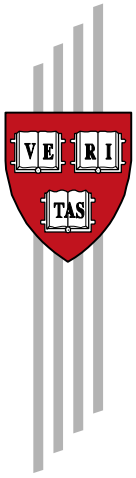
The Path to Labor Formality: Urban Agglomeration and the Emergence of Complex Industries

Neave O'Clery, Andres Gomez-Lievano and Eduardo Lora

CID Research Fellow and Graduate Student
Working Paper No. 78
October 2016



Working Papers
Center for International Development at Harvard University

The Path to Labor Formality: Urban Agglomeration and the Emergence of Complex Industries

Neave O'Clery[1,2,3], Andres Gomez-Lievano[1], and Eduardo Lora[1]

[1]Center for International Development, Harvard University - 79 JFK street, 02138 Cambridge MA, USA
[2]Present address: Mathematical Institute, Oxford University - Andrew Wiles Building, Radcliff Observatory Quarter, OX2 6GG Oxford, UK
[3]Corresponding author: neave_oclery@hks.harvard.edu

Labor informality, associated with low productivity and lack of access to social security services, dogs developing countries around the world. Rates of labor (in)formality, however, vary widely within countries. This paper presents a new stylized fact, namely the systematic positive relationship between the rate of labor formality and the working age population in cities. We hypothesize that this phenomenon occurs through the emergence of complex economic activities: as cities become larger, labor is allocated into increasingly complex industries as firms combine complementary capabilities derived from a more diverse pool of workers. Using data from Colombia, we use a network-based model to show that the technological proximity (derived from worker transitions between industry pairs) of current industries in a city to potential new complex industries governs the growth of the formal sector in the city. The mechanism proposed has robust strong predictive power, and fares better than alternative explanations of (in)formality.

Key words: labor formality, complexity, city size, diversification, networks.
JEL codes: B5, D8, J2, J4, O1 and R1

1 Introduction

Following the Second World War, the developed world experienced a rapid industrial expansion. It was expected that developing countries would follow this expansion, and traditional economies dominated by low renumeration jobs and low productivity would give way to modern capitalist production processes (Lewis, 1954). The result would be increased productivity and rising wages. These expectations, however, were not met, and the persistence of what anthropologist Keith Hart originally labeled as the 'informal

sector' (Hart, 1973), became a central topic of study and analysis in the development literature (Castells and Portes, 1989, De Soto, 1990, 2000, Doeringer and Piore, 1975, Maloney, 2004, Singer, 1973) (see Chen, 2012 for a review). Today, labor informality remains a major concern in developing countries, and is associated with low productivity (Busso et al., 2012, La Porta and Shleifer, 2014), tax evasion and reduced access to contributory social security programs (Perry, 2007).

Although central to the early economic development literature inspired by the pioneering work of (Fields, 1975) (see also Ghani and Kanbur, 2013), "the link between urbanization and formal employment is a part of the development process that is potentially important and too little studied", according to Paul Romer[1]. The role that city features such as population size and industrial composition play in the creation of formal employment (and the reduction of informality) in developing countries has been largely ignored. Variables such as minimum wages, payroll taxes and government capabilities to enforce regulations and provide public goods, which are considered important determinants of informality (La Porta and Shleifer, 2014, Levy, 2010), show little or no variation within countries. Yet, as Figure 1 shows, the variance of formality rates across cities within countries is larger than that across countries.[2] The formality rate is defined in this paper as the share of working age population that is employed by firms in observance of the legal mandates with respect to wages, social security contributions and payroll taxes. The definition of formal workers as those covered by regulations on salaried labor is standard in the literature (Busso et al., 2012, Levy, 2010).[3]

The role that urbanization processes play in the reduction of informality is clearly suggested by the fact that labor formality rates vary systematically with city size or, more precisely, with the number of potential workers as measured by the size of the working age population, see Figure 2.

Our main hypothesis is that larger cities are able to absorb a larger share of the potential labor force because increasingly complex industries take advantage of the complementary capabilities derived from a larger and more diverse pool of workers. While our hypothesis rests on the role of individual skills, it goes beyond those of managers, which have been emphasized by Lucas (1978) and La Porta and Shleifer (2014). While our approach does not deny that government interventions such as minimum wages, taxes and regulations may hinder the expansion of formal employment, it does not see informality as a result of such interventions. Our hypothesis considers *informal labor* as the default type of occupation when production is done with a small number of differentiated capabilities by family units or small businesses and *formal employment* as an emerging phenomenon that takes place mainly in cities where a larger number of differentiated, and complementary, capabilities can be combined, mainly by firms. The idea that production processes arise from the complementarity of a large collection of capabilities is implicit in a range of models including those inspired by economic complexity theory (Hausmann and Hidalgo, 2011, Hidalgo and Hausmann, 2009, Hidalgo et al., 2007), by theoretical models of economic geography and the growth of regions (Frenken and Boschma, 2007), by the O'ring model of development (Kremer, 1993a), and by the literature on production recipes (Auerswald et al., 2000, Gomez-Lievano et al., 2017).

Cities thrive because they act as a cauldron, enabling firms to mix people and combine skills in a process

[1] Personal communication, April 2016.

[2] See section 2 below for the definition of cities.

[3] However, the (in)formality rate is often calculated as share of the occupied, not of the working age population. We adopt the latter not only for lack of data on the number of occupied persons by city, but also to avoid potential endogeneity problems as labor participation may be endogenous to city size.

of incremental diversification and sophistication of production. The emphasis on skills and the processes of mixing them is not new (see, e.g., Barro and Sala-i Martin, 2003, Jones, 1995, 2002, Jones and Romer, 2010, Kremer, 1993b, Lucas, 1988, Romer, 1986, 1990, Weil, 2012). Our emphasis here, however, is on the role of skills in the formation of increasingly complex industries. Employing a simple analogy, we consider skills as letters in a game of scrabble. The more letters - or skills - a city has, the greater the number of words firms can build, and the longer and more sophisticated the words become. Under this model, as cities acquire new skills, the number of possible industries grows more than proportionally, and the processes of sector diversification and labor creation occur (Hausmann and Hidalgo, 2011, Hausmann et al., 2014b, Hidalgo et al., 2007).

Our approach is consistent also with the literature on agglomeration economies (Duranton and Puga, 2001, 2004, Friedrichs, 1993, Rosenthal and Strange, 2006), which has emphasized the role that larger cities have in better facilitating matching between employers and employees, knowledge spillovers and innovation opportunities, which translate into the creation of new firms. Entrepreneurs and firms also benefit from sharing production inputs and risks. However, what we emphasize here is that this matching process has a structure, which results in path dependence in the way cities diversify (Glaeser et al., 1992). For example, new industries in a city that produces textiles, garments and leather will be very different from industries that flourish in a city that produces cars, electronics, and machinery. This approach is related to classic regional growth models in economic geography which focus on the role of *related varieties*, or the diversification of regions into industries similar to the existing economic structure (Frenken et al., 2007, Hausmann and Hidalgo, 2011, Klimek et al., 2012, Neffke and Henning, 2013, Neffke et al., 2011a,b, Nelson and Winter, 1982).

We focus our study on Colombia. Colombia is a fitting study case, not only because of data availability (see below), but due to several features of its economic geography and its labor market. Geographical diversity and differentiated patterns of colonization gave origin to geographically-varied levels of public goods provision and prosperity still visible today (Acemoglu et al., 2015). With 76 percent of its 47.1 million inhabitants residing in urban areas (in 2013, as estimated by DANE, the National Statistical Office), Colombia has 62 cities with at least 50,000 inhabitants (see definition of cities below). The (simple) average formal occupation rate in cities is only 16.7 percent of working age population (in 2013), but the standard deviation between cities is comparatively large (9.1 percent), providing room for statistical analysis. Furthermore, important changes in urban formal occupation rates occurred between 2008 and 2013, our period of analysis: the average change was 4.7 percent points, with a standard deviation across cities of 3.4 points. Formal occupation was facilitated by a rate of GDP growth of 4.2 percent, largely fuelled by a commodity boom (65 percent of Colombia's exports in 2013 were mining products), and probably also by the reduction in May of 2013 of payroll taxes representing 5 percent of the wage bill. Labor policies and institutions (including the minimum wage, labor taxes, social security and pensions) are geographically uniform. Major increases in public social expenditures since the Constitution of 1991 have substantially reduced the gaps in access across cities to education and health. Municipality mayors are democratically elected (2007 and 2011 were election years). Although many social and economic policy areas are formally decentralized, municipalities have almost no influence on industrial, innovation or training policies.

In order to study the emergence of new sectors in a city, we adapt a measure of sophistication previously

introduced by Hidalgo and Hausmann (2009) to study exports, which we refer to as *industry complexity*. We find that the elasticity of formal employment to working age population increases with industry complexity as more sophisticated sectors, requiring many more complementary skills and inputs, emerge more easily in larger cities. We construct a network where nodes represent industries and edges represent the degree of technological similarity between industry pairs[4]. We observe that workers join the formal economy at low complexity industries (nodes) in the network. Assuming that workers are constrained to move into new industries requiring similar skills in a path dependent manner, we employ a classic model for dynamics of a random walker on a network (Barahona and Pecora, 2002, Brin and Page, 1998, Mohar, 1991) to show that workers diffuse with low probability to more peripheral or inaccessible sophisticated industries. We argue that these sectors are more likely to be reached in larger cities with a more diverse labor force.

Under this network model, the location of a city in the network (i.e., the position of its relatively large industries) governs its diversification opportunities into complex sectors. We define a measure of complexity potential, which characterizes the potential of a city to diversify into sectors which are both complex and proximate (in a network sense) to current capabilities or skills (measured in terms of existing industries). We show that this metric predicts the change of formality rates for cities in Colombia (2008-2013), a result that is robust to inclusion of relevant controls such as the initial working age population, GDP per capita and the formal occupation rate in the base year, and to a diversity of exogenous shocks due to the oil boom, government expenditure decisions and nationwide sectoral demand changes. By interacting these shocks with the main explanatory variable we explore whether complexity potential operates as a channel of transmission of the shocks or as an endogenous engine of formal employment growth. Our results lend stronger support to the former possibility, but do not rule out the latter one. Furthermore, we verify that our mechanism competes successfully with more traditional explanations of (in)formality that are testable at a sub-national level, such as the provision of infrastructure and education and government capabilities. Finally, the mechanism proposed holds when the sample is split by city size and by industry complexity level, indicating that it valid for a variety of places and industry types.

2 Data and Definitions

Our units of analysis are cities, defined as *local labor markets*. This definition of city is important for us since the potential for creating productive teams of people, we argue, depends on the available local sources of labor[5]. Cities defined in this way are usually referred to as metropolitan or urban areas. We will use the terms cities, metropolitan or urban areas interchangeably. In practice, metropolitan areas are typically based on commuting patterns and are often composed of groups of smaller political or administrative units, such as counties or municipalities. Duranton (2013) proposed a methodology to define metropolitan areas, which he applied to the case of Colombia. It consists of adding iteratively a municipality to a metropolitan area if there is a share of workers, above a given threshold, that commute from the municipality to the metropolitan area. We use a 10% threshold, which in the case of

[4] In practice, edge weights are empirically estimated via labor transitions (workers moving between industries).

[5] We note that there is a lack of consensus on what a city is (Duranton, 2013, Pickett et al., 2011, Rozenfeld et al., 2008, 2009). As open systems in which people, information and materials are constantly flowing in and out, cities lack clear boundaries. There is, however, a general recognition that for the purpose of economic analysis, cities should represent labor markets.

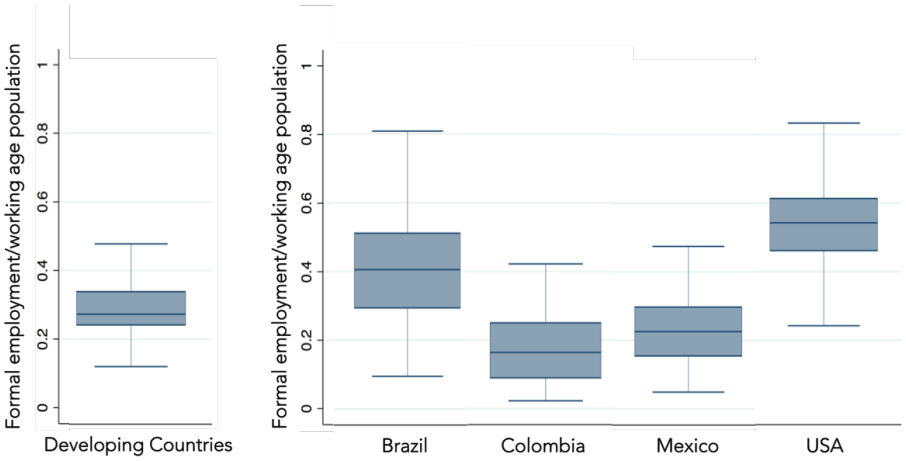


Figure 1: Box plots for the distribution of formal occupation rates in a set of developing countries, and cities in Brazil, Colombia, Mexico and the USA. All data for cities is for 2013 except Brazil which is 2010 - see Appendix for more details on data sources. The outer limits of the box correspond to the 25th and 75 percentile respectively. We observe a larger variance in formality rates across cities within countries than across countries, suggesting that the study of the determinants of formality across cities is an important area of research.

Colombia generates 19 metropolitan areas that consist of two or more municipalities, comprising a total of 115 municipalities. Similar to the standards of the US Office of Management and Budget (OMB) for metropolitan area delineations, we add to these 19 urban areas another 43 individual municipalities that have populations above 50,000 inhabitants, for a total of 62 cities.

Our analysis uses administrative data sets from the social security system of Colombia[6], part of which is publicly available at the Colombian Atlas of Economic Complexity (www.DatlasColombia.com). We count the monthly average number of individuals per city per industry in 2008 and 2013 that contributed to the social security system through firms (which excludes the self-employed). We call this the formal employment for a given industry and city. Hence, we define the formality rate of a city as *the total formal employment divided by the population older than 15*, or

$$F_{c,i} = \frac{\text{Formal Employment}}{\text{Working Age Population}} = \frac{\text{emp}_{c,i}}{w_c},$$

for industry i in city c. Analogously, $F_c = emp_c/w_c$ for city c. We operate with four-digit industry codes, using International Standard Industrial Classification (ISIC) codes (revision 3.0) with 390 industries[7].

3 The Emergence of Complexity

Do cities become 'more complex' as they increase in size? In order to address this question, we use the methodology developed by Hidalgo and Hausmann (2009) for export data, and adapt it to define a measure of sophistication of an industry. This measure is designed to capture the range of capabilities

[6] For each individual, our dataset only has information on age, industry of employment, type of social security contribution(s) and wages. The lack of data on education is a serious limitation to test hypotheses on the relation between education and the formation of productive capabilities.

[7] The full dataset includes 445 ISIC (revision 3.0) industry codes, but only 390 of these are connected to the labor flow network defined below. The remaining 55 industries represent less than 0.3% of total formal employment, and so for the purposes of consistency, we omit them from the whole analysis.

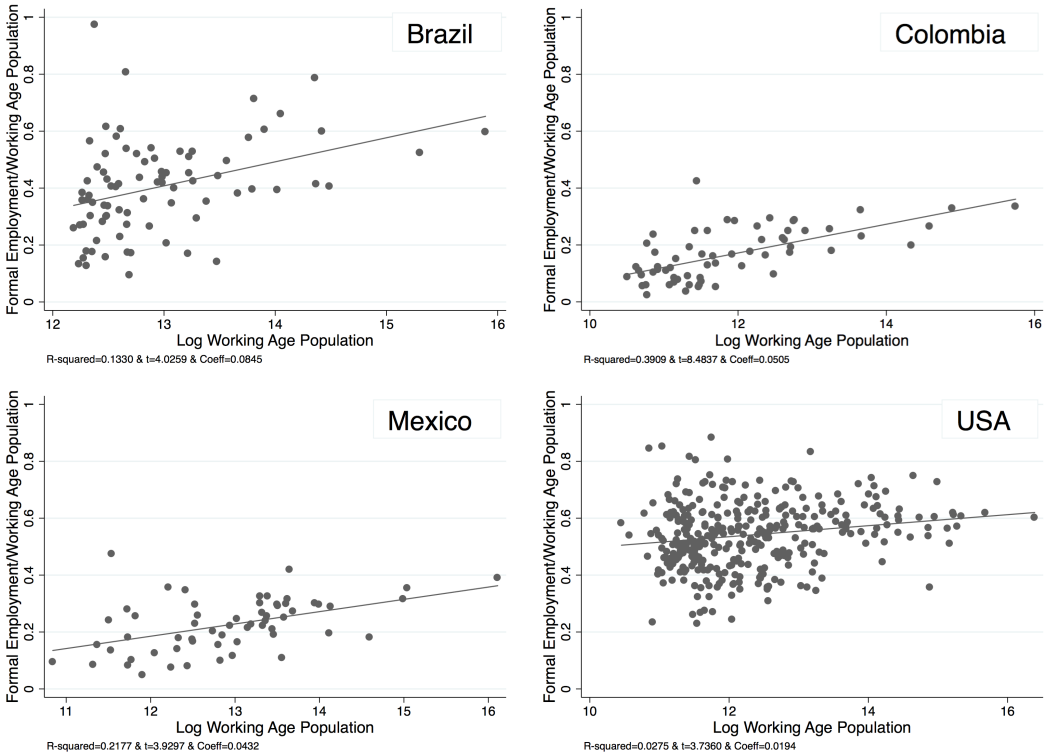


Figure 2: Formality rates increase with working age population for cities across Brazil, Colombia, Mexico and the USA. I.e., larger cities have disproportionately more workers in the formal sector than smaller cities as shown by the value of the coefficients (bottom left) respectively.

required for an industry to exist, and is computed based on an iterative model that counts the number of locations an industry is present in, and the average rarity of the industries those locations are active in. We call this measure the *complexity* of an industry. We present a mathematical derivation of this measure in the Appendix, and refer the reader to Hausmann et al. (2011), Hidalgo and Hausmann (2009) for further intuition.

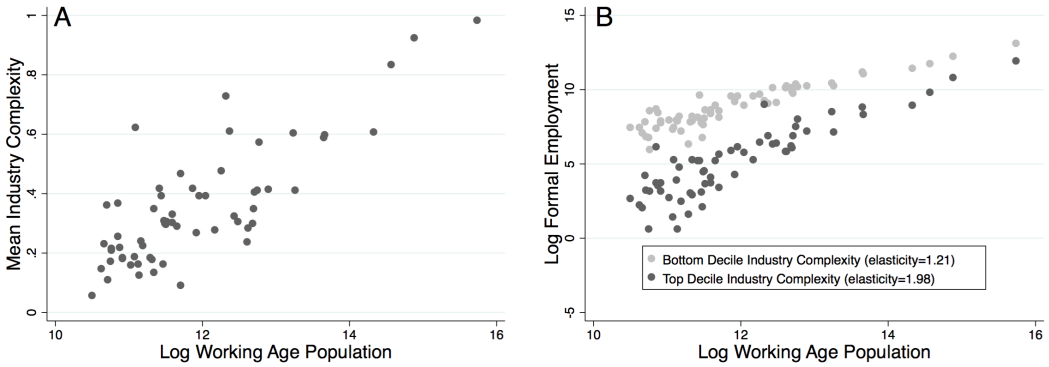
Figure 3 A shows the relationship between mean industry complexity[8] and working age population in cities. We observe that larger cities have an industrial mix that is more sophisticated. Figure 3 B shows the relationship between formal employment and working age population in cities for high complexity industries (top decile) and low complexity industries (bottom decile). We observe a steeper slope, or increased elasticity, in the former case, suggesting that the response of employment to city size is more pronounced in the more complex industries.

In order to systematically examine the changing distribution of labor by industry complexity, we compute the elasticity of industry-city employment to city size as follows. We run a regression for employment in cities c and industries i:

$$\log\left(\text{emp}_{c,i}\right) = \alpha + \beta \log(w_c) + \xi q_i + \gamma \log\left(w_c\right) q_i \tag{1}$$

where w_c is the working age population of city c, and q_i denotes the complexity of industry i. We then

[8]This can equivalently been seen as a city complexity index, analogous to the Economic Complexity Index (ECI) defined for countries by Hidalgo and Hausmann (2009).



C

Log Formal Employment	
Log Working Age Pop (β)	1.126***
	(0.021)
Industry Complexity (ξ)	-10.028***
	(0.827)
Interaction (γ)	0.525***
	(0.066)
Constant	-9.985***
	(0.253)
R2	0.38
N	15,858

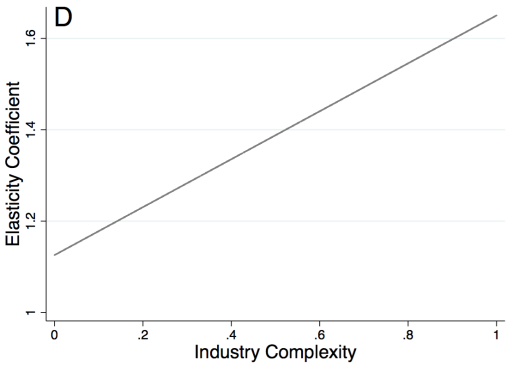


Figure 3: Subfigure A shows that mean industry complexity increases with the size of the working age population in Colombian cities. Subfigure B shows formal employment versus working age population for low complexity industries (bottom industry complexity decile) and high complexity industries (top industry decile) in all Colombian cities. The larger elasticity in the latter case indicates that employment in more complex industries increases more rapidly with city size. In order to systematically examine the changing distribution of labor by industry complexity, we compute the elasticity of industry-city employment to city size using Equation 1. Table C shows the results of this regression. Finally, in Subfigure D, we plot how the elasticity responds to change in complexity level as given by Equation 2.

compute the elasticity coefficients (i.e. taking the derivative respect to $\log(w_c)$):

$$\frac{\partial \log(emp_{ci})}{\partial \log(w_c)} = \beta + \gamma q_i \tag{2}$$

for each industry i.

Figure 3 C shows the results of the regression (1), and Figure 3 D plots the estimated elasticities as a function of industry complexity given by (2). We observe that formal employment in more complex industries rises more rapidly as the working age population increases, suggesting that large cities incubate a larger proportion of complex industries than smaller cities.

In the next section we analyze the process of urban diversification, and show using a network model that workers tend to join the formal workforce via low complexity industries, and progressively move into higher complexity sectors over time.

4 Network Analysis

So far, we have analyzed cross-sectional data for cities. In the following sections we present a network-based model to describe the growth of formality via a process where workers enter the formal labor force

at low complexity industries, and diffuse to more sophisticated sectors over time in larger cities.

The model hinges on two core mechanisms. The first, as outlined above, is that formal workers move, over time, into new more complex economic activities. The second is equally intuitive: the characteristics of existing economic activities in a city matter because new activities or industries are more likely to emerge if they use skills similar to those already deployed in the present industries.

4.1 Skill Similarity

Here we develop a network model for the mechanisms introduced above. In order to describe similarity between industries in terms of `cognitive distance' or skill transferability (Neffke and Henning, 2013), we measure worker flows (job switches) between industries. Under this model, cities diversify into new industries because workers switch between jobs that share similar know-how, as captured by the network structure.

Since all formal workers are observed in different periods, the number of times workers move from one industry to another can be computed[9]. An industry is considered related to another industry if the number of job switches between these industries is larger than what would be expected from randomizing all switches among all pairs. Formally, if $\phi_{i,j}$ is the number of job switches between industry i and industry j (between year t and year $t+1$), it can be computed as a matrix with entries

$$S_{i,j} = \frac{\phi_{i,j} / \sum_j \phi_{i,j}}{\sum_i \phi_{i,j} / \sum_{i,j} \phi_{i,j}}.$$

Although this matrix is asymmetric, it is made symmetric by averaging with its transpose, and re-scaling the values so that they range from -1 to 1:

$$A_{i,j} = \frac{S_{i,j} + S_{j,i} - 2}{S_{i,j} + S_{j,i} + 2}. \tag{3}$$

In the computations below, only positive values of this matrix (more job switches than expected) are taken into account[10].Full detail on these methodological considerations can be found in Neffke and Henning (2013).

4.2 Network Dynamics

Figure 4 is a network representation of the flows between industries in Colombia, where nodes are industries, and edges shown[11] represent inter-industry labor flows as described by Equation 3 above. Here nodes are colored by two-digit ISIC code, and edges have width proportional to the size of the labor flow between industry pairs. We observe clusters of closely related industries, driven by workers transitioning

[9] Not all industry switches correspond to transfer of skills, however. For instance, firms can change their industry classification due to acquisition, product diversification, and plant closures. Hence, industry switches that occur in blocks (i.e., many workers change industry at the same time) are discarded

[10] To be precise, if $(S_{i,j} + S_{j,i})/2 > 1$, or the average of the bi-directional switches between industries i and j is more frequent than expected, then $A_{i,j} > 0$ in Equation 3.

[11] The edges shown are determined the classic Maximum Spanning Tree algorithm, with the addition of edges with weight greater than 0.4.

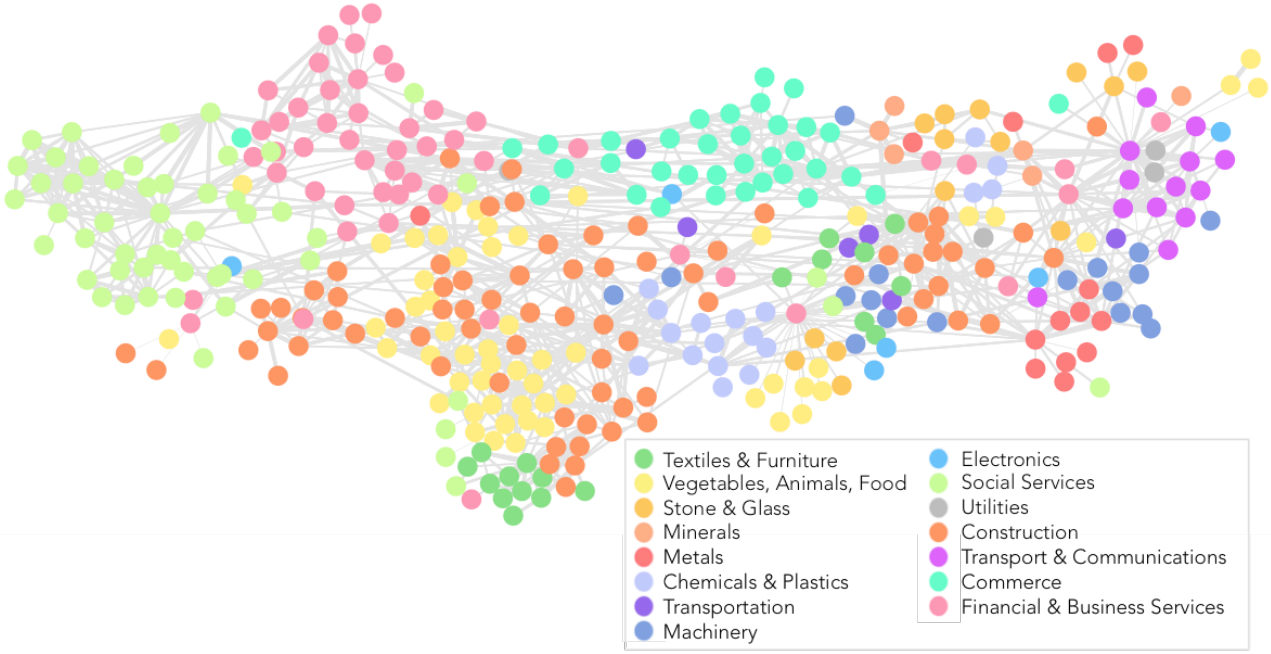


Figure 4: A network of labor flows between industries for Colombia is visualized (see also DatlasColombia.com by the same authors). Nodes represent industries, and are colored by two-digit sector code. It is observed that closely related industries tend to cluster, driven by workers transitioning between similar or complementary economic activities. This network models the flow of know-how within the Colombian economy, and can be used to model the path dependent process of industrial diversification for urban centres.

between similar or complementary economic activities. Service-orientated industries tend to be located to the left-hand side of the network, whereas heavy goods sectors dominate the far right. Retail and wholesale goods tend to inhabit the interior.

Under our hypothesis, a more diverse range of skills of workers is combined in larger cities enabling the presence of sophisticated economic activities. Less complex activities may be performed by workers with skills that are more common and less sophisticated, some of whom operate in the informal sector. We do not, however, have direct information on the ratio of formal workers to informal workers per industry - what we can observe is the entry of workers from outside the formal sector into each industry.

For each industry the number of workers previously absent from the formal labor force for a full nine-month period is computed as a share of the total worker inflow to the industry. In Figure 5 A1 nodes are colored by the share of workers who flow into the industry from outside the formal sector. Workers tend to join the formal labor force via low complexity sectors including public administration, services, retail and agriculture. Subfigure B shows a systematic relationship: the share of workers flowing in from outside the formal sector is negatively correlated with industry complexity, i.e. less sophisticated industries receive a larger share of workers previously not present in the formal sector.

We hypothesize that workers joining these low complexity sectors are unlikely to reach more sophisticated but peripheral industries. Only in large cities, with many workers, is this likely. This theory can be tested via a well-known mathematical model for diffusion of a random walker on a network (Delvenne et al., 2010, Lovász, 1993), which is alternatively known as a node ranking in other literatures (Brin and Page, 1998). Specifically, the long-run probabilities of a random walker transitioning each node, given any initial condition, are computed. Mathematically, a vector $\mathbf{x}_k \in R^n$ containing the transition probability

of a random walker transitioning each node at step k can be computed via the system

$$\mathbf{x}_{k+1} = W\mathbf{x}_k = AD^{-1}\mathbf{x}_k \tag{4}$$

with initial condition $\mathbf{x}_0 \in R^n$. The matrix D is a network of zeros with the column sums (node in-degree) of A on the diagonal (where A is defined via Equation (3) above). Hence, the probability of a random walker transitioning node i from node j is governed by the weight of the edge from i to j, normalized by the total weight of all other edges entering node i.

We seek to compute the long-run equilibrium probability of a random walker reaching each node. This value can be seen as a measure of the centrality of the node - intuitively, more central highly connected nodes are more likely to be traversed. Given two conditions on matrix W ($\mathbf{1}^T W = \mathbf{1}^T$ and all other eigenvalues within the unit disk), the system converges $\mathbf{x}_{k+1} = \mathbf{x}_k$ as $k \to \infty$. When this occurs, $\mathbf{x}_\infty = W\mathbf{x}_\infty$ and $\mathbf{x}_\infty$ may be seen as the eigenvector of W corresponding to eigenvalue 1. Hence, the eigenvector corresponding to eigenvalue 1 is the steady state vector of long-run transition probabilities.

As argued above, in large cities, workers diffuse from low complexity industries to more sophisticated activities on the periphery of the network. We have seen that workers enter low complexity industries. Subfigures A2 and A3 show that more complex industries are associated with a lower transition probability, a relationship confirmed by Subfigure C. Hence, nodes with a high share of labor inflows from outside the formal sector have high transition probabilities, meaning they tend to centrally-located, highly connected nodes in the network. Conversely, high complexity nodes are rarely traversed by a random walker. This result suggests that sophisticated (typically peripheral) industries are difficult to reach, and are thus reached with less frequency by workers, who tend with higher probability to join the network at low complexity, centrally located nodes. Over time, with a large supply of labor, workers will diffuse to high complexity but peripheral activities.

This analysis is evidence for the role of both path dependence and complexity in the growth of formal employment. Large cities provide an increased pool of workers, active in a diverse range of industries, enabling the emergence of high complexity sectors with higher probability. In the following section we will develop a network-based metric that captures these dynamics, and show that it is predictive of the growth of formality in cities.

5 A Model for the Growth of Formality in Cities

The previous sections have described the role that skill similarity across industries plays in the creation of formal employment, and how this process operates more strongly in larger cities with a more diversified skill base. We developed a network-based dynamic model that suggests that both the structure of connections between industries, and the industry complexity, influence the diversification process. Here, we show that the rate at which cities create formal employment is largely determined by the initial pool of skills (as captured by its existing industrial structure) and their proximity to complex industries.

To test the predictive power of the theory, the following model is used:

$$\Delta F_c(t+1) = \beta_0 + \beta_1 \log(CP_c(t)) + \beta_2 F_c(t) + \text{ controls} \tag{5}$$

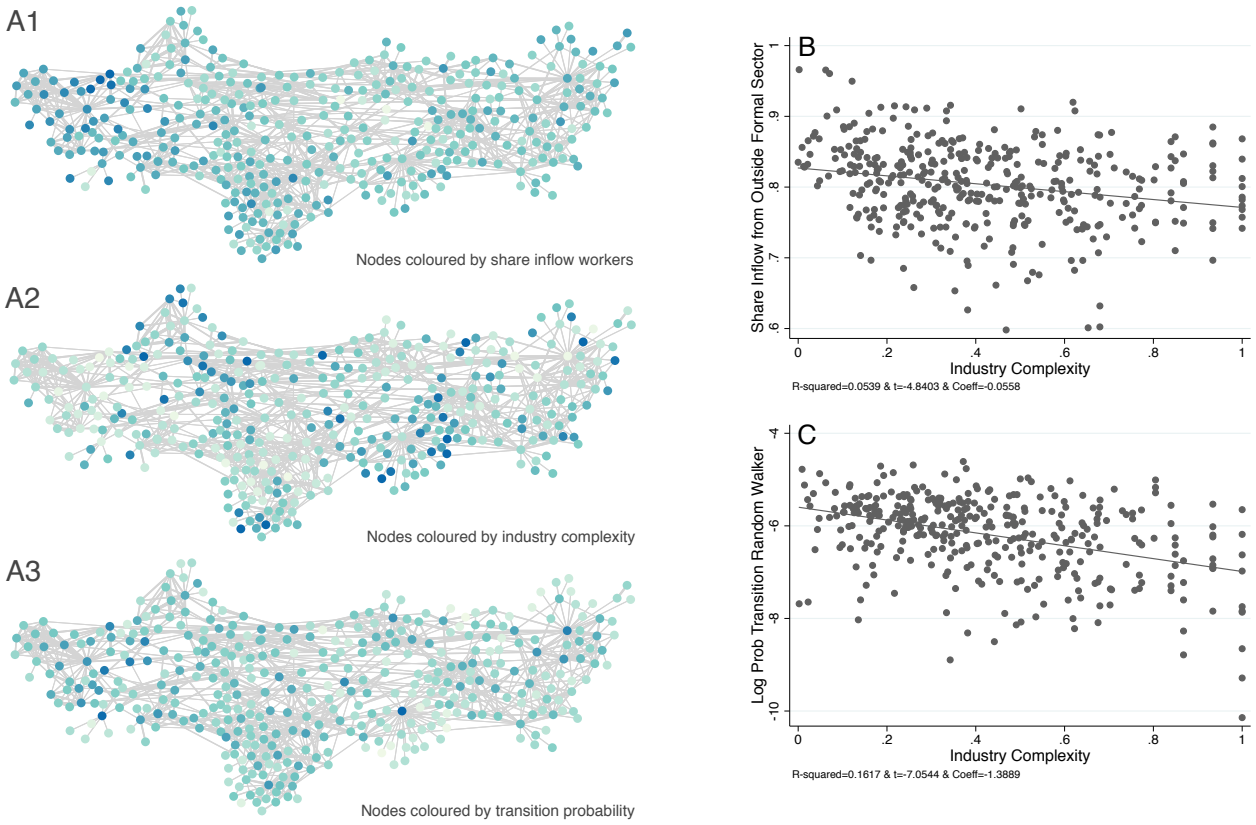


Figure 5: Panel A highlights that workers from outside the formal sector tend to enter the network at low complexity industries, which are typically located in densely connected regions of the network. Subfigure B confirms these relationships: low complexity industries have a higher share of workers flowing in from outside the formal sector. Worker inflows from outside the formal sector represent the 'initial condition' of a process where workers diffuse over time within the network. These dynamics are modeled by a classic random walker model, where the transition probability is the long-term steady state probability of a random walker traversing a node. In Subfigure C we observe that high complexity industries are associated with low transition probabilities are more inaccessible and less likely to be reached by workers joining the network from outside the formal sector. Under this model, only large cities with many workers have the capacity to occupy these sophisticated industries.

where $\Delta F_c(t+1)$ is the change of formal occupation rate between periods $t = 2008$ and $t+1 = 2013$, $CP_c(t)$ is *complexity potential* in the initial period (2008) and $F_c(t)$ is the initial formal occupation rate. $CP_c(t)$, captures information on both the potential for specific industries to grow (driven by the existence of missing but similar industries), and the complexity level of those industries. Further details of this variable are provided below. The coefficient of$F_c(t)$ will indicate if formal occupation across cities tend to converge ($\beta_2 < 0$) or diverge ($\beta_2 > 0$), conditional to the rest of explanatory variables. Controls will include initial working age population (as a proxy of the potential labor force), initial GDP per capita (an indicator of aggregate productivity) and measures of several exogenous shocks due to the presence of the oil sector, to government spending changes and to (Bartik-style - McGuire and Bartik, 1991) nationwide sectoral demand shocks (see Appendix for full details of variable definitions and data sources).

5.1 Complexity Potential

A city can be positioned in the network by locating the nodes (i.e., the industries) in which it has revealed comparative advantage with respect to the country as a whole. A revealed comparative advantage (RCA) larger than one means that the share of employment in that industry in that city is above the national

average[12] (which is essentially the same as the urban average). The relative size of a specific industry above what is expected indicates a consolidated presence of the industry in a place. The set of industries for which a city has RCA larger than one defines, in a sense, the economic or industrial profile of the city.

Hence, depending on where a city has industries located in the network, and how connected those nodes are to other industries, some cities may be better positioned to move into new industries in the future. In particular, some cities may have industries embedded in a neighborhood of complex industries, which, as discussed in the last section, absorb a relatively higher share of labor in larger as compared to smaller cities.

We propose that better positioned cities, with the potential to move into new industries that are both similar and complex, will increase their formal occupation rate faster over time. We quantify how well positioned a city is via its *complexity potential*, defined as the average distance times complexity of missing industries (those industries with RCA less than one). Distance, or skill proximity, for an industry-city pair in this calculation is the sum of the edge weights in the network connecting the industry (node) to all other 'existing' industries in the city (i.e., those with RCA greater than one), divided by the total weight of edges connected to that node.

Hence, formally, the *complexity potential* is defined as

$$CP_c = \frac{1}{|M_c|} \sum_{i \epsilon M_c} d_{c,i} C_i, \tag{6}$$

where M_c denotes the set of 'missing' industries for city c (RCA<1), and the $C_i \in [0, 1]$ is the normalized complexity of industry i. The distance weighting factor $d_{c,i}$, known as density in the literature (Hausmann et al., 2014b, Hausmann and Klinger, 2006, Hidalgo et al., 2007), is defined as

$$d_{c,i} = \frac{\sum_{j \epsilon N_c} A_{i,j}}{\sum_j A_{i,j}}.$$

where N_c is the set of industries that is present in city c.

Hence, CP_c provides a measure of aggregate potential complexity, taking into account both density and complexity of missing industries in a city, and is closely related but not identical to a previous measure constructed for trade data (Hausmann et al., 2014a). Our full model specification is given by Equation 5, and includes a range of controls discussed below.

5.2 Results

Table 1 presents the results of the model. Column 1 shows that the change in the formal employment rate is faster in cities that initially have a higher rate, which implies that formality rates tend to diverge across cities. In all the other regressions, the initial formal occupation rate is no longer significant, indicating that the (unconditional) divergence in formality rates can be explained by the other explanatory variables. Columns 2 and 3 suggest that the divergence could be the result of the initial complexity potential and/or other initial conditions (population size and GDP per capita). Columns 4 through 8 indicate that complexity potential is a robust determinant of formal employment growth. The coefficient is always

[12] The revealed comparative advantage is equivalent to the *Location Quotient* measured using employment.

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log Complexity Potential		0.031**		0.059***	0.052***	0.060***	0.053***	0.053***
		(0.013)		(0.016)	(0.012)	(0.015)	(0.014)	(0.013)
Log working age pop 2008			0.002	-0.015**	-0.009*	-0.014**	-0.012**	-0.008
			(0.007)	(0.007)	(0.004)	(0.006)	(0.006)	(0.005)
Log GDP per capita 2011			0.017**	0.013*	-0.013	0.005	0.007	-0.014
			(0.008)	(0.007)	(0.009)	(0.009)	(0.008)	(0.009)
Binary oil: one well/1000					0.074**			0.069***
					(0.028)			(0.022)
Govt spending pc						0.024		0.014
						(0.015)		(0.013)
Sectoral demand							1.017	-0.333
							(0.667)	(0.488)
Formality rate 2008	0.171**	0.082	0.088	0.103	0.109	0.129	-0.174	0.204
	(0.072)	(0.090)	(0.121)	(0.089)	(0.068)	(0.092)	(0.162)	(0.159)
Constant	0.027***	0.128***	-0.141	0.262**	0.401***	0.311**	0.266**	0.402***
	(0.007)	(0.047)	(0.088)	(0.109)	(0.117)	(0.132)	(0.115)	(0.117)
Observations	62	62	62	62	62	61	62	61
R-squared	0.131	0.208	0.193	0.340	0.488	0.410	0.377	0.506

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1: Regressions for the change in the formality rate (ratio of formal employment to working age population) between 2008 and 2013 for Colombian cities. Explanatory variables include complexity potential in 2008, the working age population in 2008, GDP per capita in 2011 (as it is not available for earlier years) and the formal occupation rate in the base year 2008. Also tested is the influence of exogenous demand shocks facing cities: due to the presence of the oil industry, to government expenditure changes and to nationwide sectoral shocks. We observe that complexity potential is a robust determinant of formal employment changes.

VARIABLES	(1)	(2)	(3)	(4)
Log Complexity Potential	0.052***	0.011	0.036**	-0.026
	(0.012)	(0.031)	(0.016)	(0.028)
Log working age pop 2008	-0.009*	-0.013**	-0.019**	-0.019***
	(0.005)	(0.005)	(0.008)	(0.007)
Log GDP per capita 2011	-0.014	0.004	0.008	-0.016*
	(0.009)	(0.010)	(0.008)	(0.009)
Binary oil: one well/1000	0.035			-0.073
	(0.150)			(0.129)
Log CP x Binary oil	-0.014			-0.047
	(0.054)			(0.049)
Govt spending pc		0.293**		0.340**
		(0.144)		(0.129)
Log CP x Govt spending pc		0.090*		0.107**
		(0.046)		(0.042)
Sectoral demand			3.071**	2.500*
			(1.491)	(1.462)
Log CP x Sectoral demand			0.798	1.149**
			(0.496)	(0.534)
Formality rate 2008	0.113	0.111	-0.054	0.410**
	(0.068)	(0.090)	(0.165)	(0.202)
Constant	0.413***	0.165	0.286**	0.294**
	(0.119)	(0.144)	(0.116)	(0.117)
Observations	62	61	62	61
R-squared	0.489	0.459	0.401	0.595

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2: Regressions for the change in the formality rate (ratio of formal employment to working age population) between 2008 and 2013 for Colombian cities including interaction terms between complexity potential and exogenous shocks in order to test if complexity potential operates as a channel of transmission of the shocks, which it does in some cases.

significant and very stable after accounting for a number of possible confounding effects, including not just the initial conditions but a diversity of exogenous demand shocks facing cities due to the presence

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
Log Complexity Potential	0.0594***	0.0506***	0.0636***	0.0573***	0.0535***	0.0512***
	(0.0160)	(0.0129)	(0.0148)	(0.0130)	(0.0121)	(0.0124)
Log working age pop 2008	-0.0146**	-0.00749*	-0.0123**	-0.0110**	-0.00969**	-0.00779
	(0.00675)	(0.00436)	(0.00491)	(0.00482)	(0.00462)	(0.00499)
Log GDP per capita 2011	0.0132*	0.00800	0.00835	0.0146**	0.00822	-0.00794
	(0.00727)	(0.00903)	(0.00777)	(0.00637)	(0.00697)	(0.00716)
Binary oil: one well/1000						0.0444***
						(0.0148)
Govt spending pc						0.00843
						(0.00975)
Sectoral demand						-0.0745
						(0.537)
Infrastructure		0.00348			0.0119**	0.0112**
		(0.00504)			(0.00462)	(0.00487)
Institutions			-0.00570**		-0.000702	-0.000888
			(0.00259)		(0.00323)	(0.00316)
Higher Education				-0.00874***	-0.0109***	-0.00914***
				(0.00190)	(0.00276)	(0.00255)
Formality rate 2008	0.103	0.0141	0.0773	0.120*	0.111*	0.144
	(0.0890)	(0.0640)	(0.0668)	(0.0618)	(0.0572)	(0.171)
Constant	0.262**	0.193	0.325***	0.229**	0.222**	0.330***
	(0.109)	(0.119)	(0.120)	(0.104)	(0.108)	(0.113)
Observations	62	55	55	55	55	55
R-squared	0.340	0.346	0.383	0.503	0.575	0.637

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3: Here we test alternative hypotheses of formal employment creation: quality of infrastructure, quality of institutions, and amount and quality of higher education. All regressions include the initial conditions (working age population, GDP per capita and formal occupation rate). Good public infrastructure contributes to formal employment creation but neither good institutions nor higher education do, which get the wrong sign. Complexity potential is robust to the inclusion of these additional variables.

of the oil industry, to government expenditure changes and to (Bartik-style - McGuire and Bartik, 1991) nationwide sectoral shocks. The results suggest that, holding everything else constant, a 1 percent increase in complexity potential is associated with a 5 percent point increase in the rate of formality in the Colombian cities over the five-year period studied (2008-13).

Furthermore, the results of Table 1 shed some light on an alternative hypothesis recently put forward by La Porta and Shleifer (2014), namely that productivity is the main determinant of labor formality. GDP per capita, which is intended to account for differences in productivity, is weakly significant and not robust. Although the evidence is ambiguous, the hypothesis may be consistent with the mechanism proposed by us because higher productivity may be a consequence of the growth of complexity.

Next we consider the possibility that complexity potential transmits shocks. For example, we ask whether a combination of complexity and government spending explains changes in formality rate (i.e., is government spending effective in increasing formal employment when the level of complexity potential is high)? Table 2 addresses the issue with the inclusion of interaction terms of complexity potential and each of the shocks. The results clearly indicate that complexity is a channel of transmission of government spending shocks (regressions 2 and 4), and less clearly of nationwide sectoral demand shocks (regressions 3 and 4). Under an equilibrium model, we would not expect complexity potential to be independently significant for the creation of formal employment when a shock is applied. Outside of the equilibrium assumption (and consistent with the theory of complexity, see Beinhocker, 2006), however, complexity potential may operate independently as an organic source of formal employment creation. Regression 4 suggests that

complexity potential does not operate as an autonomous source of formal employment growth yet due to the large number of regressors in relation to the sample size, no final conclusion can be drawn from these results.

Other alternative explanations of (in)formality are tested in Table 3. The regressors of interest are indicators of the quality of urban infrastructure, the quality of public institutions and the amount and quality of high education. Following the methodology of the Global Competitiveness Report (Schwab and Sala-i Martín, 2013), this information is collected annually since 2013 for 26 of the 33 Colombian departments including Bogota DC (Consejo Privado de Competitividad). We have attributed to 55 cities the information by department for 2013. The quality of infrastructure is significant with the right sign in regression 6, but the other two variables always get the wrong sign (significantly on occasions). Importantly, complexity potential, our main explanatory variable, is always significant with a very stable coefficient. Therefore the alternative explanations do not fare very well vis-a-vis the mechanism proposed by us.

Finally, in Table 4, the sample is split in alternative ways in order to further test the robustness of the main explanatory variable. All columns include, in addition to complexity potential for the relevant subsample, the initial conditions (namely working age population, GDP per capita and formal occupation rate). The first four columns compare large and small cities: the relevant coefficient is strongly significant in both cases (before and after controlling for shocks), but about twice as large in the smaller cities, suggesting that developing productive know-how in industries technologically close to more complex industries pays a larger dividend in terms of formal employment creation in the smaller cities. When industries are split between high and low complexity (and complexity potential recalculated accordingly), the coefficient is always significant but several times larger for high-complexity industries (columns 5 through 8). Therefore, by developing know-how closer to that deployed in more complex industries, cities can create more formal employment than by focusing on the skills needed in the low-complexity industries.

6 Conclusion

This is the first paper to assess the role that city population size plays in formal employment creation. This is an important issue because formal employment rates show a larger variance across cities than across countries. Therefore the focus on the national-level determinants of (in)formality is at least incomplete, if not misleading. We have proposed a mechanism to explain why larger cities generate more formal employment: urban agglomeration facilitates the exploitation by firms of a more diverse pool of skills in new complex industries that absorb workers into the formal sector.

Making use of the notion of 'industry economic complexity', we showed that industries that are more complex increase in size faster with working age population. More complex industries do not directly absorb workers from outside the formal sector. Rather, workers tend to enter low complexity sectors, and transition with low probability to more sophisticated economic activities. Hence, larger cities, with a deep and diverse labor pool, are more likely to develop complex industries. In order to test the predictive power of the mechanism proposed, we defined a measure of 'complexity potential' which characterizes the potential of a city to diversify into more complex sectors. A well-positioned city, exhibiting high complexity potential, is typically host to a number of industries that share skills with higher complexity industries primed to absorb new workers into the formal sector. A number of econometric tests indicates

VARIABLES	(1) High WP	(2) High WP	(3) Low WP	(4) Low WP	(5) High PCI	(6) High PCI	(7) Low PCI	(8) Low PCI
Log Complexity Potential	0.052**	0.052***	0.105**	0.103**				
	(0.020)	(0.014)	(0.042)	(0.048)				
Log Complexity Potential (low)					0.045***	0.044***		
					(0.017)	(0.013)		
Log Complexity Potential (high)							0.010***	0.008***
							(0.003)	(0.003)
Log working age pop 2008	-0.024	-0.024	-0.020**	-0.017*	-0.013**	-0.007	0.001	0.001
	(0.017)	(0.019)	(0.008)	(0.009)	(0.006)	(0.004)	(0.002)	(0.002)
Log GDP per capita 2011	0.017	-0.030*	0.006	-0.008	0.008	-0.011	0.005	-0.002
	(0.011)	(0.015)	(0.010)	(0.010)	(0.006)	(0.008)	(0.003)	(0.003)
Formality rate 2008	0.093	0.407	0.080	0.090				
	(0.130)	(0.305)	(0.088)	(0.246)				
Formality rate 2008 (low)					0.125	0.292*		
					(0.103)	(0.173)		
Formality rate 2008 (high)							-0.087	-0.085
							(0.091)	(0.145)
Binary oil: one well/1000		0.110***		0.036**		0.057***		0.010*
		(0.036)		(0.017)		(0.017)		(0.005)
Govt spending pc		0.017		0.015		0.013		0.001
		(0.013)		(0.013)		(0.011)		(0.002)
Sectoral demand		-0.965		0.014		-0.599		0.131
		(1.125)		(0.700)		(0.441)		(0.095)
Constant	0.306	0.707**	0.526**	0.595**	0.287**	0.378***	-0.014	0.033
	(0.245)	(0.270)	(0.232)	(0.252)	(0.127)	(0.124)	(0.037)	(0.035)
Observations	31	30	31	31	62	61	62	61
R-squared	0.350	0.606	0.272	0.375	0.281	0.457	0.325	0.470

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4: Here we perform further experiments to test the robustness of our findings. Columns 1 to 4 split the city sample by the size of the working age population; columns 5 to 8 keep all the cities but split the industry set in each city by complexity level. Complexity potential is always significant but has a larger impact on formal employment in smaller cities and in high-complexity industries.

that complexity potential has strong predictive power and fares better than alternative explanations of (in)formality. These results are analogous to those by Hidalgo et al. (2007) for nation-level economies, where initial export complexity is a much stronger predictor of economic growth than other variables often emphasized in the development literature, such as governance and education.

Several policy implications can be derived from these results. Most policy actions typically recommended to reduce labor informality are operable at the national level only, and therefore allow for little or no policy action by sub-national governments or councils. They include lowering minimum wages and social security contribution, as well as reducing tax rates (especially but not only on labor) and strengthening enforcement capabilities. In our alternative view, the path to reducing informality in cities lies in facilitating the mobilization and enrichment of productive skills to develop more complex industries. Policy interventions by sub-national governments may include:

1. Removing barriers, providing incentives and/or designing and implementing industrial reallocation plans for the development of industries with potential given the current availability of skills (industries that can be readily identified in the case of Colombia via www.DatlasColombia.com);
2. Implementing training and technical education programs to fill the skill gaps facing industries of high strategic value for the development of relatively complex industries (idem);
3. Facilitating skill matching by firms operating in these industries (through interventions such as making on-line job announcements compulsory, providing on-line advice to job applicants, etc.);

4. Reducing agglomeration costs due to congestion and spatial segmentation affecting these industries.

On the latter point, follow on work by O'Clery and Lora (2016) has shown that urban agglomerations with a 'diameter' of about 45-75 minutes commuting time maximizes the impact that the availability of skills has on the ability of cities to generate formal employment. Hence, providing adequate public transport is key to connecting firms to workers with a diverse skill set.

In general, an important issue for the effectiveness of these policies is that they should be designed and implemented not necessarily at the municipality level, but at the metropolitan-area level at which the local labor market operates, which may include more than one municipality and may not correspond to the officially-defined metropolitan areas.

7 Acknowledgements

We would like to acknowledge the contribution of Matte Hartog, Alfredo Guerra and Jose Ramon Morales Arilla for their help with data collection and cleaning, and Juan Pablo Chauvin and Carmen Pages-Serra for helpful comments and discussion.

8 Funding

The authors would like to thank the Julio Santo Domingo Foundation for their generous support of research activities at the Center for International Development at Harvard University.

References

Acemoglu, D., García-Jimeno, C., Robinson, J. A., 2015. State capacity and economic development: A network approach. The American Economic Review 105 (8), 2364--2409.

Auerswald, P., Kauffman, S., Lobo, J., Shell, K., 2000. The production recipes approach to modeling technological innovation: An application to learning by doing. Journal of Economic Dynamics and Control 24 (3), 389--450.

Barahona, M., Pecora, L. M., 2002. Synchronization in small-world systems. Physical Review Letters 89.5, 54101.

Barro, R. J., Sala-i Martin, X., 2003. Economic Growth, 2nd Edition. MIT Press.

Beinhocker, E. D., 2006. The origin of wealth: Evolution, complexity, and the radical remaking of economics. Harvard Business Press.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems 30 (1), 107--117.

Busso, M., Fazio, M. V., Algazi, S. L., 2012. (In) formal and (Un) productive: The productivity costs of excessive informality in Mexico. IDB Working Paper No. IDB-WP-341.

Castells, M., Portes, A., 1989. The informal economy: Studies in advanced and less developed countries. John Hopkins University Press, Ch. World underneath: The origins, dynamics, and effects of the informal economy.

CEDE, 2016. Colombian data on government spending.
URL `https://datoscede.uniandes.edu.co/`

Chen, M. A., 2012. The informal economy: Definitions, theories and policies. Women in informal economy globalizing and organizing: WIEGO Working Paper 1.

Colombian Atlas of Economic Complexity, 2016. Colombian formal employment and industry/region mapping.
URL `http://datlascolombia.com/`

Conapo, 2016. Mexican metropolitan areas.
URL `http://www.conapo.gob.mx/es/CONAPO/Zonas_metropolitanas_2010`

Consejo Privado de Competitividad, 2016. Colombian competitiveness indices.
URL `http://compite.com.co/`

County Business Patterns, 2016. United States formal employment data.
URL `http://www.census.gov/programs-surveys/cbp.html`

DANE, 2016. Colombian National Statistics Office.
URL `http://www.dane.gov.co`

Dataviva, 2016. Brazilian formal employment data.
URL `http://dataviva.info/`

De Soto, H., 1990. The other path: The invisible revolution in the third world.

De Soto, H., 2000. The mystery of capital: Why capitalism triumphs in the West and fails everywhere else. Basic Books.

Delvenne, J.-C., Yaliraki, S., Barahona, M., 2010. Stability of graph communities across time scales. Proceedings of the National Academy of Sciences 107, 12755--12760.

Doeringer, P. B., Piore, M. J., 1975. Unemployment and the dual labor market. The Public Interest (38), 67--79.

Duranton, G., 2013. Delineating metropolitan areas: Measuring spatial labour market networks through commuting patterns. Processed, Wharton School, University of Pennsylvania.

Duranton, G., Puga, D., 2001. Nursery cities: Urban diversity, process innovation, and the life cycle of products. American Economic Review, 1454--1477.

Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. Handbook of regional and urban economics 4, 2063--2117.

Fields, G., 1975. Rural-urban migration, urban unemployment and underemployment, and job-search activity in LDCs. Journal of development economics 2 (2), 165--187.

Frenken, K., Boschma, R. A., 2007. A theoretical framework for evolutionary economic geography: industrial dynamics and urban growth as a branching process. Journal of Economic Geography 7, 635--649.

Frenken, K., Van Oort, F., Verburg, T., 2007. Related variety, unrelated variety and regional economic growth. Regional Studies 41 (5), 685--697.

Friedrichs, J., 1993. A theory of urban decline: Economy, demography and political elites. Urban Studies 30 (6), 907--917.

Ghani, E., Kanbur, R., 2013. Urbanization and (in) formalization. World Bank Policy Research Working Paper (6374).

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., Shleifer, A., 1992. Growth in cities. Journal of Political Economy 100 (6), 1126--1152.

Gomez-Lievano, A., Patterson-Lomba, O., Hausmann, R., 2017. Explaining the prevalence, scaling and variance of urban phenomena, In Press at Nature Human Behaviour.

Hart, K., 1973. Informal income opportunities and urban employment in Ghana. The Journal of Modern African studies 11 (01), 61--89.

Hausmann, R., Cunningham, B., Matovu, J. M., Osire, R., Wyett, K., 2014a. How should Uganda grow? Available at SSRN 2439277.

Hausmann, R., Hidalgo, C., October 2011. The network structure of economic ouput. Journal of Economic Growth 16, 309--342.

Hausmann, R., Hidalgo, C., Bustos, S., Coscia, M., Chung, S., Jimenez, J., Simoes, A., Yildirim, M., 2011. The atlas of economic complexity: Mapping paths to prosperity. Harvard University Center for International Development, MIT Media Lab.

Hausmann, R., Hidalgo, C., Stock, D. P., Yildirim, M. A., 2014b. Implied comparative advantage. HKS Working Paper No. RWP14-003.

Hausmann, R., Klinger, B., 2006. The evolution of comparative advantage: the impact of the structure of the product space. Center for International Development and Kennedy School of Government, Harvard University.

Hidalgo, C., Hausmann, R., 2009. The building blocks of economic complexity. Proc Natl Acad Sci USA 106 (26), 10570--10575.

Hidalgo, C. A., Klinger, B., Barabási, A.-L., Hausmann, R., 2007. The product space conditions the development of nations. Science 317, 482--487.

IBGE, 2016. Brazillian population data.
URL `ftp://ftp.ibge.gov.br/Estimativas_de_Populacao/Estimativas_2014/estimativa_dou_2014.pdf`

INEGI, 2016. Mexican population data.
URL `http://www.inegi.org.mx/est/contenidos/Proyectos/encuestas/hogares/regulares/enoe/`

Jones, C. I., 1995. R&D-based models of economic growth. Journal of Political Economy 103 (4), 759--784.

Jones, C. I., March 2002. Sources of U.S. Economic Growth in a World of Ideas. The American Economic Review 92 (1), 220--239.

Jones, C. I., Romer, P. M., January 2010. The new kaldor facts: Ideas, institutions, population, and human capital. American Economic Journal: Macroeconomics 2, 224--245.

Klimek, P., Hausmann, R., Thurner, S., 2012. Empirical confirmation of creative destruction from world trade data. PloS one 7 (6).

Kremer, M., 1993a. The o-ring theory of economic development. The Quarterly Journal of Economics, 551--575.

Kremer, M., August 1993b. Population growth and technological change: One million B.C. to 1990. The Quarterly Journal of Economics 108 (3), 681--716.

La Porta, R., Shleifer, A., 2014. Informality and development. The Journal of Economic Perspectives, 109--126.

Levy, S., 2010. Good intentions, bad outcomes: Social policy, informality, and economic growth in Mexico. Brookings Institution Press.

Lewis, W. A., 1954. Economic development with unlimited supplies of labour. The Manchester School 22 (2), 139--191.

Lovász, L., 1993. Random walks on graphs: A survey. Combinatorics, Paul Erdős is eighty 2 (1), 1--46.

Lucas, R., July 1988. On the mechanics of economic development. Journal of Monetary Economics 22 (1), 3--42.

Lucas, R. E., 1978. On the size distribution of business firms. The Bell Journal of Economics, 508--523.

Maloney, W. F., 2004. Informality revisited. World Development 32 (7), 1159--1178.

McGuire, T. J., Bartik, T. J., 1991. Who benefits From state and local economic development policies? JSTOR.

Mexican Atlas of Economic Complexity, 2016. Mexican formal employment and industry/region mapping. URL `http://complejidad.datos.gob.mx/`

Mohar, B., 1991. The Laplacian spectrum of graphs. Graph theory, Combinatorics, and Applications 2, 871 -- 898.

Neffke, F., Henning, M., 2013. Skill relatedness and firm diversification. Strategic Management Journal 34 (3), 297--316.

Neffke, F., Henning, M., Boschma, R., 2011a. How do regions diversify over time? industry relatedness and the development of new growth paths in regions. Economic Geography 87 (3), 237--265.

Neffke, F., Henning, M., Boschma, R., Lundquist, K.-J., Olander, L.-O., 2011b. The dynamics of agglomeration externalities along the life cycle of industries. Regional Studies 45 (1), 49--65.

Nelson, R., Winter, S., 1982. An evolutionary theory of economic change. The Belknap Press of Harvard University Press, Cambridge.

O'Clery, N., Lora, E., 2016. City size, distance and formal employment creation. CAF (Latin American Development Bank) Working Paper.

Perry, G. (Ed.), 2007. Informality: Exit and exclusion. World Bank Publications.

Pickett, S., Cadenasso, M., Grove, J., Boone, C. G., Groffman, P. M., Irwin, E., Kaushal, S. S., Marshall, V., McGrath, B. P., Nilon, C., Pouyat, R., Szlavecz, K., Troy, A., Warren, P., October 2011. Urban ecological systems: Scientific foundations and a decade of progress. Journal of Environmental Management 92, 331--362.

Romer, P. M., October 1986. Increasing returns and long-run growth. The Journal of Political Economy 94 (5), 1002--1037.

Romer, P. M. P., October 1990. Endogenous technological change. Journal of Political Economy 98 (5), S71--S102.

Rosenthal, S., Strange, W., 2006. The micro-empirics of agglomeration economies. A Companion to Urban Economics, 7--23.

Rozenfeld, H. D., Rybski, D., Andrade, J. S., Batty, M., Stanley, H. E., Makse, H. A., 2008. Laws of population growth. Proceedings of the National Academy of Sciences 105 (48), 18702--18707.

Rozenfeld, H. D., Rybski, D., Gabaix, X., Makse, H. A., 2009. The area and population of cities: New insights from a different perspective on cities. Tech. rep., National Bureau of Economic Research.

Schwab, K., Sala-i Martín, X., 2013. The global competitiveness report 2013--2014: Full data edition. In: World Economic Forum. p. 551.

Singer, H. W., 1973. Employment, incomes and equality: A strategy for increasing productive employment in Kenya. International Labour Office.

US Census, 2016. United States working age population by county.
URL http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

US Office of Management and Budget, 2016. United States definition of metropolitan areas.
URL http://www.census.gov/population/metro/data/metrodef.html

Weil, D. N., 2012. Economic growth, 3rd Edition. Prentice hall, New York.

A Information on Data Sources

A.1 Colombia

Our main data source is the administrative data PILA (the Integrated Report of Social Security Contributions), managed by the Colombia Ministry of Health and made accessible at the Colombian Atlas of Economic Complexity. It contains information on formal employment by firms, municipalities and industries for 2008-2013. All types of sectors, including goods and services, are included (note that we use the terms 'sector' and 'industry' interchangeably). We compute the monthly average number of individuals per city per industry that contributed to the social security system - but were not self-employed. We call this the formal employment for a given industry and city. We use data aggregated at four-digit industry codes, using ISIC (revision 3.0) with 390 industries, and 62 cities.

Cities are defined using the methodology of Duranton (2013). Applying his algorithm to population and commuting data by municipality from the 2005 Census by DANE (National Statistics Office) we obtain 19 urban areas that consist of two or more municipalities, comprising a total of 115 municipalities. Similar to the standards of the US Office of Management and Budget (OMB) for metropolitan area delineations, we add to these 19 urban areas another 43 individual municipalities with urban populations above 50,000 inhabitants in 2008 according to DANE, for a total of 62 cities.

The dependent variable in the regressions of Tables 1, 2, 3 and 4 is the change in the *formality rate* as defined in the text, and computed from the social security system data mentioned above and working age population data (see below). The main explanatory variable is *complexity potential*, computed from the PILA data as explained in the text. Controls in the regressions of said tables, in addition to the initial level of the formality rate, are computed as follows:

- *Working age population* in 2008, defined as population 15 or older, is calculated from population data by age groups and municipality estimated by DANE. This variable (for 2008 and 2013) is also used to compute the formality rate by city.
- *GDP per capita* in 2011 is computed from GDP estimates by municipality by DANE. No earlier estimates are available.
- *Oil producing city*, a binary variable, which takes the value of one if the city has more than one oil well in production per thousand inhabitants. Oil well data refers to 2014, as reported by Ecopetrol (the Colombian hydrocarbon company) for their own internal records.
- *Government spending shock* is the change between 2008 and 2013 in total government spending (in 2008 prices) per working-age person. It is computed from municipality-level government spending data compiled by CEDE.
- *Sectoral demand shocks*, sds_c, is a Bartik-style measure (see McGuire and Bartik, 1991) that summarizes for each city the mix of nationwide sectoral demand shocks facing the city. It is computed as

$$sds_c = FOR_c \sum_i \frac{femp_{c,i}(2008)}{femp_c(2008)} g_{i,c} \tag{7}$$

where $g_{i,c} = log[femp_i(2013)] - log[femp_i(2008)]$ is growth of employment of industry i *excluding*

employment in industry i in city c. In other words, here $femp_i = \sum_{j \in J} femp_{i,j}$ with set J containing all cities except city c.

- *Government capabilities, higher education and transportation infrastructure* for 2013 (the earliest year available) are taken from Consejo Privado de Competitividad and Universidad del Rosario. This data was collected for 26 of the 33 Colombian departments following the methodology of the Global Competitiveness Report by Schwab and Sala-i Martín (2013). We have attributed the data for the corresponding department to 55 cities (the departments of the remaining seven cities were not covered).

A.2 Mexico, Brazil and USA

Data on formal employment for Mexico is publicly available (by municipality) from the Mexican Atlas of Economic Complexity also produced by the Centre for International Development at Harvard (accessible at www.complejidad.datos.gob.mx). Population by age groups is also publicly available for Mexico from INEGI. We use the official definition of `Metropolitan Zones' from Conapo, for which there are 59 zones.

Formal employment data for Brazil comes from RAIS (by municipality) via Dataviva (www.dataviva.info). The working age population data (by municipality/age group) is from census data 2010. A list of the 82 most populous cities based on the population of the municipality where the city is located was obtained from IBGE.

Formal employment data for the USA is provided (by metropolitan area) from the County Business Patterns for 2013. The working age population is from the US Census by county. We use the standard definition of Metropolitan Statistical Areas for the US from the US Office of Management and Budget, of which there are 388.

A.3 Other Developing Countries

The data on formality rates over a collection of developing countries is based on information from SEDLAC (CEDLAS and The World Bank) for countries in Latin America and the Caribbean, and International Labor Organization (ILO) for other countries. SEDLAC defines informal workers as salaried workers in firms of less than 5 employees, non-professional self-employed and unpaid workers. ILO defines informal workers as those who in their main or secondary jobs are own-account workers, employers and members of cooperatives or employed in their own informal sector enterprises. Working age populations (15 years or older) are taken from the same sources.

B Industry Complexity

As proposed by Hidalgo et al. (2007), Hidalgo and Hausmann (2009) and Hausmann et al. (2011), the complexity of an economy is related to the range of useful knowledge embedded in it. The industry complexity metric, analogous to the previously defined product complexity metric for export data (Hidalgo et al., 2007), attempts to estimate the range of capabilities needed for that industry to be present (in a city for example).

	(1)							
	Change FOR	Log CP	Log WPop	Log GDPpc	FOR 2008	Infrastructure	Institutions	Higher Education
Change FOR	1							
Log CP	0.433***	1						
Log WPop	0.221	0.800***	1					
Log GDPpc	0.391**	0.214	0.157	1				
FOR 2008	0.362**	0.558***	0.678***	0.485***	1			
Infrastructure	0.310*	0.266*	0.202	0.500***	0.384**	1		
Institutions	0.0483	0.355**	0.202	0.123	0.345**	0.429**	1	
Higher Education	-0.140	0.275*	0.282*	0.385**	0.487***	0.545***	0.613***	1

$^{*}\ p < 0.05,\ ^{**}\ p < 0.01,\ ^{***}\ p < 0.001$

	(1)							
	Change FOR	Log CP	Log WPop	Log GDPpc	FOR 2008	Binary oil	Govt spend	Bartik
Change FOR	1							
Log CP	0.433***	1						
Log WPop	0.221	0.800***	1					
Log GDPpc	0.391**	0.214	0.157	1				
FOR 2008	0.362**	0.558***	0.678***	0.485***	1			
Binary oil	0.525***	0.0240	-0.117	0.714***	0.194	1		
Govt spend	0.337**	-0.0966	-0.174	0.376**	-0.0223	0.553***	1	
Bartik	0.457***	0.534***	0.587***	0.591***	0.941***	0.397**	0.174	1

$^{*}\ p < 0.05,\ ^{**}\ p < 0.01,\ ^{***}\ p < 0.001$

	(1)							
	Change FOR	Log CP	Log WPop	Log GDPpc	FOR 2008	Log CP x Binary oil	Log CP x Govt spending	Log CP x Bartik
Change FOR	1							
Log CP	0.433***	1						
Log WPop	0.221	0.800***	1					
Log GDPpc	0.391**	0.214	0.157	1				
FOR 2008	0.362**	0.558***	0.678***	0.485***	1			
Log CP x Binary oil	-0.513***	-0.00731	0.130	-0.730***	-0.182	1		
Log CP x Govt spending	-0.244	0.248	0.293*	-0.325*	0.115	0.535***	1	
Log CP x Bartik	-0.396**	-0.346**	-0.413***	-0.612***	-0.903***	0.425***	0.156	1

$^{*}\ p < 0.05,\ ^{**}\ p < 0.01,\ ^{***}\ p < 0.001$

Table 5: Correlation Tables. Variables included in all tables are: the change in formality rate, log of the complexity potential, log working age population, log GDP per capita and the formal employment rate in the base year 2008. The first table also includes variables for quality of infrastructure, institutions and higher education. The second table has a binary variable for oil production in cities, government spending per capita and sectoral demand (Bartik). The third table includes these latter variables interacted with log complexity potential.

Industry complexity is measured empirically via looking at how diverse the cities are that have the industry (cities which are active in many industries tend to incubate complex industries), and how rare the industries those cities have are (complex industries tend to be rare, so this is a measure of how many rare industries these cities tend to have). Hence, it is computed by calculating the average diversity of cities that make a specific product, and the average ubiquity of the other industries present in those cities. An iterative sequence is formed under the 'Method of Reflections'.

Mathematically, if M is a matrix with entry $M_{c,i} = 1$ if city i has RCA in industry i, the diversity of city c and ubiquity of industry i are defined as

$$k_{c,0} = \sum_i M_{c,i} \text{ and } k_{i,0} = \sum_c M_{c,i}$$

Then the average diversity of city c and analogously the average ubiquity of industry i may be expressed as:

$$k_{c,1} = \frac{1}{k_{c,0}} \sum_i \hat{M}_{c,i} k_{i,0} \text{ and } k_{i,1} = \frac{1}{k_{i,0}} \sum_c \hat{M}_{c,i} k_{c,0}.$$

Continuing the iteration to step n, we reach a pair of expressions:

$$k_{c,n} = \frac{1}{k_{c,0}} \sum_i \hat{M}_{c,i} k_{i,n-1} \text{ and } k_{i,n} = \frac{1}{k_{i,0}} \sum_c \hat{M}_{c,i} k_{c,n-1}$$

which, via substitution, can be expressed in closed form for industry i:

$$k_{i,n} = \sum_{i'} \tilde{M}_{ii'} k_{i',n-2}$$

with entries of matrix $\tilde{M}$:

$$\tilde{M}_{ii'} = \sum_c \frac{M_{c,i} M_{c,i'}}{k_{c,0} k_{i,0}}$$

Hence, if $\mathbf{k}_n$ is a vector whose ith element is $k_{i,n}$ then:

$$\mathbf{k}_n = \tilde{M} \mathbf{k}_{n-2}$$

and the application of eigenvalue methods enables us to obtain the long-run solution. In particular, the Industry Complexity (C_i) is the second largest eigenvector of $\tilde{M}$.