Nairobi's Central Business District[27] explains how the notion of an extensive and intensive margin of informality extends to firm-to-firm transactions:

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

Table 3 summarises four different margins of informality that can occur in firm networks: an extensive margin at the firm-level and an intensive margin at the transaction-level. Within each category, informality can occur due to either non-compliance or simply because a unit is too small to be taxed.[28]

Table 3: 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. In our setting, as discussed in Section 3, 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. 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

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

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

to alternative definitions of informality. 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.

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.[29] Omission of transactions between two formal firms or the misreporting of trade volumes on the other hand remains a concern for us.[30] We are able to recover some of the omitted transactions 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 4 provides an overview of the data sources we draw on, while Table 5 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.[31] However, we will also rely on measures

[29] 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). Note that transaction thresholds are different from firm size thresholds that consider a firm's overall annual revenues (see Section 3).

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

[31] While the criterion for a firm to be considered formal in our setting is whether or not it pays VAT, we prefer to refer to the underlying data set as administrative data for the simple reason that some of the variables like employment, sector, location, and age are collected from tax records outside the VAT system.

of informality that only rely on mirco data and national accounts data published by KNBS.[32] 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 4: 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).

Throughout the remainder of this section, we draw on three different indicators for economic activity, namely employment figures, the number of firms, and value added (sales - purchases). Our default measure of informality will be an employment-based measure that draws on the 2019 population census (KNBS, 2019). The two alternatives based on the number of firms and value added rely on estimates of the universe of businesses in KNBS (2016)[33] and on estimates of the regional economic size captured by the Gross County Product (KNBS, 2022) respectively. While the employment-based measure is likely to primarily capture the extensive margin of informality, the value-added-based measure can be considered as an aggregation of all margins. The number of firms will provide a more nuanced picture on the extent of potential non-compliance by firms 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. Further, we can distinguish between employment in the private compared to the public sector. None of the other measures allows for this distinction.[34]

We exclude any employment or firms in the agricultural sector or non-market services when

[32] Important for our purposes is that none of these data sources are explicitly based on tax records. A list of tax filing businesses informed the data collection of the Census of Establishments (KNBS, 2017). However, this is the source of data we will rely the least on. Tax revenues might inform estimates of the Gross County Products (KNBS, 2022).

[33] 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.

[34] Sometimes the lines are not clear cut. The Census of Establishments (KNBS, 2017), for example counts a lot of private and public schools that have some form of low-revenue, but for-profit activity happening at their site. These establishments account for almost a third of the firm count in this census.

Table 5: Measures of informality

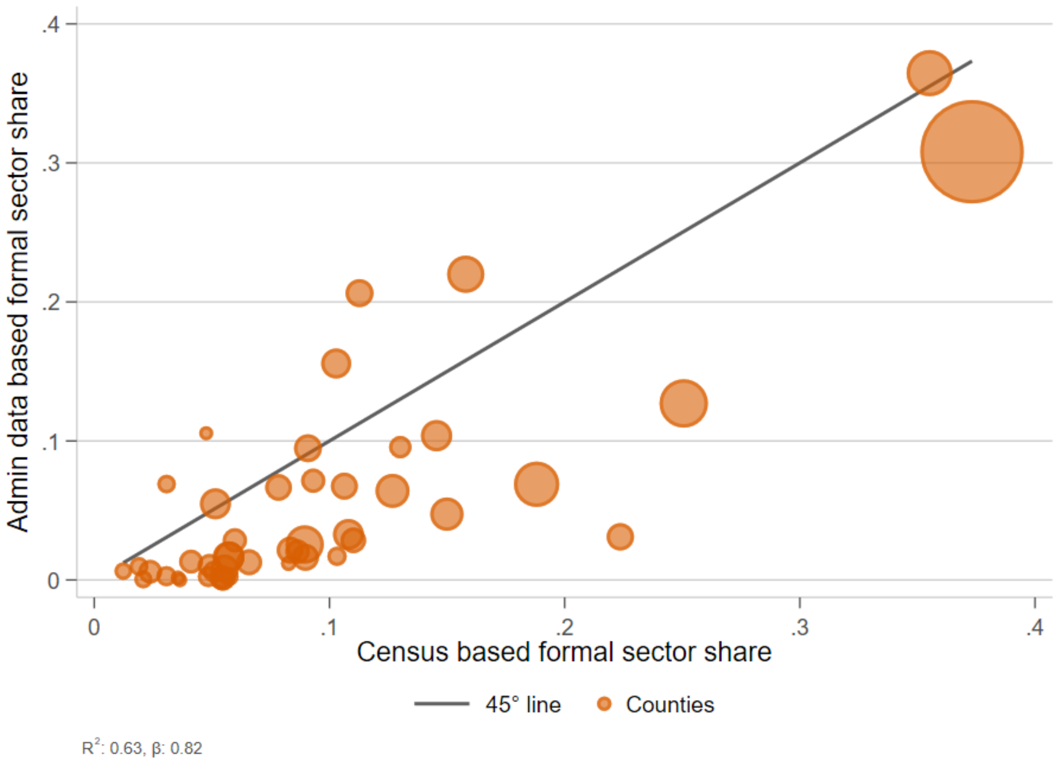
Unit	**Numerator** (formal sector)	**Denominator**	**KNBS**	**Use admin data**
Employment	No. formal priv. sector employ.	Working population	Census	✗
Employment	No. employ. in licensed firms	No. employ. in all firms	MSMEs	✗
Employment	No. employ. VAT firms	No. employ. in licensed firms	MSMEs	✓
No. firms	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 4. The term "all firms" refers to both licensed and unlicensed businesses based on KNBS (2016) estimates and county government records.

measuring informality at the regional level.[35] In both cases, the tax records cover a small and very specific sub-population of firms and employees only. In the case of agricultural firms, the administrative data cover large-scale commercial agriculture only. The vast majority of these firms are export oriented (Chacha et al., 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.

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

Figure 6: 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.

Our preferred measure of informality considers formal sector employment as per the 2019 population census. It correlates strongly ($\rho = 0.83$, Table 6) with its counterpart based on the administrative records (also see Figure 6).[36] 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. Table 6 summarises the correlation coefficients of the alternative informality measures. 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.

[36] To avoid mechanical correlation between the two measures we use total employment in licensed firms as the denominator for the administrative data. The KNBS estimate for employment in licensed firms is based on micro data that is distinct from the population census. Alternatively, one could use total employment in all MSMEs, which, however, includes many self-employed people. The correlation results are very similar for both alternatives.

Table 6: 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.

5.3 Five stylized facts on informality in space

We document five stylised facts about the informal sector.

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 B1).[37] This is substantially lower than for high income economies like Chile, where 80% of the country's GDP can be attributed to VAT-paying firms (Huneeus, 2018). 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.[38] The second reason is informality. In the following, we first discuss the influence of sectors that are not well-represented in the data in explaining the gap, then informality and its different margins.

If we exclude sectors that are to a large extent exempt from VAT (non-market services,[39] agriculture) or have special reporting rules applied to them (financial services), the VAT sector accounts for 64% of residual economic activity, implying an overall informal sector share of 36%.[40] This estimate suggests a larger informal sector compared to the 26% estimated by Elgin

[37]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.1.

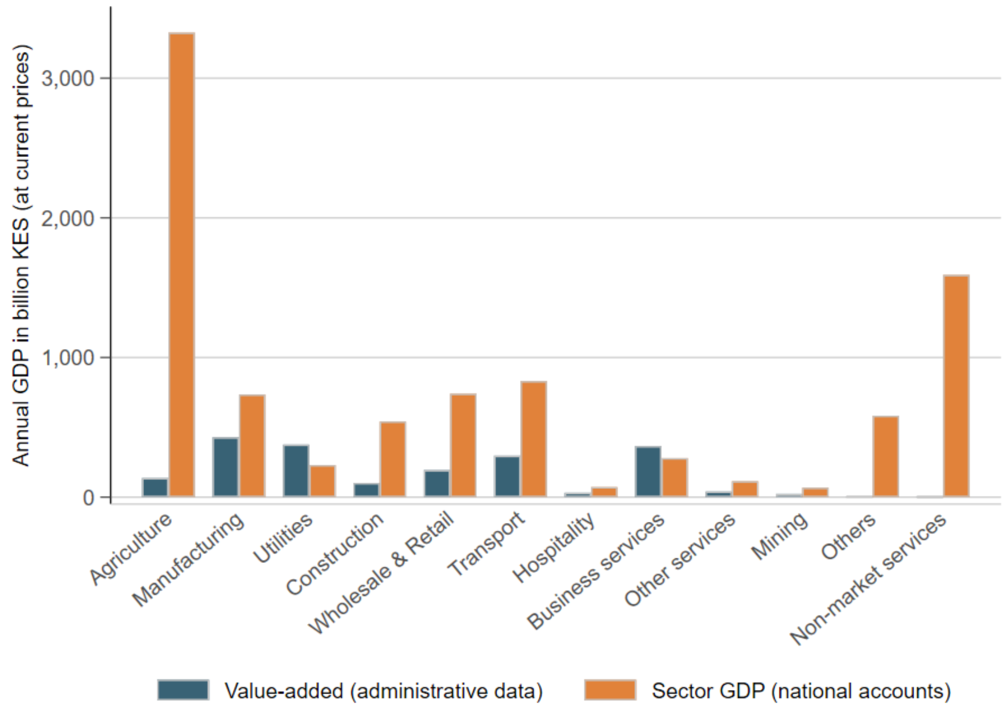
[38]Financial service firms, for example, are the biggest contributor to value added by VAT firms in Chile Huneeus (2018), although they are small in numbers. But, we do not consider financial services in Kenya due to differences in reporting requirements relative to other sectors.Alfaro-Urena et al. (2018) exclude them for similar reasons in Costa Rica.

[39]Non-market services contribute another 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 Appendix Table B1). Non-market services include education, health, public administration and real estate (Herrendorf et al., 2022). Figure 7 highlights another sizeable gap for "others", which includes international organisations, unclassified firms, and financial services.

[40]To arrive at this number, we exclude financial services, non-market services, and agriculture from total GDP. Appendix Table B1 details how the GDP share of the VAT sector changes if each of them is added or removed

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. In addition to our relatively more simple approach to arrive at 36%, we also focus on the VAT sector only. By doing so we apply one of the most stringent possible definitions of informality for firms.

Figure 7: 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 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 7 shows that only a fraction of the sector's GDP is captured in the VAT data.

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 7, shows that manufacturing and professional services (a sub-component of business services) are best represented in the administrative data.[41] This pattern aligns with the fact that both sectors are

sequentially.

[41] The ranking of sectors and variation in the formality share observed in Kenya is closely aligned with the findings of Murthy (2019) in India.

well-connected in the firm network and mostly buy from and sell to other formal firms.[42]

The gap between value added and GDP is an aggregate of all margins of informality discussed in Section 5.1.[43] We are unable to quantify the extent to which each margin contributes to aggregate informality. However, we are able to study the importance of each margin across sectors. To quantify the prevalence of informality on the extensive margin, we compare the number of VAT-paying firms to a number of alternative firm counts by KNBS.

The first type of extensive margin informality arises due to non-reporting firms with revenues above the VAT threshold. Figure 8 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, ICT, and professional services show the least deviation between the two data sources (see Table B2 for the same comparison in tabular form).[44] Wholesale & retail stands out as the sector with the largest number of non-compliant firms.[45] This is unsurprising given VAT self-enforcement is weakest for consumer facing firms (Naritomi, 2019), a theme we explore further below.

While the VAT-paying sector accounts for a substantial share of GDP, VAT paying firms are easily outnumbered by a large volume of small businesses with annual revenues below the VAT threshold. 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.[46] 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. Figure 8 plots the business count by data source and sector. For agriculture, utilities, and construction, the overall number of licensed businesses aligns closely with former two sources. 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. For both cases of firms that do not show up in the administrative data, firms in the wholesale & retail sector show the highest incident of informality.[47]

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

[43]The most standard measure for informality in the literature is employment. Private sector firms observed in our sample only employ 0.92 million employees. This corresponds to 4.6% of the overall working population in the 2019 population census. The number improves to about 10.8% in the un-filtered raw data, which covers parts of the government sector, most of the non-profit sector, as well as the financial sector in the tax returns.

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

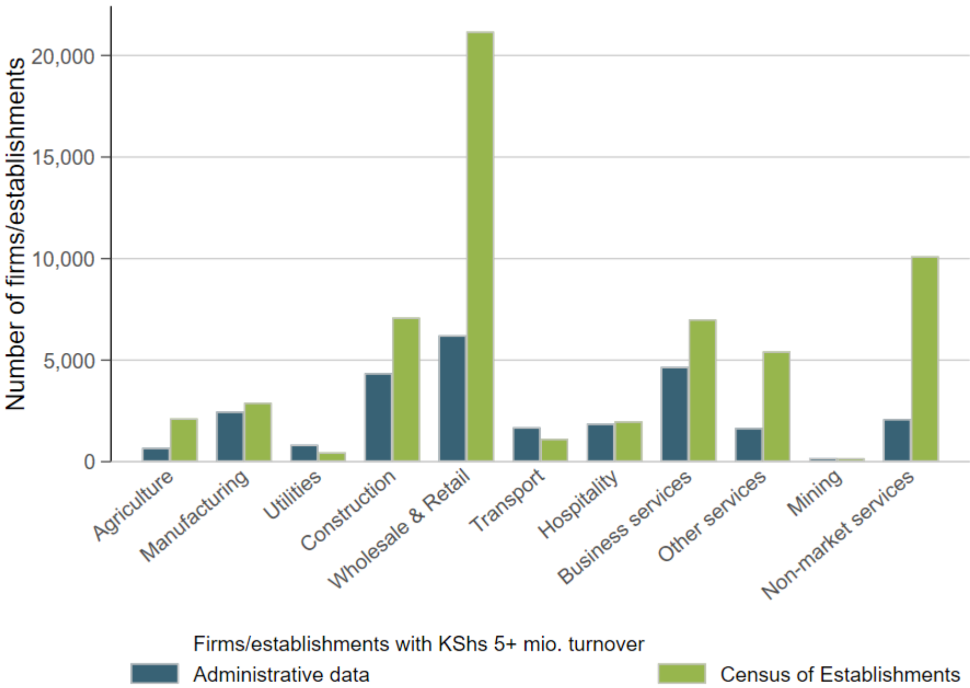
[45]In line with the discussion above, the gap for education & health is mainly a result of VAT exemptions that apply to most firms in this sector. The CoE, for example, includes public sector establishments such as public schools that offer educational services.

[46]The majority of unlicensed businesses is not permanent. 94% have a monthly revenue of less than KShs 50,000 (about $400). KNBS (2016) estimates they contribute about 2.4% of Kenya's GDP.

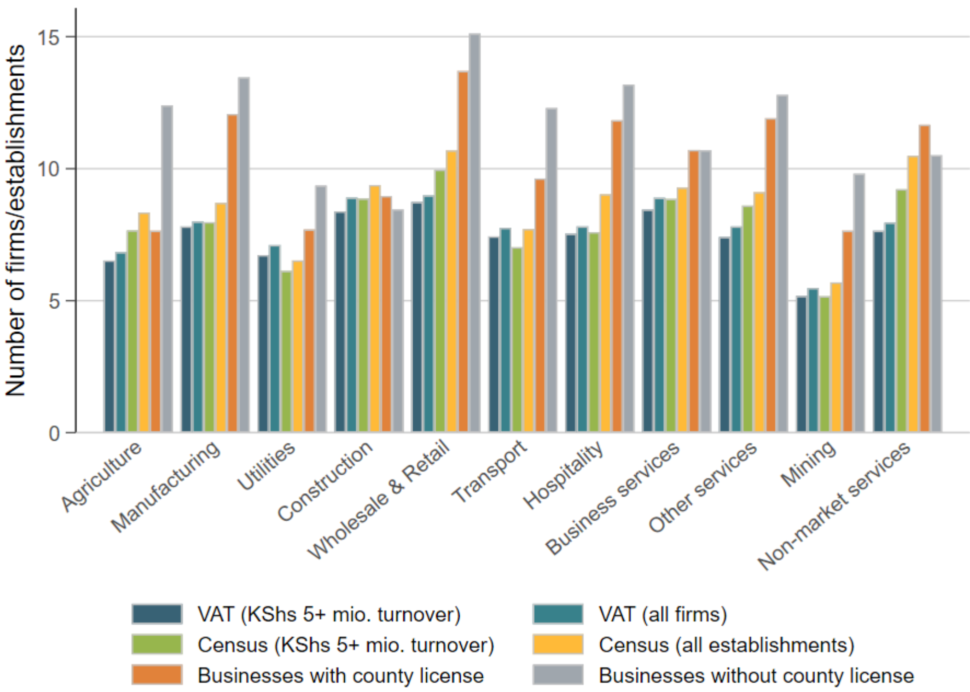
[47]We do, however, not find any other notable pattern in the data when we correlate the sector-level deviation

Figure 8: 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).

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 8). Therefore, most of the observed value added gap is a result of informality on the intensive margin, i.e. mis-reporting in VAT returns.

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: Informal firms are usually located downstream of large 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).[48]

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.[49] Figure 9 shows that the pattern holds across sectors.[50] 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. While a dual economy view on the interaction between the formal and informal sector is widespread (La Porta and Shleifer, 2014), the lines between two are often blurred. Böhme and Thiele (2014), for example, find that trade between formal

of the administrative data from firms above the VAT cut-off and the total number of firms in Kenya (see Figure B3).

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

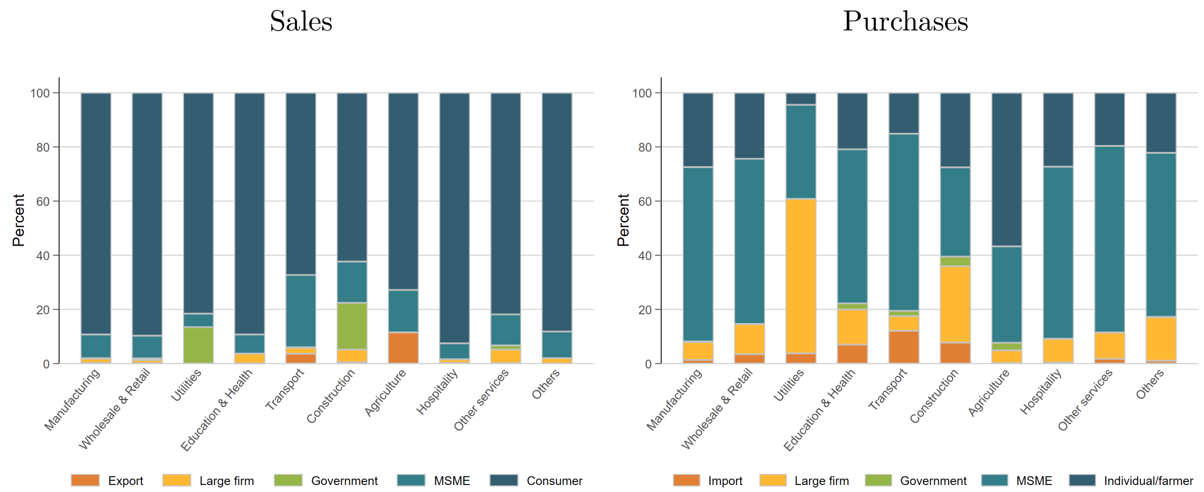
[49] KNBS (2016) defines non-MSMEs/large firms as entities with more than 99 employees. Other survey response options for inputs are: MSMEs, farmer, direct import, individual supplier, government. The corresponding options for customer types are: MSMEs, direct exports, individual consumer, government.

[50] 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.

and informal firms varies with the degree of informality. More formalised firms (unregistered firms with high captial stocks) purchase as much as 25% of their inputs from formal industrial enterprises and thus a much higher proportion than the average informal firm. The hierarchical structure of formality suggests that the vast majority of informal firms ultimately sources formal sector inputs through both indirect and direct linkages.[51] The point estimates for direct linkages between the formal and informal sector can therefore be interpreted as a lower bound on the degree of integration between the two types of firms. Nevertheless, despite large and small firms indeed interacting with each other, the VAT system itself creates wedges in firms' supply chains (De Paula and Scheinkman, 2010; Gadenne et al., 2022; Zhou, 2022). A VAT-registered retailer in a third tier town in Machakos County said:

> *"We don't buy from the local distributor as they are not filing tax returns and hence do not issue receipts. So we cannot transact with them. [..] To make it easier for us to buy from manufacturers and access trade credit, we opted for our shop to be VAT-registered."*

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 wholesale & retail sector, as well as informal firms being more likely to purchase from larger firms

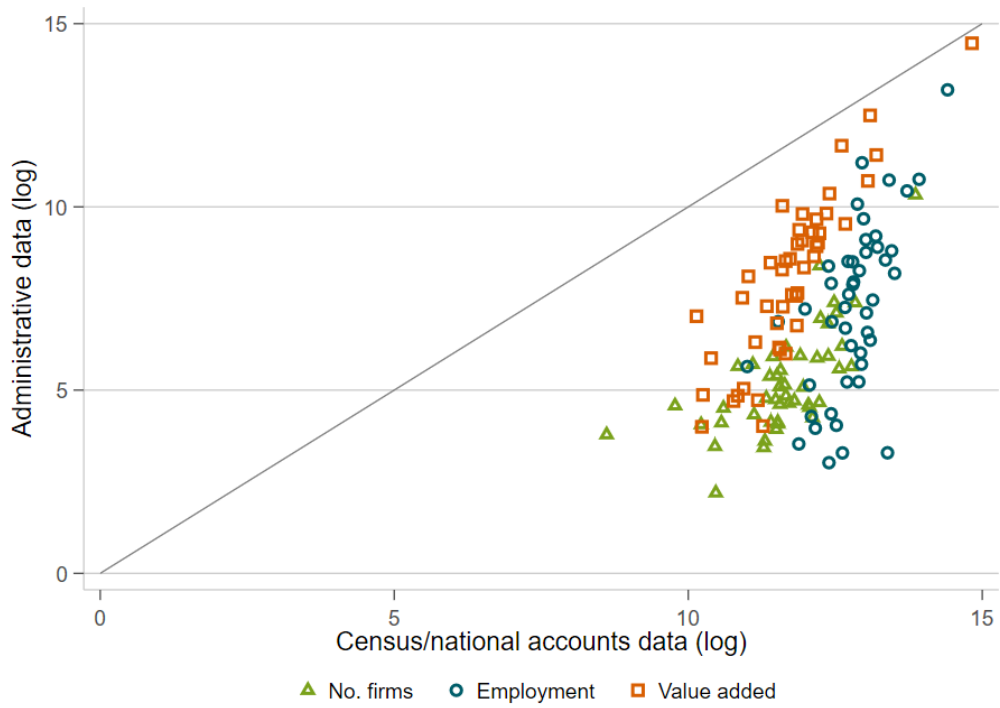
[51] Think of a small retailer selling soap. While the retail might have purchased the soap from an informal wholesaler, the wholesaler itself or its suppliers will have sourced the soap from a formal manufacturer or importer further up the chain.

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.

5.3.3 Fact 3: The VAT-sector share correlates positively with regional economic size and income levels

Moving on from sectors, where do we find informal firms in space? Here the answer is: in smaller markets. We find a strong correlation of both the overall economic size and income levels (Gross County Product per capita) with the county's formal sector share. A county's economic size explains between 35% and 52% of the variation in the the county's formal sector share (see Figure B4). We visualise the pattern in the top graph of Figure 10 where we plot the aggregate number of firms (green), value added (orange), and employment (turquoise) against their respective KNBS benchmark statistics. Each marker in the scatter plot corresponds to one of the 47 counties. In case of a 1:1 mapping the markers would lie on the 45-degree line (in blue). For cases where the KNBS benchmark exceeds the corresponding figure from the administrative data, the scatter lies below the 45-degree line and vice versa. Larger counties lie closer to the 45-degree line than smaller ones. This pattern is observed across all three of the available indicators we are able to benchmark against KNBS data: employment, value added, and the number of establishments. We note, however, that the markers for employment stay further away from the 45° line for economically larger counties. This finding is in line with the general notion previously shown in cross-country studies that informality of output declines less steeply with income levels than the share of informal workers (Kose et al., 2019). 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 11. 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 10: Count-level comparison of the number of firms, employment, and value added in KNBS benchmark data with the administrative data



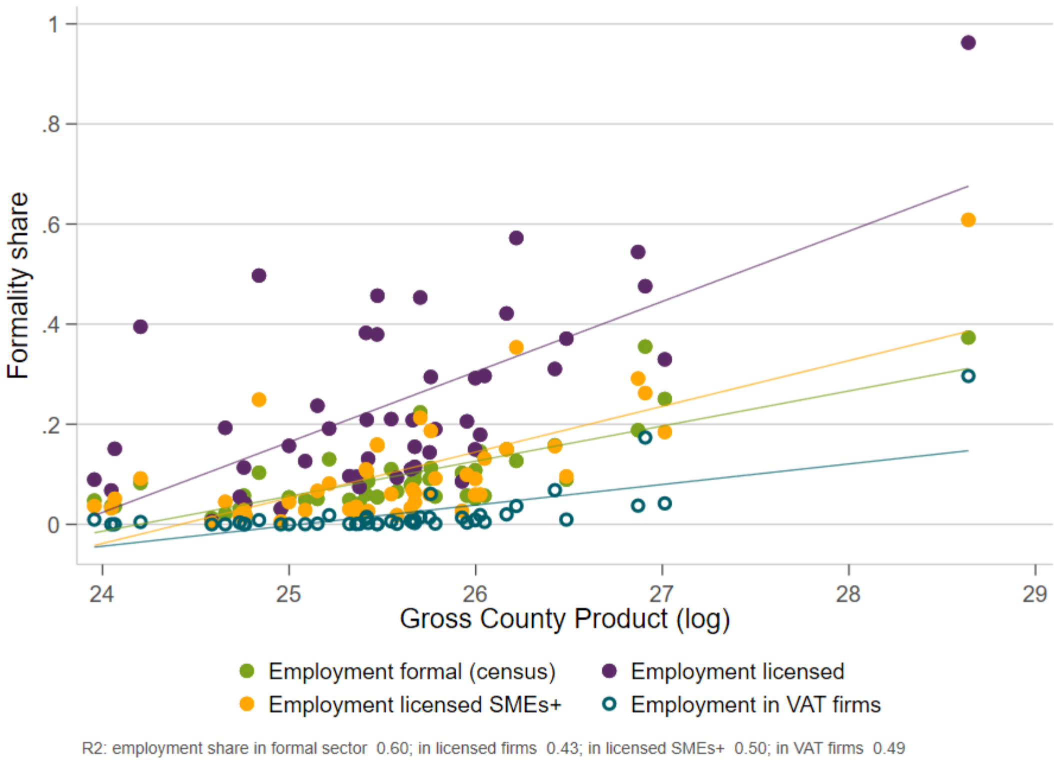
This figure plots county-level indicators derived from the administrative firm data to other county-level statistics published by the Kenya National Bureau of Statistics. Each marker represents one of Kenya's 47 counties.

5.3.4 Fact 4: 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.

Regarding the role of information frictions in spatial concentration, one interpretation of this finding is that these frictions have a greater impact on formal sector firms. Formal firms tend to be less transient than micro-firms (KNBS, 2016), and their owners have likely incurred higher fixed costs to establish their businesses. Consequently, they may seek to safeguard against the risk of exiting due to rapidly changing market conditions. In a similar spirit, studies by Kumar

Figure 11: 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 5

et al. (1999) and Laeven and Woodruff (2007) suggest a positive correlation between the quality of local institutions, particularly the judicial system, and firm size. Considering that proximity to the capital often ensures better access to the judicial system, we would expect larger firms to concentrate in Nairobi and nearby counties.

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): $\log \text{rank} = \log A + \alpha \log x$.

5.3.5 Fact 5: 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?

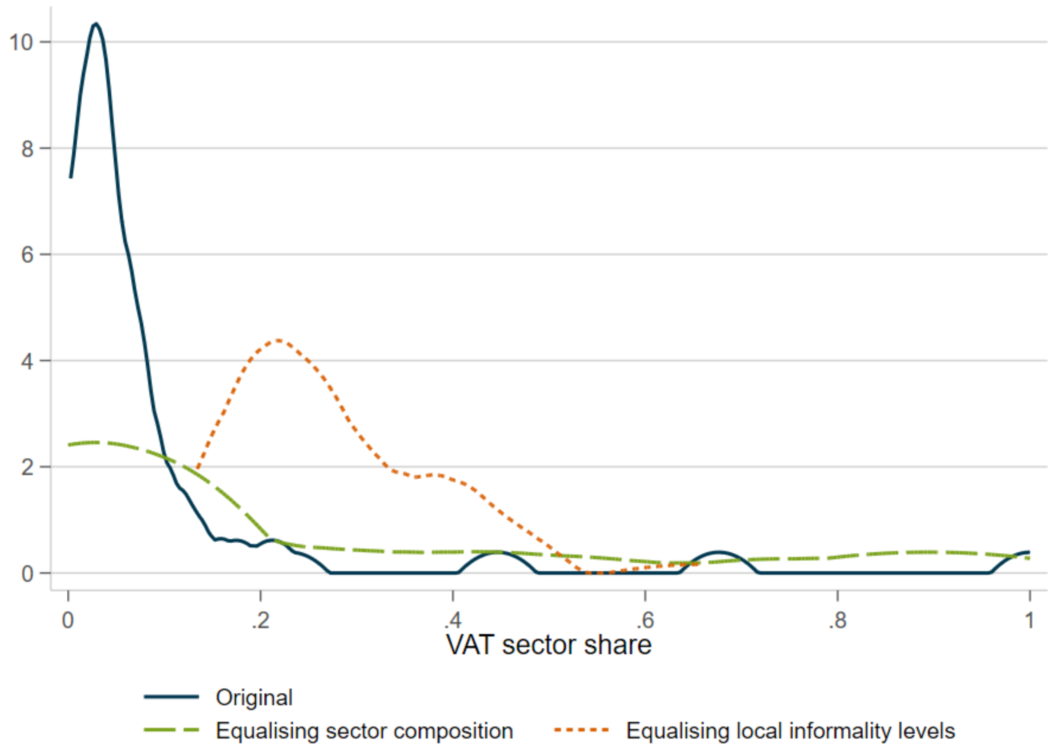
We rely on an accounting exercise with two counterfactual scenarios to study this question. We use the value-added based measure of informality for the purpose of this exercise. We start by noting that the aggregate formal sector share f_c is each county c is the sum of each sector s's degree of formality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

$$f_c = \sum s_{sc} \times f_{sc} \quad (3)$$

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 12 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.[52] 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 12 shows, the counterfactual pushes the formal sector share f_c above 20% in virtually all counties. The variance in outcomes shrinks drastically.

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

Figure 12: 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.

What have we learned from the five stylized facts? The level of informality varies drastically across space. Informal firms can be found in downstream activities, which in turn are relatively more important in smaller markets. We would therefore expect that not observing informal firms would results in us underestimating the connectedness of remote counties, especially in terms of the number of outlinks of their firms. To explore this hypothesis we rely on a network formation model that allows us to run counterfactuals.

6 A model of network formation with sectors and regions

We apply the network formation model proposed by Bramoullé et al. (2012) to our setting. The purpose of the model is to study how the spatial inequality in trade links would look like if we were able to observe informal firms. Bramoullé et al. (2012) is particularly well-suited for our purposes for three reasons: First, it allows us to easily incorporate two key dimensions of firm heterogeneity - sectors and geography. Sectors reflect the underlying input-output structure, while the geographic dimension allows us to study the question of spatial inequality. Second, the framework is explicitly tailored towards studying a network formation process of networks whose degree distribution follows a power law. In this class of models, the power law distribution is generated by so-called preferential attachment, i.e. link formation via existing links rather than via search mechanisms that do not rely on the network environment. The framework allows us to estimate the share of firm-to-firm links formed via existing network links 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 allowing 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). Lastly, we opted for a simple tractable framework to stay focused on our main goal to estimate the spatial dispersion in firm-to-firm links. We abstract from 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 pairs.[53] i.e. all firms in the same sector and county 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

[53]In the original paper by Bramoullé et al. (2012) the authors apply the framework to a setting where nodes correspond to academic papers, links to citations, and types to research fields.

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:

$$\pi^t_{t_0} = m\frac{r}{1-r}(f(t,\mathbf{B}) - \mathbf{I}) \tag{4}$$

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.[54] As a result the outdegree of existing firms, i.e. the suppliers of the newly entered firms, evolves over time. Thus, the matrix $\pi^{\mathbf{t}}_{\mathbf{t_0}}$ gives the expected outdegree (i.e. number of buyers) of each column node born in time t_0 to a row,

[54] 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.

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 and county as well as all interaction probabilities $p(\theta, \theta')$ between sector-county 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.[55]

First, we classify firms into types defined by unique sector-location combinations where the location is given by the county in which the firm is located. For example, all firms in the manufacturing sector in Nairobi 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 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'.[56] 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 4 that $\pi_{t_0}^t$ only tells us the expected outdegree of types born in t_0 evaluated at time t. Since a new firm is born in every period up until period t, we need to aggregate these matrices across all time periods leading up to t to get the type-by-type adjacency matrix of the network as a whole. The matrix of connections at time t is given by $\pi_t = \sum_{t_0}^{t} (\mathbf{p} \cdot \pi_{t0}^{t}{}')'$ 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

[55] 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. Including firms who never buy from any other firm does not align with this feature of the model.

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

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= 50897 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 $t_0 = 50879$. 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 firms with all firms observed in the data. As a result, we compute π_t= $\sum_{t_0=1:100:50897}(\mathbf{p} \cdot \pi^t_{t0}{}')'$. This implies that we will under-predict the average degree of the network as our model ignores firms born between specific time periods. At the same time, it ensures that our estimate of r is not dependent on including or excluding specific types of older or younger firms. Even if the network is scaled down in terms of number of firms, the features of the network are kept intact. Second, based on Bramoullé et al. (2012)'s formula for predicting π_t also requires us to compute a geometric series of matrix $\mathbf{B}$.[57]

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.[58] 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

[57]For ease of computation, we restrict this to the first five entries of the geometric series as the matrix has very small entries afterwards.

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

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 13).

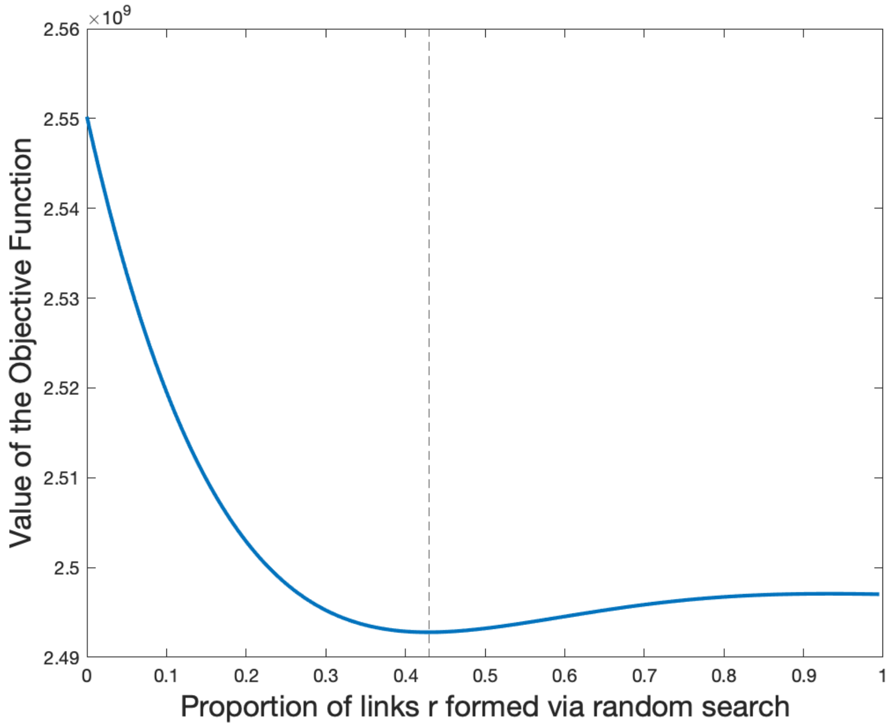
6.3 Estimation results

Our estimation strategy yields a result of $r^* = 0.42$. It suggests that a newly entered firm chooses 42% of its m suppliers randomly, and the remaining 58% among the suppliers of its existing suppliers. A network with 58% 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 also suggests a substantial, but not quite as prominent role for information asymmetries.[59]

In order to validate the estimated value of r, we plot the objective function in Figure 13 and show that it reaches a global minimum as $r = r^* = 0.42$.

[59] 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.

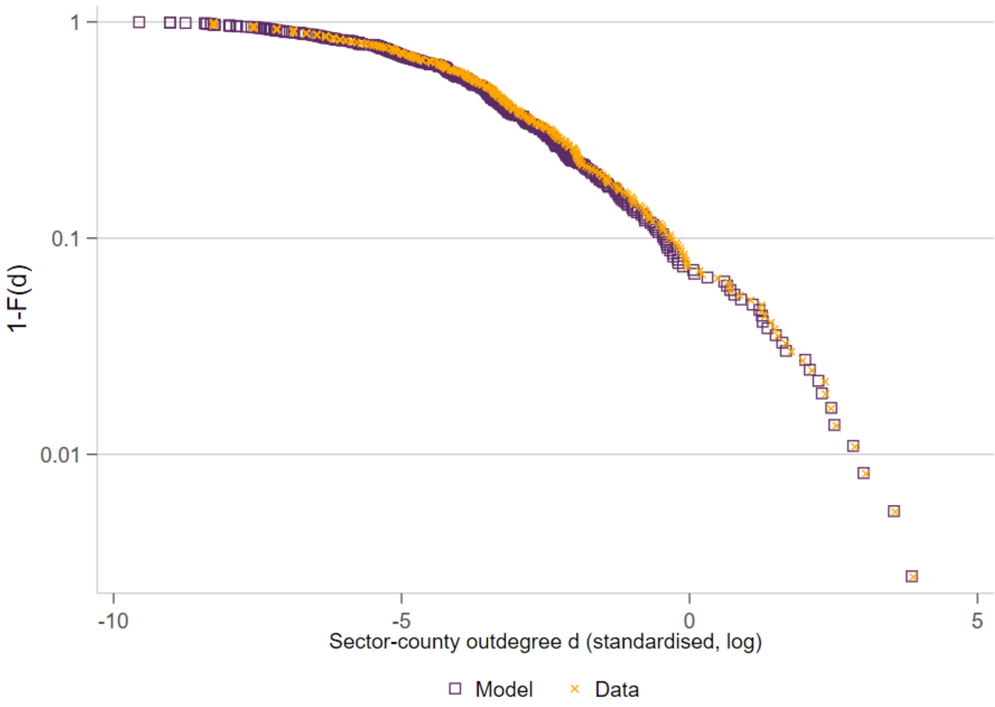
Figure 13: Objective function for various values r



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 $r^* = 0.42$ 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 14 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 Parto 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 that both coefficients are equal.

Figure 14: 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 pair). 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 - a counterfactual

With r in 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. What would happen to the outdegree distribution of various types θ? In model terms, our counterfactual shifts the probabilities $p(\theta)$ with which we observe nodes of certain sector-region types θ to be born. Against the background of the stylized facts presented in Section 5, we would expect our current estimates for $p(\theta)$ based on the administrative data to under-account for downstream firms in counties with smaller local markets. We will therefore use our population census-based measures of informality from Section 5 to update our estimates for all $p(\theta) \in \Theta$. Knowing r^* and our updated $p(\theta)$s, we can then once again predict the matrix π, keeping everything else constant.

The proposed counterfactual hinges on a strong assumption. 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 $p(\theta, \theta')$. For example, Nairobi's manufacturing sector, which has high interaction probabilities with wholesalers in Garissa county continues to do so even as a wave of new informal wholesale firms enters in

Garissa. Holding the interaction probabilities of the two sector-region types constant can be interpreted as a reflection of the economy's fundamental input-output structure and geography. Further, anecdotal evidence suggests that smaller retailers often benefit from being located in the vicinity of larger establishments who receive deliveries from larger distributors or manufacturers. At the same time, however, the argument can be made that informal firms in Garissa are unable to directly link with Nairobi-based firms thereby reducing their interaction probability. This argument would be in line with a hierarchical structure of supply chains within a sector-county cell, i.e. larger wholesalers sourcing directly from Nairobi, while selling to smaller retailers in Garissa, who in turn do not end up sourcing directly from Nairobi themselves. We will revisit the implications of keeping $p(\theta, \theta')$ unchanged when we discuss the counterfactual results.

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 pair. We will refer to the updated, alternative versions of $p(\theta)$ that incorporate the sector-county characteristics of informal firms as $p(\theta_a)$. To update $p(\theta)$, we ideally would want to observe the number of firms N_{cs} in each sector s and county c 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.[60] Therefore, instead of the firm count, we rely on the number of people who work in the private sector in each sector-county cell to compute $p(\theta_a)$:

$$p(\theta_a) = \frac{\text{No. people who work in sector s in county c}}{\sum_c^{47}\sum_s^{12}\text{No. people who work in sector s in county c}}$$

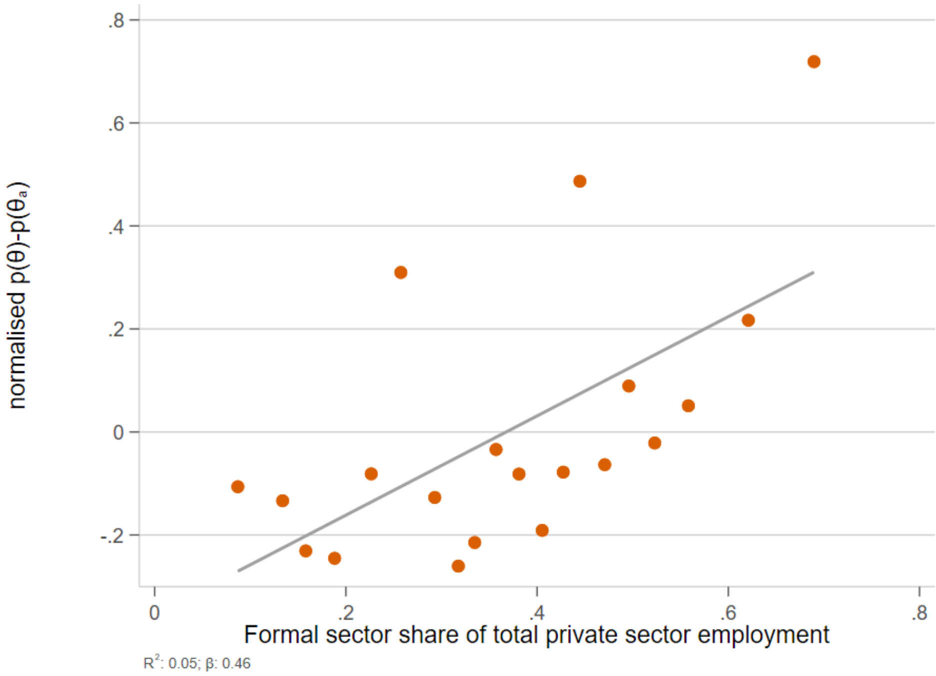
The denominator is equivalent to the total number of people who work in the private sector (employed and self-employed, formal and informal) across Kenya. For both $p(\theta)$ and $p(\theta_a)$ the sector-region probabilities sum to one. This is important to keep in mind for the interpretation of the 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 niche in both cases (see discussion in Section 5.2).[61] Amending the $p(\theta)$s of agriculture and non-market services using records on total employment (instead of formal private sector employment

[60] 12 and 47 refer to the 12 aggregate sectors and 47 counties respectively.

[61] 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.

Figure 15: 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. For a version of the same graph as a local polynomial with confidence intervals and including two alternative measures see Figure C1.

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 15, we plot the share of total private sector employees who have a formal sector job[62] against the difference between the baseline $p(\theta)$ and the augmented $p(\theta_a)$s.[63] A 10 percentage point increase in formality leads to an increase of $p(\theta)$-$p(\theta_a)$ by one percentage point (0.46 standard deviations).[64]

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

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

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

[64] To estimate the slope in Figure 15, 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.

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 two alternative approaches to compute $p(\theta_a)$, which entail re-scaling the number of employees using the average firm size in each sector-county cell. In the first alternative we use the administrative data to compute the average number of employees for each sector-county. For the second one, we use the 2016 Medium, Small and Mirco Enterprise survey to compute the average number of employees (KNBS, 2016).[65] Given both measures of firm size by employment do not cover the entire firm size distribution (the survey covers the left tail, while the administrative data cover the right tail), we rely on the plain-vanilla employment shares as the default measure and consider the other two robustness checks. We plot the default measure and the two alternative measures 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.[66]

7.2 Counterfactual results

We find that accounting for informal firms increases the variation in outdegrees across counties by 21% (see Table 8). Meanwhile, the standard deviation to mean ratio in outdegrees across all sector-county types increases by 17%. The dispersion of outdegrees across sectors on the other hand falls by 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). Rather than thinking of the counterfactual as adding new firms, we adjust the weights of each sector-region type.

Table 8: Changes in the dispersion of outdegrees

Aggregation	Δ sd/mean (in %)
County outdegree	20.8
Sector outdegree	-6.9
Sector-county outdegree	17.1

The above table reports the result from our baseline counterfactual that adjusts entry probabilities of nodes in sector-county 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 two different proxies for the firm size distribution are reported in Table C1.

The increase in inequality of outlinks across counties is driven by sectors and counties that the newly accounted for firms are more likely to link to. Nairobi and Mombasa-based firm

[65] 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.

[66] The only notable difference is that correcting for firm size by relying on the MSME survey yields significantly higher entry probabilities in sector-county cells with the lowest formal sector share.

types are 12 percentage points and 4 percentage points more likely to supply to firm types whose augmented entry probabilities are higher than before. Firm types in almost all remaining counties have a lower connection probability to these previously unaccounted for firms. We re-compute the county-level variance in outlinks by excluding Nairobi and Mombasa in Table C1 as these counties have the highest average interaction probabilities. We find an increase in the variance to mean ratio of only 8% rather than 21% across counties.

The aggregate patterns of course mask substantial heterogeneity across counties. Counties with the largest percentage increment in outlinks are, in decending order, Nairobi, West Pokot and Garissa. Garissa and West Pokot are counties with low levels of formality. Garissa is a geographical gateway for some of Kenya's northern counties, the region with the highest concentration of counties with a low GDP per capita and high levels of informality. Overall, the top ten[67] counties in terms of gains are counties with either high levels of informality (Marsabit, Elgeyo Marakwet, West Pokot, Mandera, and Isiolo) or because they serve as a service or distribution hub either nationally or for their respective regions (Nairobi, Garissa, Mombasa, Uasin Gishu, and Kisumu). All else constant, ignoring informal firms underestimates the importance of the ex ante well-connected counties as hubs and the exposure of the ex ante less-connected counties to the rest of the economy. This pattern is reflected by the U-shape correlation between the change in the number of outlinks and a county's formal sector share shown in Figure 16.[68] The U-shape disappears (see Figure C2) once we control for the probabilities with which new firms want to link withe the county across all county-sector types ($\sum_i p(\theta, \theta')$). In Figure C3 we plot the change in outdegrees against the interaction probabilities themselves. The steep slope for counties popular with previously under-counted buyer types mirrors the upward slope in the U-shape graph, i.e. the rise in outdegrees for counties with a high formal sector share.

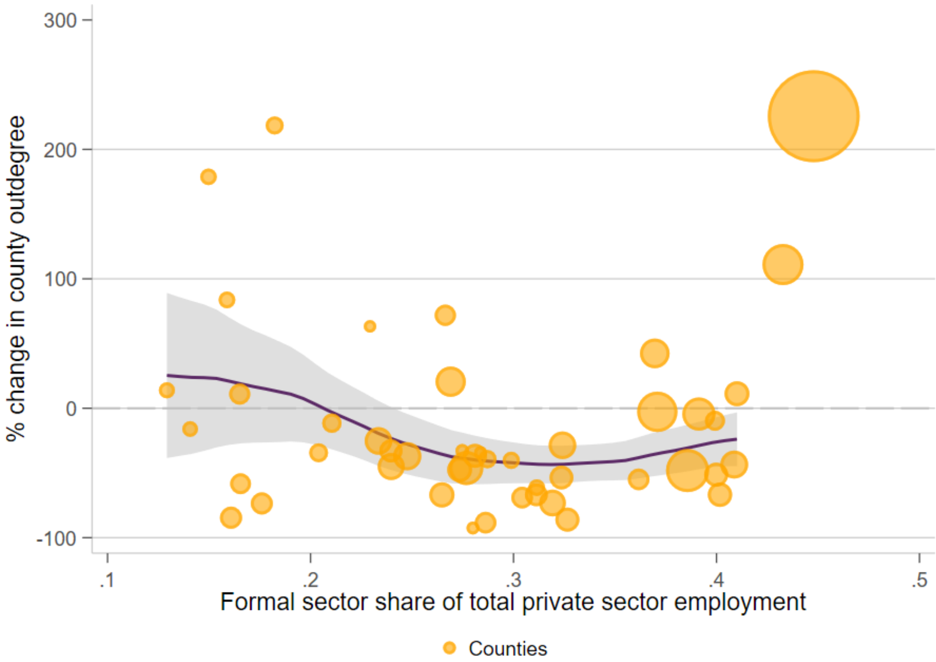
At the sector level, business services (which includes ICT), construction, mining,[69] and utilities sectors gain the most links in relative terms. Note our earlier observation in Figure 3 of Section 4, which showed that business services and utilities are the sectors with the highest degree of spatial concentration. Their gains therefore align with the increase in outdegrees of counties that act as hubs in the network and host a disproportionate number of firms from these sectors. The result for the construction sector can be explained by the fact that construction sector firms have particularly strong linkages to or within counties with otherwise high levels of informality. Formalisation of construction sector links is often incentivised by the participation of the gov-

[67]The top ten are: Nairobi, West Pokot, Garissa, Mombasa, Marsabit, Elgeyo Marakwet, Isiolo, Uasin Gishu, Kisumu, and Mandera.

[68]The U-shape pattern is also present if we look at the change in outlinks across sector-region types rather than aggregating outlinks to the county level in Figure C2. Moreover, it is less pronounced if we exclude Nairobi and Mombasa, but there is still a noticeable rise due to gains in outlinks by other counties that serve as hubs such as Kisumu, Machakos and Nakuru.

[69]The observed relative increase in links for the mining sector is influenced by small numbers since the administrative data include only a few hundred mining firms. The population census adds small-scale mining activities.

Figure 16: Predicted change in county-level outdegree and formal sector shares



Each marker represents a county. The size of each marker is proportional to the respective county's Gross County Product. The local polynomial is, however, an unweighted version treating each county equally.

ernment in construction projects. The wholesale and retail sector, i.e. the sector that gained the largest number of new firms in the counterfactual exercise, is the median sector in terms of its increase in outdegrees.

Our suggested counterfactual accounts for informal firms being born into the network based on their sector-region profile. At the same time, our exercise keeps linking probabilities between sector-region types $p(\theta, \theta')$ constant. We thereby treat $p(\theta, \theta')$ as a fundamental production technology. The U-shaped response to changes in outdegrees suggests that spatial inequality actually increases if we account for informal firms and that linking probabilities play a key role in shaping spatial inequalities in firm-to-firm links across regions. If a policy goal was more regional economic integration, examples of policies that might shift these linking probabilities include the expansion of communication infrastructure like the fiber optic network, the strengthening of transportation infrastructure between secondary cities and towns or the introduction of a physical address system that reduces the cost of shipping goods. The contextual relevance of our result is highlighted by the fact that all counties among those that gain relatively more links and feature a lower ex ante formal sector share, are currently targeted by a 100,000 km extension of Kenya's fiber optic network.[70] At the same time our set-up is not able to speak

[70] The programme targeting relatively underserved counties is part of Kenya's 2022 Digital Master Plan (Ministry of ICT, Innovation and Youth Affairs, 2022). https://www.connectingafrica.com/author.asp?section_id=816&doc_id=784428

to the optimal spatial allocation from an aggregate welfare point of view. Given the current market frictions, agglomeration and a strong hub-based network could indeed be optimal for private sector development.

The counterfactual exercise brings us closer towards the question of the firm network's structure if we were to observe informal firms, but it abstracts from the up- and downstream dynamic of supply chains within sectors, in particular transportation, wholesale and retail. To give an example, a formal wholesaler in West Pokot County might be more likely to source directly from a manufacturer in Nairobi than a smaller, informal firm. Using the formal wholesaler's linking probabilities in the counterfactual might therefore overpredict how many more links between West Pokot and Nairobi we are missing out on. In reality, Nairobi might obtain relatively fewer links. At the same time, we are likely to underestimate downstream links to other businesses rather than consumers as we rely on observed linking probabilities of formal firms and cannot distinguish between their sales to final consumers and sales to non-VAT-registered businesses. We might therefore underpredict the number of outlinks West Pokot itself actually gains.
While we advise against taking our stylized model too literally, it serves its main purpose - to test our intuition on the role of the informal sector based on the stylized facts presented in the empirical part of the paper. Our initial intuition that the spatial inequality in network links would reduce if one were to observe informal firms, potentially requires an additional assumption about the manner in which informal firms form links with the formal sector. The most plausible is that informal firms mostly link with other local firms due to their size.

8 Conclusion

Memon (1976) discusses how the debate on the economic dominance of the metropolitan areas in Kenya was as much of a concern, but to some extent also an accepted status quo, back in the 1970s as it is today:

> *"A recent document on the country's urban development strategy, for instance, rationalises the dominant position of Nairobi and Mombasa in relation to other towns as a demonstration of "the role primate cities play in a transitory economy" (MoF, 1974, as cited in Memon (1976)). Acceptance of urban primacy as a necessary evil is also reflected in the "large-city bias" of the policies governing resource allocation for urban infrastructure."*

He goes on to show that the same policy report by the Ministry of Finance outlines the following urban planning goals:

> *"... to encourage the expansion of several large towns in addition to Nairobi and Mombasa in order to promote a geographical dispersal of benefits arising from urban development amongst a number of centres." (MoF, 1974, as cited in Memon (1976))*

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. At the same time, counties that serve as trading hubs also gain. This is because informal firms are particularly prevalent in sectors and regions that ex ante have a high probability to link with those hubs. This result aligns with Blanchard et al. (2021)'s finding that people who travel from more rural to more urbanised areas in Kenya favour visits to Nairobi and Mombasa over secondary cities and towns. They are able to rationalise this pattern with a model where people place a premium on variety in amenities, which in turn tends to be higher in cities. In the firm context this can translate to a higher density of potential trading partners, a greater variety in available inputs, and access to complementary services like transportation. In a nutshell, our counterfactual results suggest that data on the formal firm network likely underestimate the connectivity and thereby also vulnerability of smaller regions to shocks which pass through hubs such as Kenya's capital Nairobi.

A limitation of our counterfactual is that we are only able to adjust entry probabilities rather than linking probabilities between different sector-county pairs. While the model would allow us to implement such a counterfactual, we lack information on how these probabilities change with the addition of informal firms. Future extensions of this work could use shocks to formalisation to estimate the linking probabilities of marginally formal firms.

References

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

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

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

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.

Allen, T. (2014), 'Information frictions in trade', *Econometrica* **82**(6), 2041–2083.

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.

Alonso, W. (1968), 'Urban and regional imbalances in economic development', *Economic development and cultural change* **17**(1), 1–14.

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.

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

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

Baqaee, D. R. and Farhi, E. (2019), 'The macroeconomic impact of microeconomic shocks: Beyond hulten's theorem', *Econometrica* **87**(4), 1155–1203.

Bergquist, L. F., Faber, B., Fally, T., Hoelzlein, M., Miguel, E. and Rodríguez-Clare, A. (2022), Scaling agricultural policy interventions, Technical report, National Bureau of Economic Research.

Bergquist, L. F., McIntosh, C. and Startz, M. (2021), 'Search cost, intermediation, and trade: Experimental evidence from ugandan agricultural markets', *eScholarship, University of California* .

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

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

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

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*

Berry, B. J. (1961), 'City size distributions and economic development', *Economic development and cultural change* **9**(4, Part 1), 573–588.

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

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

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).

CAK (2019), Competition in shipping, trucking and haulage sector study in east africa, Technical report, Competition Authority Kenya.

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

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.

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.

Coşar, A. K. and Demir, B. (2018), 'Shipping inside the box: Containerization and trade', *Journal of International Economics* **114**, 331–345.

Coşar, A. K., Demir, B., Ghose, D. and Young, N. (2022), 'Road capacity, domestic trade and regional outcomes', *Journal of Economic Geography* **22**(5), 901–929.

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., Kikkawa, A. K., Mogstad, M. and Tintelnot, F. (2021), 'Trade and domestic production networks', *The Review of Economic Studies* **88**(2), 643–668.

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

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

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

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

Erol, S. and Vohra, R. (2022), 'Network formation and systemic risk', *European Economic Review* **148**, 104213.

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.

Grassi, B. et al. (2017), 'Io in io: Size, industrial organization, and the input-output network make a firm structurally important', *Work. Pap., Bocconi Univ., Milan, Italy* .

Grigoli, F., Luttini, E. and Sandri, D. (2023), 'Idiosyncratic shocks and aggregate fluctuations in an emerging market', *Journal of Development Economics* **160**, 102949.

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**.

Head, K. and Mayer, T. (2014), Gravity equations: Workhorse, toolkit, and cookbook, *in* 'Handbook of international economics', Vol. 4, Elsevier, pp. 131–195.

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* .

Ishikawa, J. and Tarui, N. (2018), 'Backfiring with backhaul problems: Trade and industrial policies with endogenous transport costs', *Journal of International Economics* **111**, 81–98.

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.

King, M., Tarbush, B. and Teytelboym, A. (2019), 'Targeted carbon tax reforms', *European Economic Review* **119**, 526–547.

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.

Lane, N. (2023), 'Manufacturing revolutions - industrial policy and industrialization in south korea', *Quarterly Journal of Economics* .

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.

McMillan, J. and Woodruff, C. (1999), 'Interfirm relationships and informal credit in vietnam', *The Quarterly Journal of Economics* **114**(4), 1285–1320.

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

Ministry of ICT, Innovation and Youth Affairs (2022), *The Kenya National Digital Master Plan 2022-2032*.

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.

MoF (1974), Urban population projections during 1969-2000, within the context of urban development strategy, Technical report, Republic of Kenya, Ministry of Finance and Planning.

Murthy, S. V. R. (2019), 'Measuring the informal economy in india'.
URL: *https://www.imf.org/-/media/Files/Conferences/2019/7th-statistics-forum/session-ii-murthy.ashx*

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

Newman, M. (2018), *Networks*, Oxford university press.

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.

Otiso, K. M. (2005), 'Kenya's secondary cities growth strategy at a crossroads: Which way forward?', *GeoJournal* **62**, 117–128.

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

Pred, A. (1980), *Urban growth and city systems in the United States, 1840-1860*, Vol. 11, Harvard University Press.

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.

Venables, A. J. (1996), 'Equilibrium locations of vertically linked industries', *International Economic Review* pp. 341–359.

Wong, W. F. (2022), 'The round trip effect: Endogenous transport costs and international trade', *American Economic Journal: Applied Economics* **14**(4), 127–66.

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.

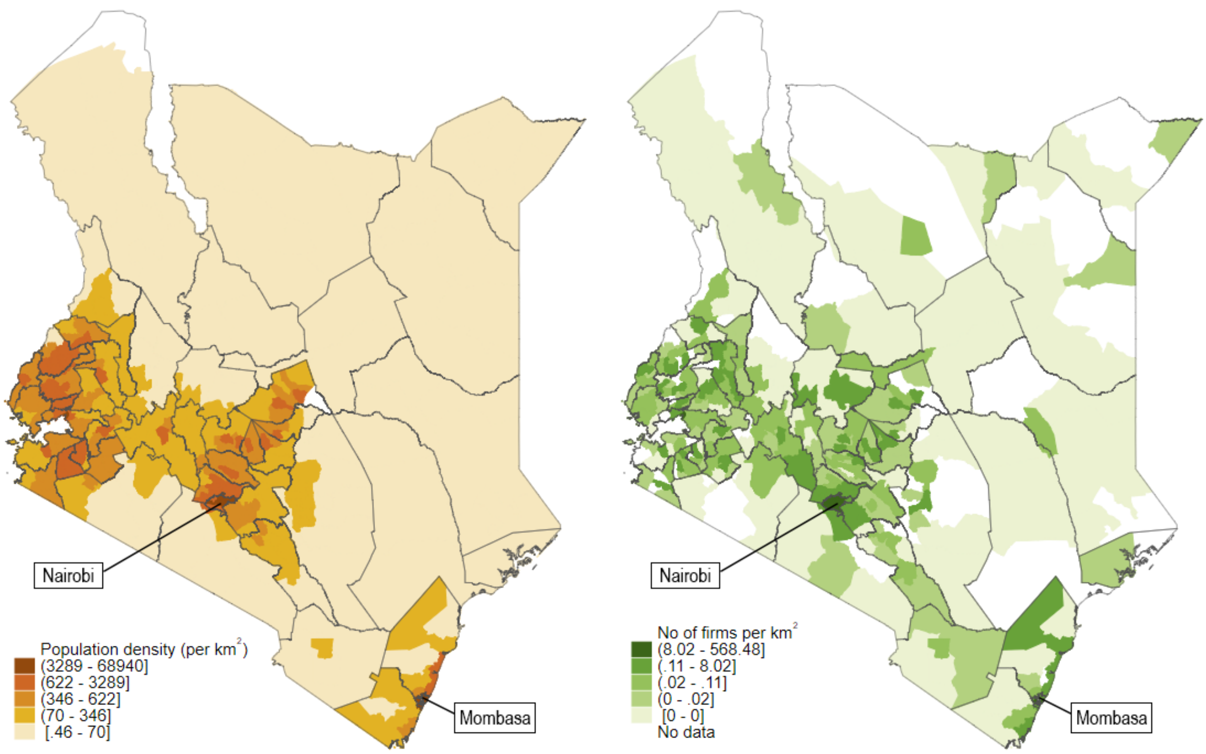
Appendices

Appendix A Complementary material for Section 4 on spatial trade patterns

A.1 Additional graphs and tables

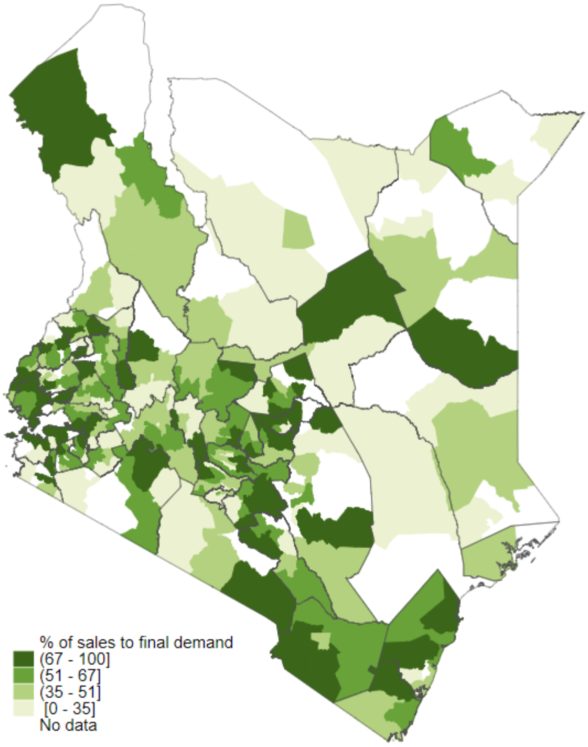
Figure A1: Firm headquarter locations and population density

Population density | Geographic density of firms



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

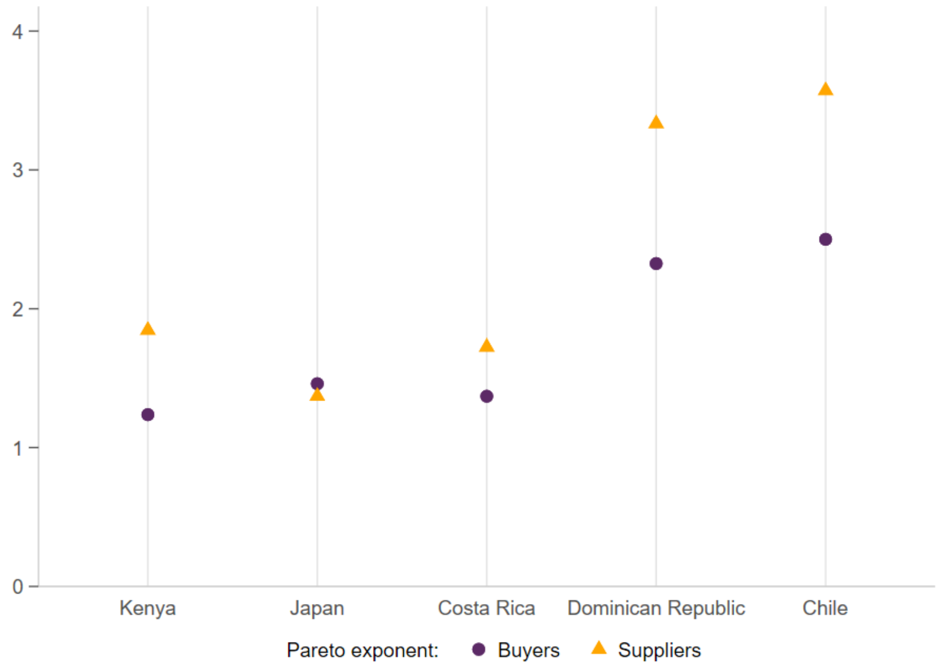
	Total trade volume		# of links		Avg # of transactions per link		Avg volume per transaction	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Travel time	-2.353***	-1.137***	-1.906***	-1.012***	-0.448***	-0.176***	0.001	0.051
	(0.101)	(0.138)	(0.050)	(0.056)	(0.035)	(0.051)	(0.053)	(0.082)
Social connectedness		1.042***		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.

Figure A3: The higher inequality in outdegree is common across contexts



The figure plots the pareto exponents for Kenya (this paper), Japan (Bernard et al., 2019), Costa Rica (Alfaro-Urena et al., 2018), Chile (Grigoli et al., 2023), and the Dominican Republic (Cardoza et al., 2023). The latter three papers estimate and report the inverse of the pareto coefficient, which we invert here for the purpose of comparability.

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[71] 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

[71] 2015 is the earliest year for which the VAT records have been fully digitised.

manufacturing firms reported by Storeygard (2016) for Tanzania.[72] 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).

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

Figure B1 and Table B1 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.[73] 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.

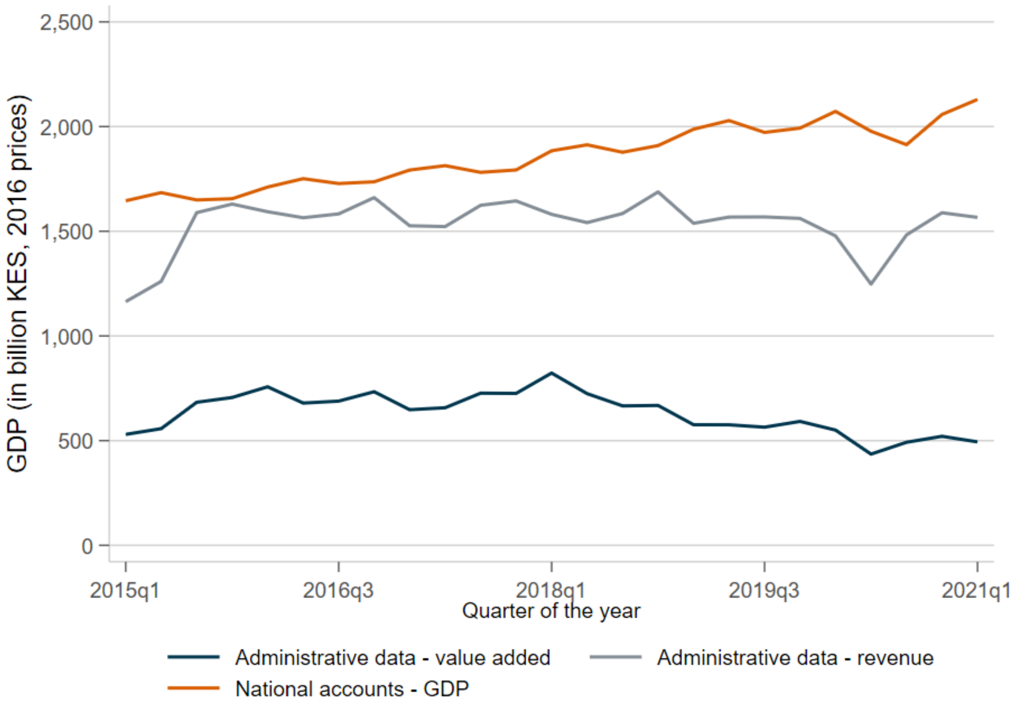
Table B1: Share of GDP covered in the administrative records

	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/.

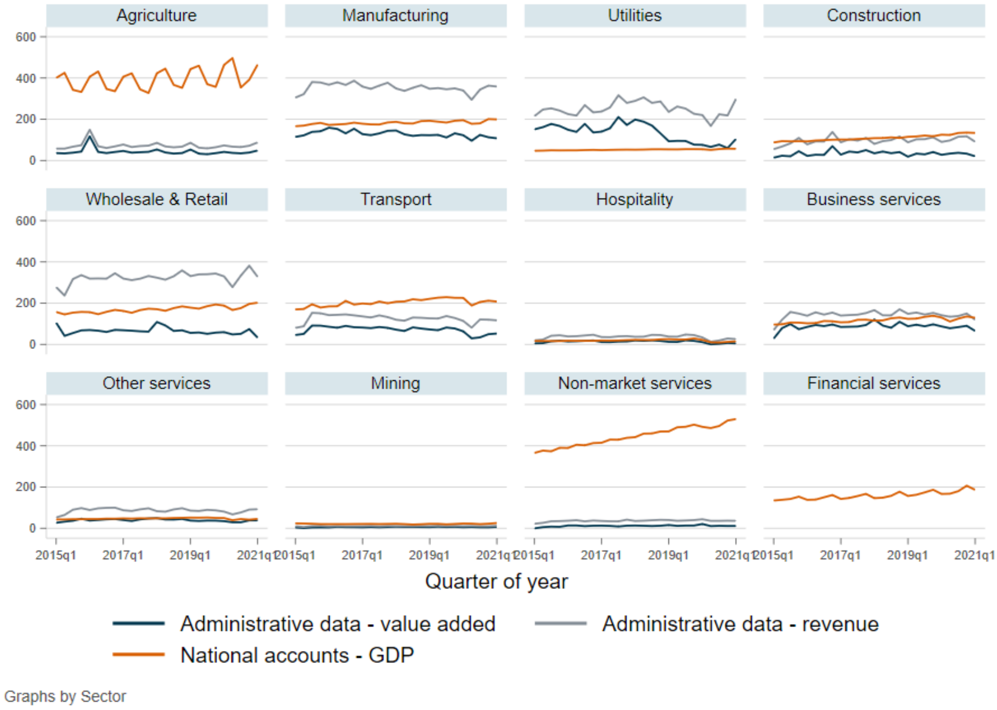
[73] The impact of this tax is particularly evident in the plot for utilities (electricity, gas, and water) in Figure B2. However, this sector alone cannot fully explain the overall downward trend and kink in the data.

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)



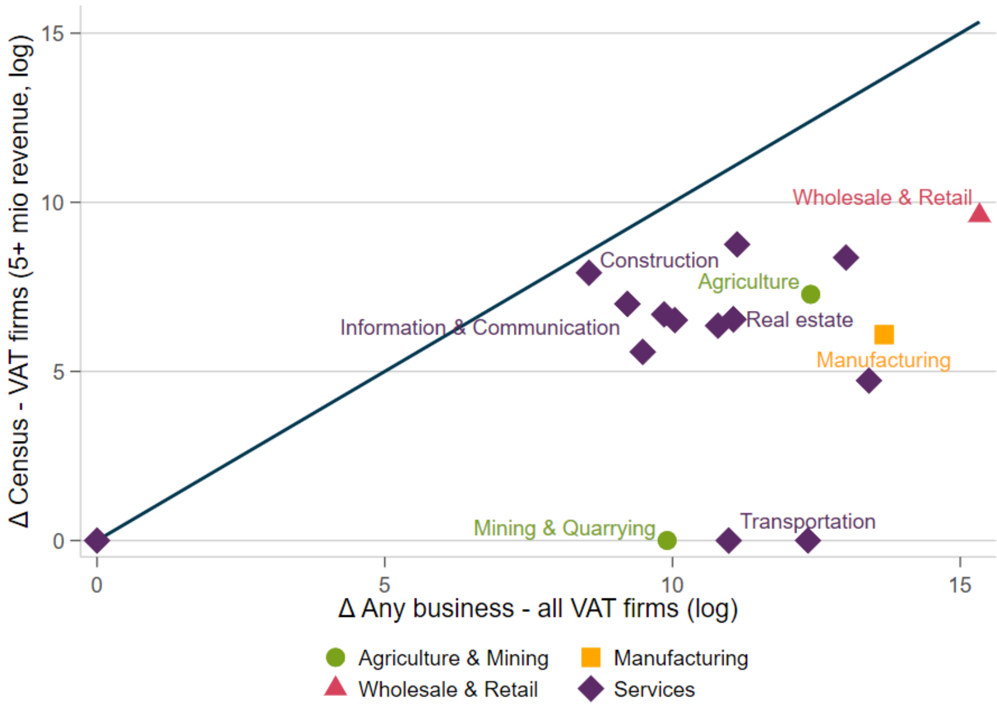
This figure compares the quarterly sector-level GDP in 2016 prices to the quarterly value added (domestic sales + exports - domestic purchases - imports - salaries) from the administrative data (deflated).

Table B2: 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



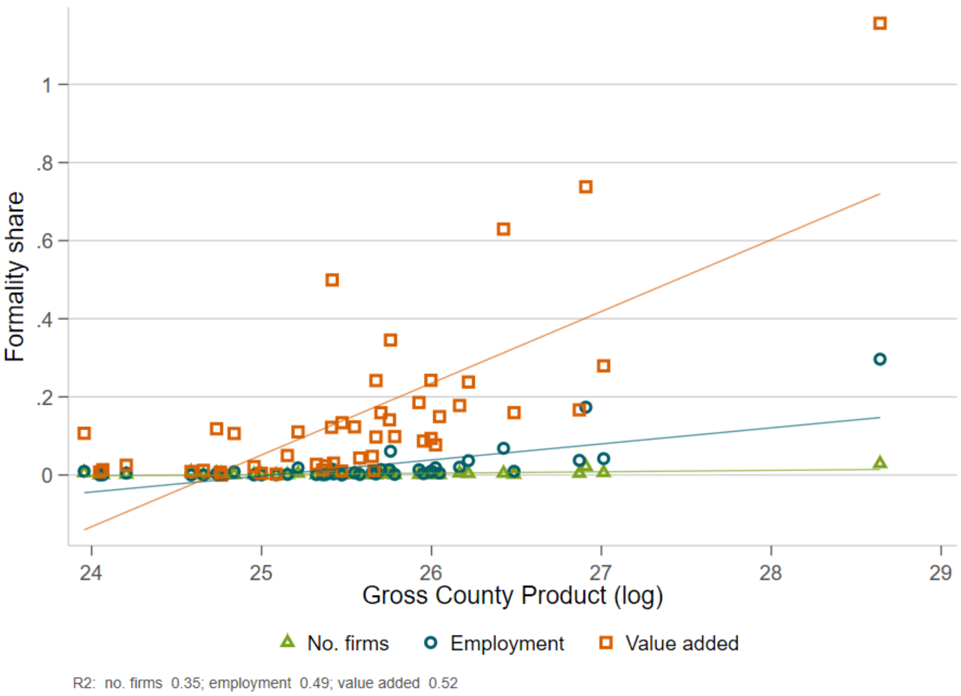
The above graph correlates the two types of extensive margin informality at the sector level. On the y-axis we put the gap between the number of firms with more than KShs 5 million of revenue from the Census of Establishments (KNBS, 2017) and the administrative data. The x-axis shows the gap between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

B.2 Informality, market size, and income levels

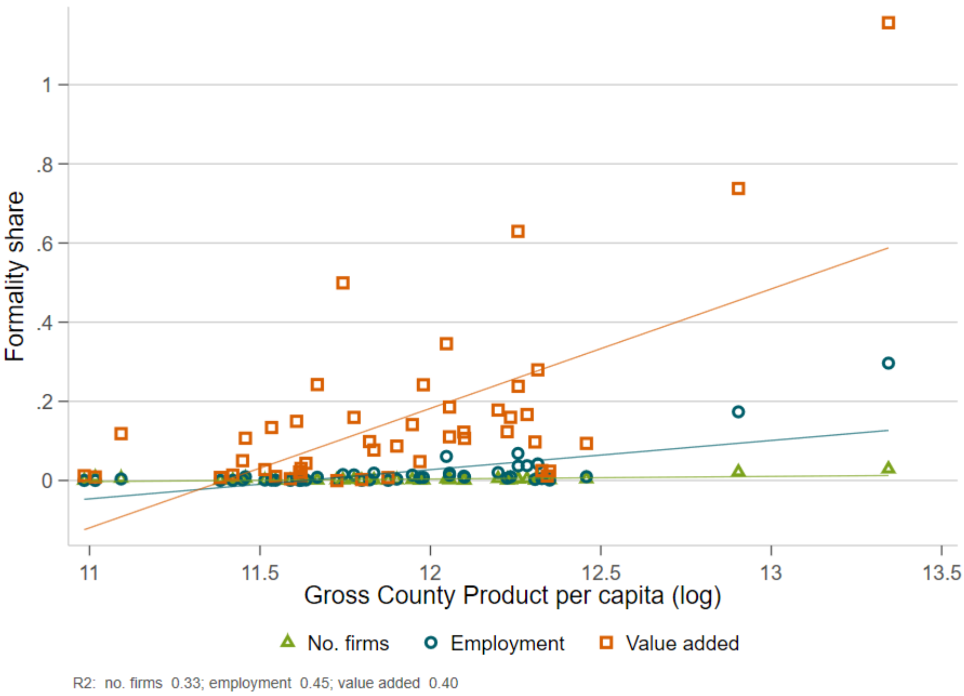
Figure B4: Informality, market size, and income levels

Correlation of the formal sector share and ...

... Gross County Product



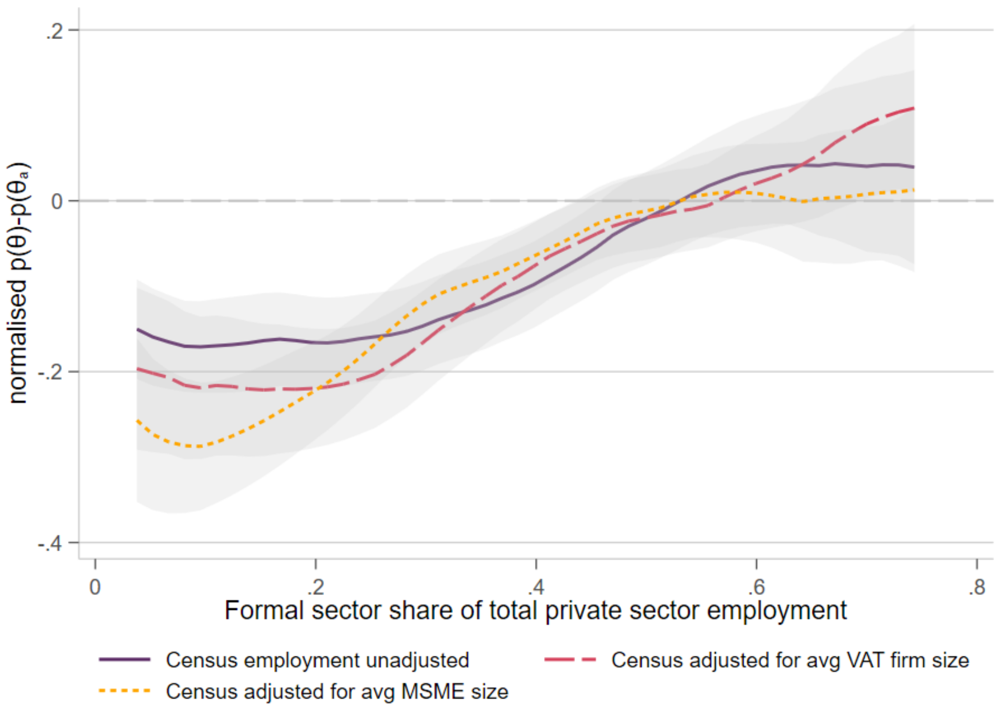
... Gross County Product per capita



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.

Appendix C Complementary material for Section 7 on the counterfactual

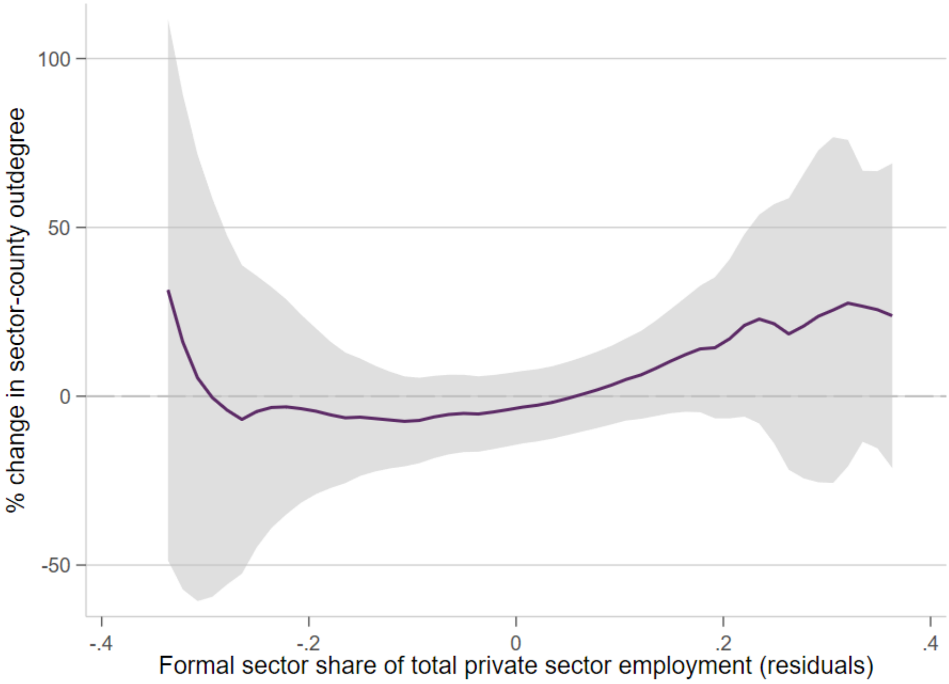
Figure C1: Sector-region probabilities and formal sector shares



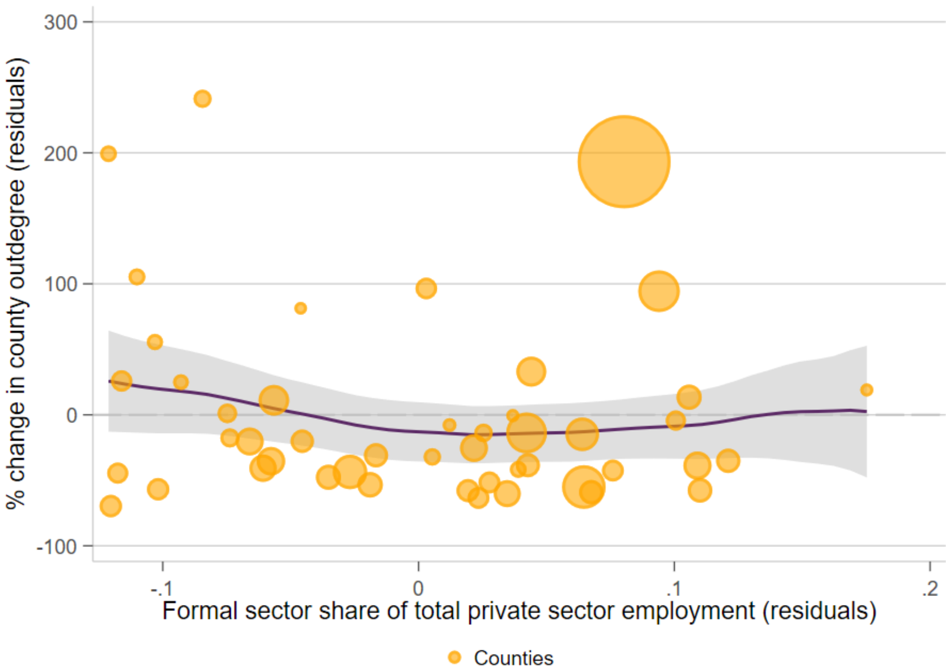
The graph plots each sector-regions formality share against the normalised difference between the baseline $p(\theta)$ and the augmented version $p(\theta_a)$ that takes into account informal firms. $p(\theta)$-$p(\theta_a)$ is reported in terms of standard deviations. In addition to the default $p(\theta_a)$, the graph plots two additional versions, which account for differences in the firm size distribution across sector-county cells drawing on the average firm size based on the administrative data and a survey of medium, small and micro enterprises (KNBS, 2016) respectively.

Figure C2: Predicted change in outdegree and formal sector shares

By sector-county type

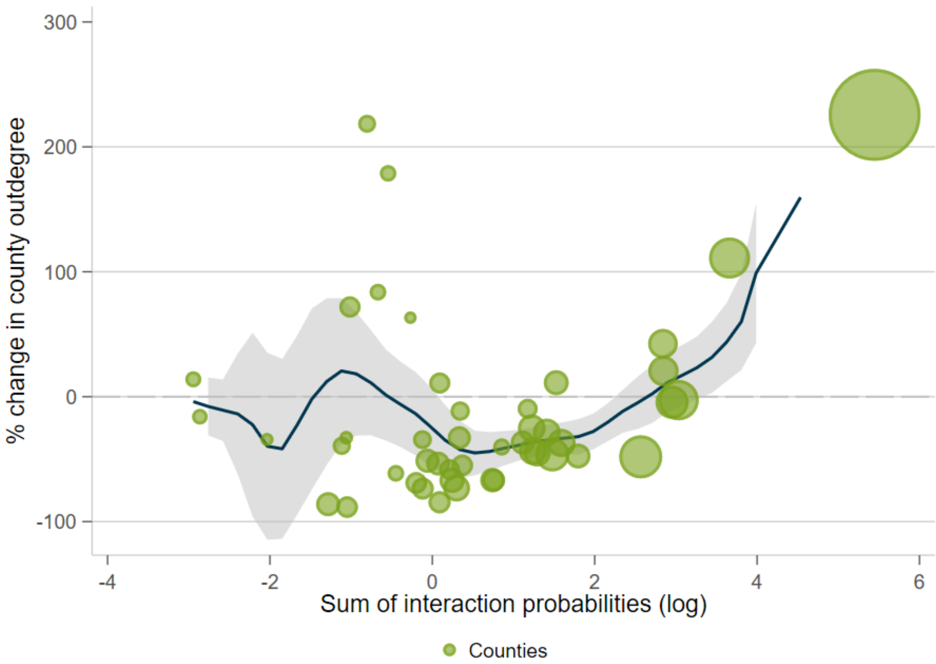


Controlling for baseline linking probabilities ($\sum_i p(\theta, \theta')$)



We plot the percentage change in outdegrees at the sector-county level in the top graph and at the county level in the bottom graph. Relative to Figure 16 we control for the linking probabilities of a county at baseline in the bottom graph, i.e. the aggregate probability sector-county types link with the county ($\sum_i p(\theta, \theta')$). Each marker represents a county. It's size reflects the county's Gross County Product. The local polynomial is unweighted.

Figure C3: Predicted change in county-level outdegree and baseline linking probabilities $(\sum_i p(\theta, \theta'))$



Each marker represents a county. Its size reflects the county's Gross County Product. The local polynomial is unweighted.

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

County outdegree	Δ sd/mean (in %)
Baseline	20.8
Baseline without Nairobi & Mombasa	7.7
Adjust for firm size (admin data)	25.5
Adjust for firm size (MSME survey)	18.5

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 and four report the results for alternative distributions of $p(\theta)$ that correct for differences in the firm size distribution across sector-county types.