Why do industries coagglomerate? How Marshallian externalities differ by industry and have evolved over time

Dario Diodato[1], Frank Neffke[1], and Neave O'Clery[2]

[1]Center for International Development, Harvard University - 79 JFK street, 02138 Cambridge MA, USA
[2]Mathematical Institute, University of Oxford, Woodstock Road, OX2 6GG Oxford

Abstract

The fact that firms benefit from close proximity to other firms with which they can exchange inputs, skilled labor or know-how helps explain why many industrial clusters are so successful. Studying the evolution of coagglomeration patterns, we show that which type of agglomeration benefits firms has drastically changed over the course of a century and differs markedly across industries. Whereas, at the beginning of the twentieth century, industries tended to colocate with their value chain partners, in more recent decades the importance of this channels has declined and colocation seems to be driven more by similarities industries' skill requirements. By calculating industry-specific Marshallian agglomeration forces, we are able to show that, nowadays, skill-sharing is the most salient motive in location choices of services, whereas value chain linkages still explain much of the colocation patterns in manufacturing. Moreover, the estimated degrees to which labor and input-output linkages are reflected in an industry's coagglomeration patterns help improve predictions of city-industry employment growth.

JEL classifications: J24, O14, R11
Keywords: Coagglomeration, Marshallian externalities, labor pooling, value chains, manufacturing, services, regional diversification

Acknowledgments: Dario Diodato acknowledges the financial support of NWO, Innovation Research Incentive Scheme (VIDI). Project number: 45211013. Frank Neffke is grateful to the MasterCard Center for Inclusive Growth & Financial Inclusion for their financial support. The authors would like to thank Ron Boschma, Ricardo Hausmann, David Rigby, Nicoletta Corrocher and the attendees at the Global Conference on Economic Geography 2015 in Oxford, the Geography of Innovation 2016 conference in Toulouse and the CID Growth Lab seminar for their useful comments and suggestions.

1 Introduction

In spite of congestion, elevated factor costs and the risk that trade secrets leak to competitors, firms in the same industry frequently locate close to one another (Ellison and Glaeser, 1999; Rosenthal and Strange, 2001). As a consequence, in many industries we observe pronounced geographical clusters. The existence of these clusters are often attributed to the presence of three different types of externalities, namely the sharing of inputs, labor and knowledge. However, all industries may not benefit equally from all agglomeration. Moreover, given the dramatic decline in transportation and communication costs and the progressive spatial fragmention of value chains in the previous century, it is likely that the balance of costs and benefits for industries to cluster has changed drastically. However, our understanding of how agglomeration externalities differ across industries and have changed over time is still limited. The aim of this paper is to assess the relative importance of different drivers of agglomeration across industries, and how the main forces behind agglomeration externalities have changed over the course of a century.

Marshall (1920) ascribed "the advantages which people following the same skilled trade get from near neighbourhood to one another" (p.225) to three different types of agglomeration externalities: the benefits of a large pool of skilled labor, easy access to local customers or suppliers and local knowledge spillovers.[1] However, in spite of the early recognition and ample subsequent research of this topic, the relative importance of each of these Marshallian externalities has fueled debate for over a century. In part, this is due to the so-called "Marshallian equivalence" (Duranton and Puga, 2004): all three Marshallian agglomeration theories yield the same prediction for the spatial distribution of an industry, namely, that, because they generate benefits for one another, economic establishments engaged in similar activities will tend to agglomerate. This confluence of agglomeration benefits makes it difficult to determine which theory carries most weight as an explanation for the observed tendency of industries to concentrate in space.

A major stride forward is marked by Ellison, Glaeser, and Kerr (2010), henceforth EGK, who study not the agglomeration of *individual* industries but the coagglomeration of *pairs* of industries. The rationale for this is that industries that are similar along some dimensions may differ along others. For instance, whereas some industries will benefit from being colocated because they employ similar labor, others may colocate because of input-output or technological linkages. By analyzing the relationship between locational similarity (i.e., coagglomeration) and similarities that reflect different Marshallian agglomeration benefits, EGK disentangle the strength of three different types of Marshallian externalities. They find that input-output linkages are the most important explanation for why industries coagglomerate, closely followed by opportunities to share labor. Least empirical support is found for sharing know-how as a rationale for coagglomeration.

[1] Whereas theoretical models tend to divide agglomeration externalities into benefits from sharing, matching and learning in local economies (Duranton and Puga, 2004), most of the empirical literature on the topic categorizes Marshallian externalities as economies in transportation, coordination or communication when acquiring one of three factors: labor, (intermediate) capital goods or knowledge.

The effects EGK report represent averages across all industries, and, as such, may conceal marked differences among industries. For instance, whereas making musical instruments requires specialized workers with years of on-the-job training, workers in food-processing firms are often employed on short-term contracts through temporary work agencies, without much regard for their skills. Similarly, while car manufacturers often closely collaborate with their local suppliers (Morgan and Cooke, 1998), the principal inputs for steel mills, coal and iron ore, are acquired on anonymous exchanges with little need for buyer-supplier interaction. Finally, although knowledge spillovers may be important drivers of the clustering of biotechnology firms (Zucker et al., 1994), they will be less important in industries in which technology progresses less rapidly. Meta-studies reviewing the empirical literature on agglomeration externalities since the foundational papers by Glaeser et al. (1992) and Henderson et al. (1995) confirm the existence of considerable variation in empirical findings (Beaudry and Schiffauerova, 2009; Groot et al., 2015). We expect that these differences are in part driven by variation across industries in how much they rely on specific agglomeration forces, but also by changes in these forces over time.

Building on EGK, we explore whether this hypothesized heterogeneity in agglomeration benefits is expressed in industrial coagglomeration patterns. We start by replicating key parts of the original work by EGK, which was based on US manufacturing industries in the late 1980s and 1990s, using similar data for the 2000s. Mimicking EGK, we do so using both Ordinary Least Squares (OLS) estimation and Instrumental Variables (IV) based strategies that instrument inter-industry linkages among US industries by analogous measures constructed from data external to the US, mainly drawing from data for the Mexican economy. We confirm that labor linkages and input-output linkages are still more or less equally important explanations for coagglomeration, whereas evidence for technological (knowledge) spillovers is relatively weak. However, there are good reasons to expect that coagglomeration patterns in services will be different from those in manufacturing. First, unlike manufactured goods, services are often difficult to trade over large distances. Consequently, service firms will need to colocate with their customers. Second, services tend to be labor intensive, and their quality often crucially depends on face-to-face interactions between a firm's employees and its customers.[2] For both reasons, access to adequate human capital may be particularly important in services. This conjecture is confirmed when we extend the analysis to industries beyond manufacturing: the estimated effects of input-output linkages on coagglomeration are at least as large in services as in manufacturing, while labor-linkages are a much stronger factor in the coagglomeration of services than of manufacturing industries.[3]

Next, we allow Marshallian externalities to vary fully by industry. Doing so reveals an even wider variation in agglomeration effects. For some manufacturing activities, such as furniture and food production, coagglomeration patterns can be attributed neither to labor nor to input-output linkages.

[2]See, for instance, Kolko (1999) on the importance of agglomeration in services.

[3]For reasons explained more fully later on, unlike EGK, we disregard technological linkages and shared natural advantages as explanations for coagglomeration. Whereas our reservations about technological linkages are related to measurement issues – the most readily available source of information on such linkages, patents, are barely used in the majority of services – our reservations about natural advantages (as defined in EGK) are more fundamental: first, many natural advantages in EGK are actually man-made, second, EGK use local prices, which are endogenous to the agglomeration process, to quantify natural advantages.

For other manufacturing industries, such as those in electronics or in the pharmaceuticals & medical sector, coagglomeration patterns are primarily driven by input-output linkages. However, the greatest variety in agglomeration effects exists in services, where some industries, such as those in arts and culture, cluster along both externality channels, whereas in other services, like media and knowledge-intensive business services, labor pooling opportunities dominate coagglomeration.

Finally, we turn to changes in coagglomeration forces over time. The US economy and its geography have undergone major changes in the twentieth century. Urbanization rates rose from 40.5% to over 75% (Black and Henderson, 2003), work shifted from over 40% of male employment in agriculture,[4] to the current dominance of service jobs, and the US population has become increasingly educated, with the 18% secondary school enrollment rates of 1910 rising to 73% of youth enrolling in high schools by 1940 (Goldin and Katz, 2009, p. 195). Meanwhile, value chains started fragmenting through outsourcing and offshoring spurred by the drastic reduction in transportation and communication costs due to the roll-out of highway systems, airport and internet infrastructure. Using data on the coagglameration of industries in US states between 1910 and 2010, we explore how, combined, these processes have altered agglomeration forces. We find that the importance of value chain linkages in industrial coagglomeration patterns has strongly declined, while the relevance of labor pooling opportunities has – if not mildly increased – at least stayed constant. These findings suggest that the aforementioned historical changes indeed led to a shift in agglomeration forces. Arguably, whereas the contraction of space has eroded the importance of geographical proximity among customer and suppliers, the skill accumulation and specialization that came with the expansion of the education system and a deepening of the economy's knowledge base has amplified the relative benefits of sharing a common pool of labor.

By studying the heterogeneity of Marshallian externalities, we contribute first and foremost to the literature on agglomeration externalities in general, and to the fast-expanding strand of research that focuses on coagglomeration to untangle different agglomeration forces in specific (Ellison et al., 2010; Howard et al., 2015; Behrens, 2016; O'Sullivan and Strange, 2017). Our work is closely related to the work of Faggio et al. (2017), who study heterogeneity in Marshallian externalities in the UK. These authors find that agglomeration effects are particularly strong among small firms. Moreover, they report differences in agglomeration forces between high- and low-tech sectors, differences driven by the level of education of an industry's workforce, and – using quantile regressions – differences by industries' general tendency to coagglomerate. Although we also conclude that different industries experience different agglomeration forces, we arrive at this conclusion from a markedly different empirical approach. Whereas Faggio et al. (2017) explicitly look for differences related to the organization and technology of industries, we take a more agnostic approach and study heterogeneity in agglomeration forces in a nonparametric fashion. One finding that emerges from this is that labor sharing tends to be more important in services than in manufacturing, whereas the opposite holds for value-chain linkages. Given the growing importance of services in the economy and the fact that – using a century worth of data

[4] Authors' own calculations using IPUMS data.

– our historical analysis suggests that labor linkages are gaining in relative importance, labor sharing may soon become the dominant rationale for the cities of the future. However, at the same time, for some services coagglomeration are also strongly driven by value-chain linkages,[5] possibly reflecting the fact that the output of services is to a large extent nontraded and, therefore, associated with very high transportation costs. Moreover, we show that the heterogeneity in coagglomeration translates into differential growth of local industries. With this exploration of how the estimated differences in coagglomeration forces can help gauge how sensitive an industry's local growth rate is to a given type of agglomeration benefit, our paper also contributes to a growing literature on agglomeration externalities and local industry growth in urban economics (Glaeser and Kerr, 2009; Dauth, 2010; Jofre-Monseny et al., 2011, Hanlon and Miscio, 2017), cluster research (Porter, 2003; Delgado et al., 2010), and an emerging literature on related diversification in economic geography (Neffke et al., 2011; Hausmann et al., 2014).

2 Methodology

2.1 Data

To measure coagglomeration patterns in the current economy, we use datasets that describe employment by region-industry pair in the US and Mexico. For the US, these data are derived from the County Business Patterns (CBP) for the years 2003 and 2008.[6] Employment data for Mexico are taken from the economic censuses in 2003 and 2008.

We analyze these data at three different geographical levels: US counties (of which there are 3,190), metropolitan areas (922 including both Metro- and Micropolitan Statistical Areas) and states (51). In Mexico, our geographical units are 2,455 municipalities, 58 metropolitan areas and 32 states. Because we regard metropolitan areas (labelled 'cities' hereafter) as the most appropriate spatial unit for defining labor markets and economically integrated regions, we will focus the discussion on the results derived at this level, and report county and state level results as supporting evidence where practical. Furthermore, to define industries, we adjust the North American Industry Classification System (NAICS) to correct for small inconsistencies between US and Mexican data (see Appendix B). From the resulting list of 184 industries, we exclude nontraded industries with spatial distributions that are strongly driven by the distribution of population, such as retail, auto repair, construction work and elementary schools. Note, however, that we do not exclude extractive activities such as mining. Although these activities are obviously restricted in their location choice, it is still informative to see how

[5] This evidence is particularly salient when using a locational correlation index instead of the Ellison-Glaeser index of coagglomeration.

[6] The CBP data censor small industry-region cells. For such cells, we assign employment numbers following Holmes and Stevens (2004). For further details, we refer to Appendix G.

other industries choose to colocate with them. This leaves us with a set of 120 traded industries.[7] As a robustness check, we repeat the main analysis using the full sample of 184 industries. The outcomes, which are reported in Appendix F, are in line with those based on the restricted sample.

To analyze historical coagglomeration patterns, we use the US census samples provided by IPUMS USA (Ruggles et al., 2017) for the 1910-2010 period. Given the between 1% and 5% sampling rates, estimated industry-region employment data will be too noisy for small spatial units. Therefore, we restrict our analysis in this section to the coagglomeration of industries within US states. The industry classification in IPUMS (IND1990) distinguishes more than 200 activities, from which we select a set of 104 industries following the same logic as for CBP data.[8]

2.2 Dependent variable: coagglomeration

Our main object of interest is the degree to which industries coagglomerate. That is, to what extent do pairs of industries employ workers in the same regions? To quantify the tendency of industry i to coagglomerate with industry j, EGK propose the following measure:[9]

$$EG_{ij} = \frac{\sum_{r=1}^{R} (s_{ir} - x_r)(s_{jr} - x_r)}{1 - \sum_{r=1}^{R} x_r^2}. \tag{1}$$

where $s_{ir} = \dfrac{E_{ir}}{\sum_{r'=1}^{R} E_{ir'}}$ – with E_{ir} industry i's employment in region r – is the employment share of industry i in region r, while x_r is the mean of these shares in region r across all industries. Using a model of location choice, Ellison and Glaeser (1997,1999) motivate this index as a measure of the likelihood that establishments in industries i and j generate spillovers for one another. The index has the advantage that it should not be affected by the size distribution of establishments in an industry or by the granularity of spatial units. Following this logic, we calculate EG indices for all pairs of industries in the US.

The EG index is similar in spirit to a measure used by Porter (2003), who quantifies the coagglomeration of two industries as the correlation between industries' locational employment vectors:

$$LC_{ij} = corr(s_{ir}, s_{jr}), \tag{2}$$

[7] Traded industries represent about 38% of employment nationwide. That is, 100% of manufacturing, 100% of extractive activities, 0% of agriculture, utility or construction. Among services, we consider as traded a set of activities that constitute about 30% of employment in the tertiary sector. We list all excluded and included industries in Appendix D.

[8] More details on the construction of variables from IPUMS data can be found in Appendix B, while the list of included IPUMS industries is provided in Appendix D.

[9] This measure is equivalent to the coagglomeration index in Ellison and Glaeser (1999) for the coagglomeration of pairs of industries.

Hausmann et al. (2014) show that the locational correlation of industries i and j, LC_{ij}, can be interpreted as an estimate of the similarity of the industries' technology requirements in a Ricardian trade model. For brevity, we focus our discussion on the analysis that uses the EG index. The results using the LC index are reported in Appendix E, unless they lead to qualitatively different conclusions, in which case we discuss them in the main text.

2.3 Independent variables

EGK use four different inter-industry linkages to explain why industries may coagglomerate. The first three represent Marshallian externality channels: value-chain linkages, similarities in labor requirements and technological similarities among industries. The fourth linkage type captures the fact that the location of some industries reflects a need to access natural resources. Although these "natural advantages" do not represent Marshallian externalities, they may result in nonrandom coagglomeration of industries that require similar natural resources.

Input-output links

Value chains allow individual firms to specialize. However, such specialization also creates costs: intermediates need to be shipped between firms, and innovation efforts must be coordinated with suppliers (Richardson, 1972; Abdel-Rahman, 1996; Porter, 1998). Because the costs of transportation and coordination typically rise with distance, colocating different parts of a value chain can be an effective cost-reduction strategy.

We measure the strength of input-output relations between a pair of industries using the same indicator as EGK. That is, the input-output proximity of industries i and j is defined as the maximum relative importance of i as a customer or as a supplier of j and vice versa. Let IO_{ij} be an input-output matrix, i.e., IO_{ij} represents the value of goods and services that industry j sources from industry i. We now measure the proximity between i and j in terms of input-output linkages as

$$P_{ij}^{IO} = max\left(\frac{IO_{ij}}{\sum_k IO_{kj}}, \frac{IO_{ji}}{\sum_k IO_{kj}}, \frac{IO_{ij}}{\sum_k IO_{ik}}, \frac{IO_{ji}}{\sum_k IO_{ik}}\right) \tag{3}$$

To estimate current value-chain linkages, we use tables provided by the Bureau of Economic Analysis (BEA) for the year 2002 in the US and by the Mexican statistical office, Instituto Nacional de Estadística y Geografía (INEGI) for the year 2008 in Mexico. For the historical analyses, we use input-output tables provided by BEA for the period 1947-2012. A detailed description of the data processing and subsequent calculations is provided in Appendix B.

Labor market pooling

A large local pool of specialized labor benefits both firms and workers. First, larger pools of skilled workers (and firms that want to hire them) may result in a better matching of workers to firms (Helsley and Strange, 1990). Second, workers may demand a wage premium as a compensation for moving to regions that offer few alternative employment opportunities in case they lose their jobs. In contrast, having many firms and industries that can absorb one another's redundant workers acts as an implicit insurance scheme and may, therefore, lower wage demands (Marshall, 1920; Duranton and Puga, 2004).

Following EGK, we measure the extent to which two industries can draw from the same pool of workers using industry-occupation employment matrices. In particular, we compute correlation coefficients between the occupational employment of industries i and j, E_{io} and E_{jo}, where E_{io} represents the number of workers in occupation o employed (nationwide) in industry i:

$$P_{ij}^{L} = corr(E_{io}, E_{jo}). \tag{4}$$

For the US, we use the industry-occupation data for the year 2002 reported in the Occupational Employment Statistics (OES), while for Mexico, we use the Encuesta Nacional de Ocupación y Empleo (ENOE). Historical estimates are derived from the samples of the US census provided by IPUMS USA. For more details, we refer the reader to Appendix B.

Technological similarity

EGK estimate inter-industry technological similarity using cross-industry patent citations and the technology-flow matrix constructed in Scherer (1984). The latter yields the, conceptually, most attractive metric. To build the technology-flow matrix, science-degree students assessed for a large sample of patents which industries were most likely to be beneficiaries of the inventions described therein. Given that this exercise has not been repeated, we fall back to patent citations. Using the NBER patent citations dataset (Hall et al., 2001), we construct the overrepresentation of industries i and j in bilateral citation flows, and measure the extent to which industries exchange knowledge, P_{ij}^{T}, in a similar fashion as P^{IO}, as

$$P_{ij}^{T} = max\left(\frac{C_{ij}}{\sum_k C_{kj}}, \frac{C_{ji}}{\sum_k C_{kj}}, \frac{C_{ij}}{\sum_k C_{ik}}, \frac{C_{ji}}{\sum_k C_{ik}}\right) \tag{5}$$

where C_{ij} is the number of citations from patents associated with industry i to patents associated with industry j. We estimate P_{ij}^{T} twice: once as a main covariate using patents by US inventors, and once as an instrument using patents by foreign inventors. Details of this procedure are provided in Appendix B.

Natural advantages

In some industries, firms' location choices are limited by factors that are driven by natural advantages. For instance, shipbuilding requires access to waterways and some tourism activities rely on the beauty of a place's natural environment. To correct for such factors, EGK calculate an index of how similar industries are in terms of their reliance on natural resources. This index is based on work by Ellison and Glaeser (1999) that predicts for each industry what share of its employment will locate in a given region from information on, on the one hand, the industry's intensity of use of a given resource, and, on the other hand, on the price of this resource in the region. Next, the predicted shares are used to estimate *predicted* coagglomeration patterns as a metric of similarity in natural advantages:

$$EG_{ij} = \frac{\sum_{r=1}^{R} \left(\hat{s}_{ir} - x_r\right)\left(\hat{s}_{jr} - x_r\right)}{1 - \sum_{r=1}^{R} x_r^2}. \tag{6}$$

Note that equation 6 is the same as equation 1 but for the substitution of $\hat{s}_{jr}$ for s_{jr}. For reasons explained hereafter, we will not repeat this exercise in full. Instead, we recalculate equation 6 using two different versions of $\hat{s}_{ir}$. First, to replicate the EGK results, we update the estimated employment shares by industry taken from the supplementary online material of EGK in such a way that they match the US economy's employment composition of 2003 and the NAICS classification system. Next, we create our own estimates to extend the analysis to services, which are disregarded by EGK. These estimates combine information on an industry's *use* of primary resources (using IO tables) and unskilled labor (using the industry's occupational employment vector) on the one hand, and the regional *availability* of these primary resources (proxied by the presence of the associated extractive industries) and unskilled labor (taken from BLS tables on urban employment by occupation), on the other hand. Both procedures are explained in more detail in Appendix B.

Discussion

As pointed out by EGK, a possible cause for concern is that the inter-industry similarity matrices are endogenous. That is, if two industries located close to each other due to an historical accident, this coagglomeration may have prompted them to adjust their production technologies, for instance, in such a way that they could use one another's outputs or skilled workers. Another concern is that similarity matrices are only imperfect proxies for the degree to which industries can exchange products, or use the same workers, technologies or natural advantages. Such measurement error would lead to a downward bias in our estimates. In line with EGK, we therefore also estimate the models in equations 7 and 8 using an IV approach. As instruments, we use analogously constructed variables based on information from outside the US economy.[10] Similar to the UK-based instruments in EGK, our instruments are

[10]For value chain and labor similarity linkages, we use data on the Mexican economy, for technological similarity we use patents by inventors who reside outside the US. Natural advantages, should, by definition, be exogenous and therefore

valid, as long as idiosyncratic patterns in the input-output, labor or technological linkages *outside* the US are exogenous to coagglomeration patterns *in* the US.

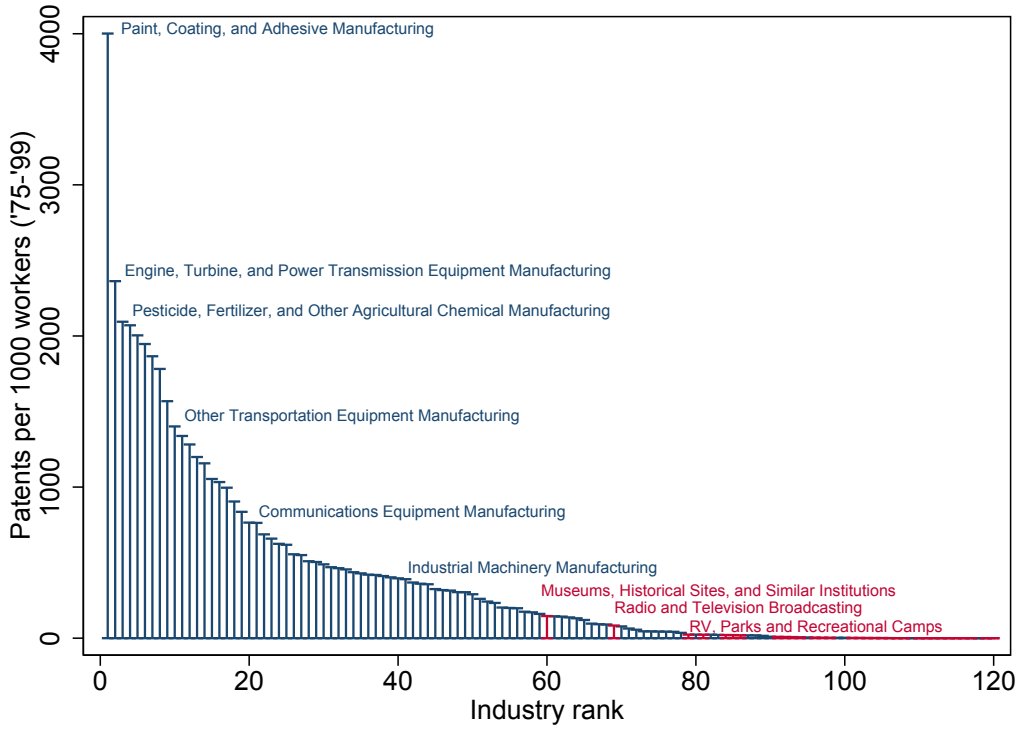
A second concern is related to the main goal of this paper. To explore the heterogeneity of Marshallian externalities across industries and over time, it is insufficient to generate proxies for each type of inter-industry linkage; we must also ensure that these proxies can be created with some accuracy for all industries. Although value-chain linkages and occupational similarity matrices can be constructed with relative ease for any industry, the same cannot be said for technological linkages because industries differ widely in how much they rely on patents to protect their intellectual property (Cockburn and Griliches, 1988). To illustrate this, Figure 1 shows the number of patents in the period 1975-1999 per 1,000 employees by industry. Clearly, although most manufacturing industries (in blue) produce at least some patents, most service industries do not patent any inventions at all. What is more, the technologies that get patented are arguably unrepresentative of the vast set of technologies firms use in their day-to-day processes.[11] As a result, the accuracy with which inter-industry technological similarity can be estimated from patent data will differ tremendously by industry. Because this heterogeneity in measurement error will translate mechanically into heterogeneity in the estimated effect of inter-industry knowledge spillovers on coagglomeration patterns, it compromises what our paper sets out to do.[12]

don't require instrumenting. Below, however, we will raise some doubts about the exogeneity of the natural advantage variable in EGK.

[11] EGK are well aware of this: "Our patent citation measure is a proxy for the importance of exchanging technology rather than a proxy for all forms of intellectual spillovers. Since our measures of idea sharing are weaker than our measures of input-output linkages, we anticipate their connection with coagglomeration to be weaker." (p. 1202)

[12] That is not to say that we believe that the other inter-industry proximity measures are unaffected by heteroscedastic measurement-errors. However, for reasons explained above, we believe that the variation in the precision with which we can characterize labor and input-output linkages is dwarfed by the variation in measurement error for technological linkages. Moreover, although the IV estimations will correct for some of the errors-in-variables bias, the scope for this is limited in the case of technological linkages, where the instrument is based on patent data as well.

Figure 1. *Patenting rates by industry*



Patents in the period 1975-1999 per 1,000 employees by industry. Industries sorted by patenting intensity. Blue bars refer to manufacturing industries, red bars to services.

A third and final concern arises due to the way natural advantages are measured in EGK. Although we agree that natural advantages may have driven some of the historical coagglomeration patterns in the US, the metric EGK use to correct for this has problems of its own. First, it is unclear whether prices should be considered when constructing a measure of natural advantages as in EGK. After all, local prices are endogenously determined by supply and demand in a region. For instance, low prices of a natural resource may make a location attractive to industries that heavily rely on this resource, or reflect that demand for them is low because such industries have shunned the region for other reasons. Second, the majority of production factors EGK list as natural resources are, in fact, man-made. For instance, EGK do not just label cheap electricity and abundant farmland and cattle as natural advantages. They also include low local manufacturing wages, educational attainment, unionization rates and population density. Most of these factors would also be reflected in similarities in industries' labor requirements. Moreover, a number of other resources, such as coal and timber – although arguably related to natural endowments – are more accurately regarded as outputs of

extractive industries. As such, they should be considered as parts of value chains. Upon closer inspection, the only truly exogenous factor in EGK appears to be a coastal dummy.[13] Therefore, the measure of similarities in natural-resource use is – if not endogenous – redundant given that it uses much of the same information as what is provided in input-output and labor-use matrices. For these reasons, after having compared our results to the ones in EGK, we will drop both technological and natural-resource-use similarities as explanations for coagglomeration patterns.

2.4 Descriptive statistics

Table 1. *Summary statistics.*

Variable	Obs	Mean	Std. Dev.	Min	Max
		United States			
EG index	7140	0.0001	0.0047	-0.0232	0.0846
LC index	7140	0.7166	0.1209	0.4854	0.9915
Input-output	7140	0.0124	0.0353	0.0000	0.6776
Labor	7140	0.6013	0.0997	0.4956	0.9919
Technology	7140	0.0082	0.0267	0.0000	1.2346
Technology (foreign)	7140	0.0082	0.0266	0.0000	1.1838
Nat.Advantage LC (original)	3828	0.9703	0.0357	0.7487	1.0000
Nat.Advantage EG (original)	3828	-0.0000	0.0011	-0.0060	0.0084
Nat.Advantage LC (new)	7140	0.9784	0.0262	0.7819	1.0000
Nat.Advantage EG (new)	7140	0.0000	0.0010	-0.0056	0.0058
		Mexico			
EG index	7140	0.0005	0.0477	-0.2579	0.5149
LC index	7140	0.7784	0.1657	0.4399	0.9994
Input-output	7140	0.0143	0.0503	0.0000	1.0000
Labor	7140	0.5565	0.1094	0.4828	1.0000

Summary statistics for coagglomeration (EG and LC metrics) and input-output and labor linkages for $120 \times (120-1)/2 = 7140$ *unique industry pairs for the US (top) and Mexico (bottom).*

Excluding the diagonal, there are 7,140 unique industry pairs in our sample of 120 industries. Table 1 contains descriptive statistics. It is interesting to note that, whereas means are similar for all variables in the US and Mexico, the dispersion is larger in Mexico, and sometimes substantially so. Although the greater dispersion in the Mexican variables may be structural, it could also mean that the variables constructed from US data are measured more accurately.

[13] However, this dummy has no significant impact on industries' locational choices (see Ellison and Glaeser (1999)).

Table A.6 reports correlation coefficients between the various dependent and independent variables. Interestingly, the correlations of the same variables in a different country are typically higher than of different variables in the same country. For instance, there is a 0.24 correlation between US labor and input-output similarities, whereas these variables exhibit correlations of over 0.5 with their Mexican counterparts. This is reassuring because it suggests that the similarity measures capture distinct, yet general, relations among industries.

2.5 Estimation framework

To infer the strength of agglomeration forces, we follow EGK and analyze the relationship between the coagglomeration patterns and various types of inter-industry linkages using the following econometric model:

$$EG_{ij} = \alpha + \sum_{rel} \beta^{rel} P_{ij}^{rel} + \epsilon_{ij}, \tag{7}$$

where EG_{ij} is the EG-index of coagglomeration for industries i and j, and $rel \in \{IO, L, T, NA\}$. Matrices P^{IO}, P^{L}, P^{T} and P^{NA} contain the inter-industry value-chain linkages, labor-requirement, technology and natural-advantage similarities defined in the previous sections.

While EGK only study coagglomeration among manufacturing industries, we extend the analysis to other sectors of the economy. We do so by expanding both the columns and rows of the EG matrix. Next, we split the rows into a manufacturing and a services section. Within these subsets, we study how industries coagglomerate with any other industry. That is, we study the coagglomeration patterns of a manufacturing industry (or of a service) with all other industries. Finally, note that each industry can coagglomerate with 119 other industries. Therefore, there are sufficiently many degrees of freedom to estimate agglomeration effects that vary freely by industry. We do so using the following two-way fixed effects model:

$$EG_{ij} = \alpha_i + \delta_j + \sum_{rel} \beta_i^{rel} P_{ij}^{rel} + \epsilon_{ij} \tag{8}$$

This regression yields vectors of industry-specific agglomeration effects, $\hat{\beta}^{rel}$. The elements of these vectors represent the extent to which a given inter-industry linkage type is expressed in an industry's coagglomeration patterns.

3 Empirical findings

3.1 Replication of EGK results

We begin our analysis by assuming that agglomeration effects are homogeneous across industries. That is, we assume that Marshallian externalities affect all industries in the same way. Table 2 replicates the results of EGK. The table shows the results of univariate OLS specifications for manufacturing industries.[14] To facilitate the interpretation of estimated effect sizes, all variables are rescaled such that they are expressed in units of standard deviations. We will apply this rescaling to all subsequent analyses as well. The left part of the table (columns 1 to 3) shows our own estimates; the original findings by EGK (table 3, p.1204) are shown on the right (columns 4 to 6).

Table 2. *OLS univariate regressions (EGK replication).*

	EG index (our estimates)			EG index (EGK estimates)		
	(1)	(2)	(3)	(4)	(5)	(6)
	state	city	county	state	city	county
Input-output	0.214	0.177	0.144	0.205	0.167	0.130
	(0.035)	(0.032)	(0.029)	(0.037)	(0.028)	(0.022)
Observations	3655	3655	3655	7381	7381	7381
R^2	0.060	0.060	0.047	0.042	0.028	0.017
Labor	0.181	0.164	0.195	0.180	0.106	0.082
	(0.018)	(0.018)	(0.016)	(0.014)	(0.016)	(0.013)
Observations	3655	3655	3655	7381	7381	7381
R^2	0.029	0.035	0.060	0.032	0.011	0.007
Technology	0.108	0.109	0.116	0.081	0.100	0.085
	(0.016)	(0.013)	(0.012)	(0.012)	(0.016)	(0.013)
Observations	3655	3655	3655	7381	7381	7381
R^2	0.015	0.022	0.030	0.007	0.010	0.007
Nat.Advantages	0.109	0.085	0.082	0.210	0.188	0.222
	(0.018)	(0.014)	(0.009)	(0.020)	(0.017)	(0.014)
Observations	3655	3655	3655	7381	7381	7381
R^2	0.014	0.012	0.014	0.044	0.036	0.049

Robust standard errors in parentheses. Columns 1-3 replicate the univariate regression results of Ellison et al. (2010) (shown in columns 4-6) using equivalent data for the US from 2003 for manufacturing industries and three levels of spatial aggregation: states, cities and counties. In all cases, the dependent variable is the pairwise EG coagglomeration

[14] EGK restrict their sample to the manufacturing sector. To enhance comparability, we define manufacturing industries in this table at the 4-digit level of the original US NAICS classification. In the remainder of this section, we switch to the (US-Mexican) harmonized NAICS classification.

index (equation 7).

In spite of the fact that our sample refers to a different period and more aggregated industry definitions (manufacturing codes are less detailed at the 4-digit NAICS level than at the 3-digit SIC level), most of our estimates are remarkably close to those of EGK. Moreover, we replicate EGK's main conclusions: all three externality channels are important, but the strongest impact on coagglomeration is recorded for value-chain linkages, the weakest for technological linkages. The main difference between our results and the ones in EGK is that we find larger effects of labor pooling at the city and county levels. Furthermore, natural advantages have a weaker influence according to our estimates.

Next, we extend the EGK framework in several directions. First, we expand the sample of industries by including services. This results in a sample of 120 industries in the (adjusted) NAICS classification, roughly two thirds (83) of which belong to the manufacturing and one third to the services sector. Doing so also forces us to abandon the natural-advantages variable taken from EGK. Instead, we use the alternative implementation described in the data section. Furthermore, we switch to a multivariate regression framework and add IV estimations, using the similarity matrices based on non-US data as instruments. Marshallian effects on coagglomeration measured by the EG index are reported in Table 3. For results using the LC index, we refer to Appendix E.

Adding services and moving to a multivariate setting changes some of the point estimates, but does not alter the main conclusions. The most substantive change is that the effect of technological similarities on coagglomeration becomes insignificant, with some of the point estimates even turning negative. This change is in part due to the fact that we now estimate the effects of all externality channels simultaneously. However, also in univariate regressions merely including services in the sample halves the technological similarity effect (see Table F.2 of Appendix F). This reinforces our concerns about the use of this variable outside the manufacturing sector. Furthermore, the effects of the new natural-advantages variable are over three times stronger than those reported in Table 3.[15,16] Moreover, adding this control variable cuts the estimated effect of labor linkages in half. This is unsurprising in light of our earlier discussion of the fact that what EGK call natural advantages includes characteristics of the local labor force. Given these concerns about the technology and natural-advantages measures, we henceforth drop these variables from the analysis.[17]

Using IV regressions, the labor channel regains some of its prominence in the state- and city-level es-

[15]To some extent, this increase seems to be driven by some exceptionally large locations. When we drop the twenty largest counties, ten largest cities, or five largest states, the coefficients drop to levels closer to the ones reported in Table 3.

[16]Note that, given that natural advantages should be exogenous, these variables enter the models of columns 4-6 uninstrumented.

[17]We also ran robustness tests, where we retain natural advantages as a control variable to correct for potential omitted variable bias. The only difference we find is that the estimated OLS effect of labor linkages in services is attenuated when using the EG-index as a dependent variable. All other estimates – which are available upon request – are consistent with what we report here.

timates.[18] Unlike EGK, who base their instruments on data from the UK, our instrumental variables are derived from Mexican data and, therefore, further test the robustness of the original EGK results. Overall, however, the discrepancies between OLS and IV estimates are minor, suggesting that endogeneity plays no major role in these estimations.[19] Balancing the trade-off between bias and efficiency, our preferred estimates are the ones from OLS regressions. At the city level, these estimates imply that a one-standard-deviation increase in similarity in the human capital requirements of two industries increases their coagglomeration by 0.1 standard deviations, whereas a one-standard-deviation increase in the strength of value-chain linkages increases their coagglomeration index by 0.16 standard deviations.

Table 3. *OLS and IV multivariate regressions on extended sample.*

	(1) OLS state	(2) OLS city	(3) OLS county	(4) IV state	(5) IV city	(6) IV county
	EG index - All traded					
Input-Output	0.227	0.149	0.125	0.244	0.128	0.105
	(0.042)	(0.026)	(0.028)	(0.048)	(0.033)	(0.036)
Labor	0.144	0.109	0.077	0.173	0.166	0.080
	(0.012)	(0.010)	(0.009)	(0.034)	(0.023)	(0.023)
Technology	-0.001	-0.028	-0.026	-0.003	-0.030	-0.024
	(0.010)	(0.008)	(0.009)	(0.011)	(0.008)	(0.008)
Nat.Advantages	0.239	0.323	0.272	0.223	0.300	0.274
	(0.018)	(0.023)	(0.031)	(0.021)	(0.023)	(0.031)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.136	0.127	0.081	0.135	0.124	0.081

Robust standard errors in parentheses.

3.2 Manufacturing versus services

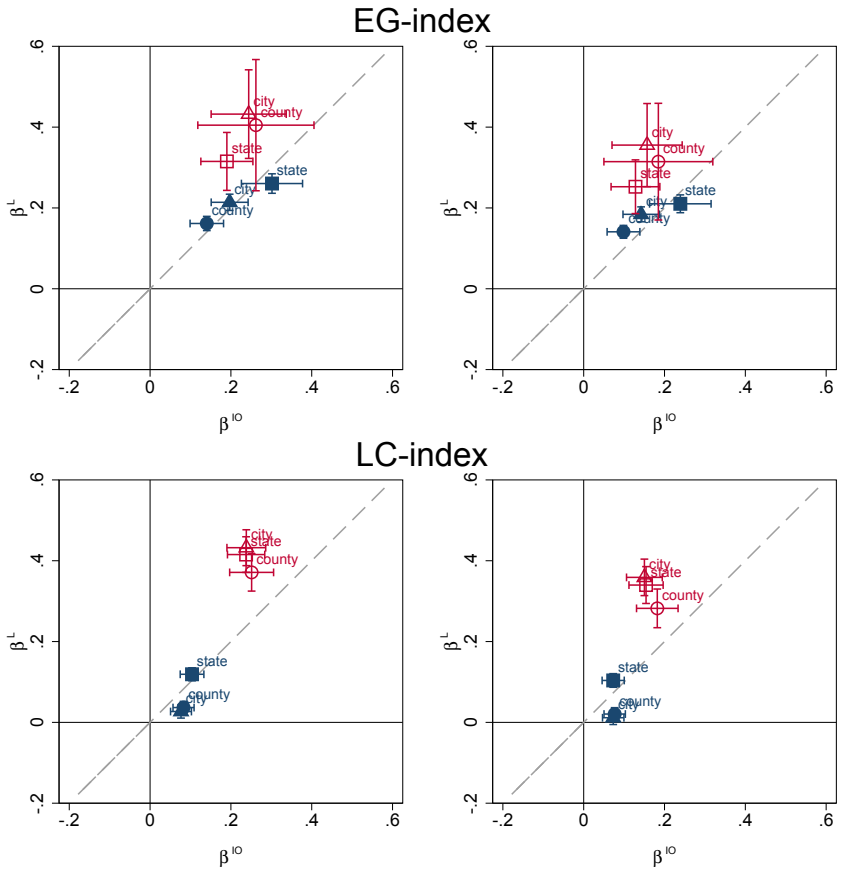
We start our exploration of sectoral differences in Marshallian externalities by splitting the sample into a manufacturing and a services subsample. Figure 3.2 shows the effects of labor market pooling and value chain linkages in each subsample. The different panels show estimated effects, together with their 95% confidence intervals, on coagglomeration patterns for manufacturing industries in blue and

[18] Note that this also occurs in a univariate analysis (Appendix F). Adding natural-advantages as a control variable thus also lowers estimated labor pooling effects in IV regressions.

[19] We formally test for endogeneity in each Marshallian channel using the difference-in-Sargan statistic, a heteroscedasticity robust variant of the standard Hausman test. This test fails to reject the null hypothesis that regressors are exogenous in all cases, except for the labor pooling channel in city-level estimates.

services in red. Estimates using the EG index are shown in the top panel, those using the LC index in the bottom panel. On the left, we plot univariate, on the right, multivariate estimates.[20]

Figure 2. *Coagglomeration effects, manufacturing versus services.*



Estimated effects on coagglomeration of services are plotted with red hollow markers; of manufacturing industries with blue solid markers. Labor pooling effects are depicted on the vertical axis, value chain effects on the horizontal axis, using the univariate estimations of equation 7. State-level estimates are marked by squares, city-level estimates by triangles and municipality-level estimates by circles. IV estimates use labor pooling and input-output linkages based on Mexican data as instruments. The crosshairs represent 95% confidence intervals based on robust standard errors. The dashed gray line is a 45 degree line.

[20]Tables F.3 and F.4 report the analogous coefficients derived from multivariate IV regressions. However, due to the collinear nature of the inter-industry proximity matrices our instruments are too weak to identify effects in services, where the Kleibergen-Paap statistic drops below 1.

The most striking aspect of Figure 2 is the difference between services and manufacturing. Above the (dashed) 45 degree line, labor-pooling effects exceed value-chain linkages as an explanation for coagglomeration. Below this line, the opposite holds. Whereas services typically appear above the 45 degree line, manufacturing industries tend to be located on, or below, the line. This means that coaggomeration patterns in services are more driven by labor-linkages than by value chains, whereas the opposite holds for manufacturing industries. Moreover, the point estimates for services always lie above and – in the bottom, LC-based, plots – to the right of the manufacturing estimates. This confirms our earlier contention that the nontraded nature and labor-intensity of services may make them particularly sensitive to Marshallian externalities: both externality channels are as or even more pronounced in the coagglomeration patterns of services than of manufacturing.[21]

3.3 Effect heterogeneity

To explore the observed sectoral heterogeneity in greater detail, we allow the effects of labor and value chain linkages to vary freely by industry, as in equation 8. To present outcomes in a more compact way, we average the resulting 120 industry-specific labor (β^L) and input-output (β^{IO}) effects for 27 broad subsectors and plot these averages in Figure 3.[22] Estimates for services are colored red, for manufacturing blue. To avoid cluttering the graph, we only show city-level estimates.

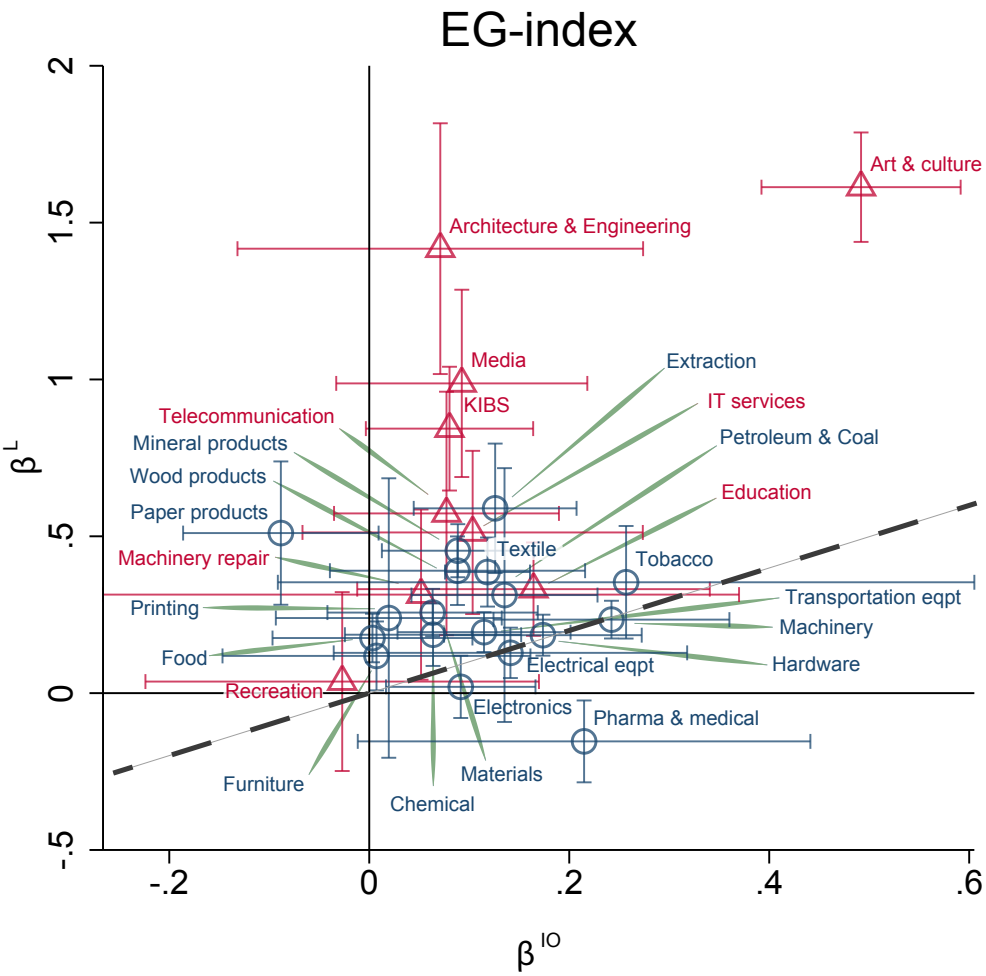
The general patterns roughly track those foreshadowed in Figure 2. Both types of inter-industry similarities tend to exert a greater influence on coagglomeration in the subsectors of services than of manufacturing. However, there is pronounced variation in agglomeration effects between (and, as shown in Appendix D, also within) subsectors, especially in labor-pooling effects. For instance, the labor channel's greatest impact on coagglomeration patterns is found in industries in *arts & culture*, *architecture & engineering*, *media* and *knowledge-intensive business services* (KIBS). In these subsectors, a one standard-deviation rise in the human capital similarity of two industries increases their EG index by up to 1.5 standard deviations.

Although less pronounced than in services, manufacturing industries also exhibit heterogeneous agglomeration effects. For instance, industries in *hardware* and *machinery* manufacturing tend to be driven by value chains *and* labor pooling relations. In contrast, *pharma & medical* and *electronics* industries coagglomerate with value-chain partners but not with industries that employ similar labor.

[21] Note that in multivariate regressions of the LC index (Appendix E), the effect of labor-linkages even turns slightly (although not significantly) negative in two of the manufacturing samples.

[22] We report the full list of 120 industry-specific estimates in Appendix D.

Figure 3. *Coagglomeration effects for 27 subsectors (EG-index)*



Average point estimates for industries in 27 subsectors (see Appendix D for results by industry) of labor pooling effects (vertical axis) and value chain effects (horizontal axis) on EG coagglomeration index, using OLS regression. Estimates for services are marked with red triangles, for manufacturing industries with blue circles. The crosshairs represent 95% confidence intervals for these averages, based on robust standard errors.

4 Marshallian externalities and regional diversification

Using the EGK framework, we have shown that the strength of labor-market-pooling and value-chain-based agglomeration externalities differs widely across industries. To test the predictive validity of

the underlying estimations, we now analyze whether the documented differences across industries manifest themselves in industries' local growth paths. If the measured effect-heterogeneity reflects the importance of agglomeration forces, it should not just affect how industries coagglomerate, but also help predict local growth patterns.

To investigate this, we build on a large and growing literature that focuses on the role of Marshallian externalities in the evolution of regions' industrial structures. According to this literature, inter-industry linkages are important for understanding diversification processes, because regions tend to branch out into new economic activities that build on a region's pre-existing strengths. This claim is supported by the fact that industries grow faster in regions with substantial employment in related industries (Porter, 2003; Greenstone et al., 2008; Delgado et al., 2010; Dauth, 2010; Jofre-Monseny et al., 2011; Neffke et al., 2011; Hausmann et al., 2014).

Most papers in this tradition implicitly assume that all industries benefit equally from inter-industry spillovers. Moreover, they typically do not distinguish among different types of inter-industry linkages. In this section, we use our estimated industry-specific effects on coagglomeration to enhance the econometric models of local-industry growth-rates used in this line of research. To do so, let the relatedness between industries i and j be measured by one of two proximity measures, $P_{ij}^{rel} \in \left\{P_{ij}^{L}, P_{ij}^{IO}\right\}$. We can now calculate the proximity-weighted employment for a region-industry, (i, r), in year t as:

$$E_{irt}^{rel} = \sum_{j} \frac{P_{ij}^{rel}}{\sum_{k \neq i} P_{ik}^{rel}} E_{jrt}.$$

This expression can be interpreted as the amount of related employment that already exists in the local economy, where what we call "related" depends on rel. For example, in the case of labor, E_{irt}^{L} is an index that reflects the size of the local workforce with skills and know-how that are relevant to industry i.

Letting G_{ir}^{03-08} refer to the logarithm of employment growth of industry i in region r between 2003 and 2008, i.e., $G_{ir} = \ln\left(E_{ir08}\right) - \ln\left(E_{ir03}\right)$ and restricting the sample to local industries with non-zero employment in 2003, we estimate the following:

$$G_{ir}^{03-08} = \delta \ln\left(E_{ir03}\right) + \sum_{rel} \gamma^{rel} \ln\left(E_{ir03}^{rel}\right) + \iota_i + \rho_r + \epsilon_{ir03} \tag{9}$$

where δ captures mean reversion effects, and ι_i and ρ_r are industry and region fixed effects.

The parameter of interest is γ^{rel}. It quantifies to what extent the growth in a local industry can be predicted from the amount of related activity that already exists in the local economy.

Table 4. *Growth in local industries – Intensive margin.*

	City-industry employment growth								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
lnE_{ir03}	-0.2003	-0.2037	-0.2003	-0.2037	-0.2007	-0.2040	-0.2045	-0.2046	-0.2054
	(0.0034)	(0.0034)	(0.0034)	(0.0034)	(0.0034)	(0.0034)	(0.0035)	(0.0035)	(0.0035)
lnE^{IO}_{ir03}	0.0283		0.0285		0.0282		0.0153	0.0156	0.0154
	(0.0045)		(0.0045)		(0.0045)		(0.0046)	(0.0046)	(0.0046)
lnE^{L}_{ir03}		0.0616		0.0619		0.0630	0.0575	0.0576	0.0590
		(0.0050)		(0.0050)		(0.0050)	(0.0051)	(0.0051)	(0.0051)
$lnE^{IO}_{ir03}\hat{\beta}^{IO}_{i,EG}$			0.0053					0.0045	
			(0.0022)					(0.0022)	
$lnE^{L}_{ir03}\hat{\beta}^{L}_{i,EG}$				0.0045				0.0043	
				(0.0020)				(0.0020)	
$lnE^{IO}_{ir03}\hat{\beta}^{IO}_{i,LC}$					0.0090				0.0098
					(0.0020)				(0.0020)
$lnE^{L}_{ir03}\hat{\beta}^{L}_{i,LC}$						0.0108			0.0117
						(0.0023)			(0.0023)
Obs.	38124	38124	38124	38124	38124	38124	38124	38124	38124
$Adj.R^2$	0.1302	0.1332	0.1303	0.1332	0.1305	0.1335	0.1334	0.1335	0.1342

Robust standard errors are reported in parentheses. The dependent variable is the logarithm of employment growth in local industries between 2003 and 2008. E_{ir03} *is the employment in industry* i*, city* r *and the year 2003.* E^{IO}_{ir03} *and* E^{L}_{ir03} *define the employment related to industry* i *according to input-output and labor linkages, respectively, in region* r *in 2003. The terms* $E^{IO}_{ir03}\hat{\beta}^{IO}_{i,EG}$ *and* $E^{L}_{ir03}\hat{\beta}^{L}_{i,EG}$ *refer to the interaction of related employment and the industry-specific effect-estimates of input-output or labor linkages in the coagglomeration regressions. Industry-specific coagglomeration effects are centered on their means and scaled by their standard deviations. All regressions include industry and region dummies.*

Table 4 contains results when using cities as the spatial unit of analysis. In line with prior studies (e.g., Delgado et al., 2010; Hausmann et al., 2014), we find significant and negative mean reversion effects, and positive effects of related employment, regardless of whether we measure relatedness in terms of labor or input-output linkages.

However, if the estimated $\hat{\beta}_i^{rel}$ parameters of section 3.3 capture any information on the importance of Marshallian channels, we would expect that industries with a higher $\hat{\beta}_i^{rel}$ will benefit more from the associated agglomeration externality when it comes to their growth.

To explore this, we augment 9 by interacting the amount of related employment in a city with $\hat{\beta}_i^{rel}$. To facilitate interpretation, $\hat{\beta}_i^{rel}$ values are expressed in units of standard deviations away from their respective means. This yields the following:[23]

$$G_{ir}^{03-08} = \delta \ln\left(E_{ir03}\right) + \gamma^{rel} \ln\left(E_{ir03}^{rel}\right) + \gamma_{\beta}^{rel} \hat{\beta}_i^{rel} \ln\left(E_{ir03}^{rel}\right) + \iota_i + \rho_r + \epsilon_{ir03} \tag{10}$$

where we expect γ_{β}^{rel} to be positive. Table 4 shows that this is indeed the case. The interaction effect, γ_{β}^{rel}, is always positive and significant.

The marginal effects of related employment are given by

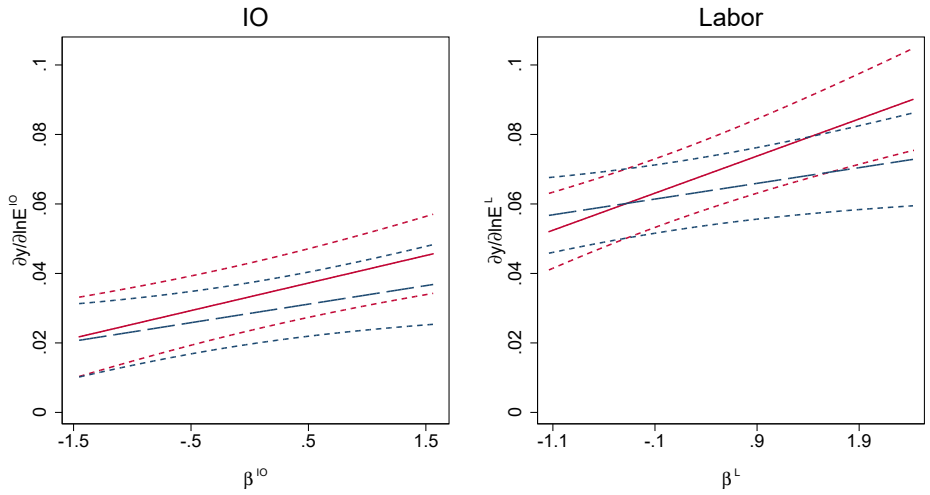
$$\frac{\partial G_{ir}^{03-08}}{\partial E_{ir}^{rel}} = \gamma^{rel} + \gamma_{\beta}^{rel} \hat{\beta}_i^{rel}.$$

Figure 4 plots these partial derivatives against $\hat{\beta}_i^{rel}$ on a range that reflects the span of coefficient estimates obtained in section 3.3. The interaction effects are strongest when the $\hat{\beta}_i^{rel}$ are based on regressions of the LC index. In this case, the elasticities for value-chain related employment run from as low as 2% to as high as 4%. The range of elasticities is slightly wider for employment that is related in terms of human-capital similarities, starting from a low 5% for industries with coagglomeration patterns that are relatively insensitive to human capital similarities and reaching over 8% where such similarities are strong drivers of coagglomeration. The fact that the industry-specific coagglomeration effects are also reflected in industries' local growth patterns validates the effect-heterogeneity reported in Section 3.3.[24] In Appendix C, we re-estimate the growth model in equation 9 for growth at the extensive margin, estimating the entry probability for local industries that do non yet exist in 2003. Here, the range of elasticities runs from about -1% and 2% for value-chain effects, whereas labor linkages have a constant effect of about 1%. This suggests that a pool of local suppliers or buyers may also attract new industries to a city, but only those industries whose location patterns are strongly affected by the spatial distribution of value-chain partners.

[23] Note that level effects of β_i^{rel} will be absorbed in ι_i.

[24] The industry-specific parameters estimated using the LC index outperform the ones using the EG index in these growth projections. This suggests that the LC coagglomeration metric may be a better indicator of inter-industry spillover potentials than the EG index. Although interesting in its own right, we will not pursue this issue further.

Figure 4. *Marginal effects of related employment.*



Marginal effects of related employment with 95% confidence intervals, computed as the partial derivative of growth with respect to related employment. The left panel depicts IO effects, the right panel labor-pooling effects. The horizontal axis plots β_{IO} and β_L, respectively, on a range that reflects the distribution of coefficient estimates obtained in section 3.3. Blue dashed lines use estimates from coagglomeration regressions using the EG-index, red solid lines using the LC-index.

5 Trends in Marshallian forces

So far, we have concentrated on cross-sectional heterogeneity in agglomeration externalities. However, in the past 100 years, the US economy underwent significant transformations, including urbanization, mass education, de-industrialization, tremendous reductions in transportation and communication costs, fragmentation and internationalization of supply chains, and a secular shift from agriculture and manufacturing activities to services. Have these phenomena collectively affected the relative strength of different Marshallian externalities? For instance, the increasing spatial fragmentation of supply chains that was facilitated by new transportation and communication technologies may have made spatial proximity to value-chain partners less crucial than it once was. Similarly, if the reliance on a highly skilled and specialized work force has increased over time, is this reflected in an increasing role

of labor similarity in determining coagglomeration? To shed light on these issues, we turn to historical data on the US economy.

5.1 Descriptive statistics

We start by describing how the agglomeration, occupational structures, and input-output structures of US industries have changed over time. Figure 5 shows weighted average statistics for industries in the US economy across decades. To ensure that we are picking up changes in industry characteristics, and not in the economy's industrial composition, weights are kept constant at industries' average employment shares across all decades.

The left panel shows the evolution of the geographical concentration of industries. The graph plots the weighted average effective number of US states in which an industry is active, as expressed by the inverse of the Herfindahl-Hirschman Index (HHI) of an industry's employment shares across states.[25] Until the 1980s, on average, industries have spread their employment more evenly across space. However, since then, the curve has flattened, indicating that this deconcentration process has ceased.

The center graph shows how the occupational structure of industries has changed in the same period. The blue line depicts the effective number of occupations per industry, the red line shows the (weighted) average similarity of industries' occupational structures to the occupational composition of the US economy as a whole.[26] The effective number of occupations per industry rises over the decades – again, until roughly the 1980s – suggesting that the division of labor within industries increases.[27] However, at the same time, this division of labor has become more uniform, with industries' occupational compositions becoming more similar to the economy's overall employment composition, suggesting that workers can be shared across a wider set of industries.

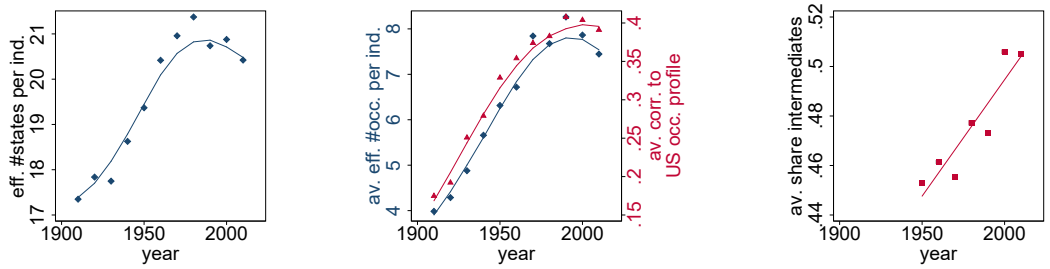
The final graph shows how value chains have developed over time. It plots the value of intermediates that industries use as a share of total output. Because input-output data are not available before 1947, this graph is limited to the period from 1950 to 2010. In spite of this, in line with the notion that value chains have become progressively fragmented, the trend towards greater outsourcing is clearly visible. Over the course of six decades, the weighted average input share rose by four pp, from 43% to 47%.

[25] If an industry's employment is spread out equally across N_s states, the inverse of the HHI, $\frac{1}{HHI_{it}} = \frac{1}{\sum_s \left(\frac{1}{E_{ist}}\right)^2}$, attains a value of $\frac{1}{\sum_s \left(\frac{1}{N_s}\right)^2} = \frac{1}{N_s \left(\frac{1}{N_s}\right)^2} = N_s$. Therefore, a given value of $\frac{1}{HHI_{it}}$ can be interpreted as the "effective" number of states in which an industry is present, i.e., the number of states needed to achieve this value had the industry's employment been distributed equally across them.

[26] The latter is calculated as $corr\left(E_{iot}, \sum_i E_{iot}\right)$, where the correlation is taken across occupations.

[27] The flattening of the curve towards the end of the period may reflect real changes in specialization patterns, or simply that the historical occupational classification system becomes less and less adequate to describe our modern economy's job structure.

Figure 5. *Historical trends, descriptive statistics*



Left: geographical concentration. Effective number of states per industry. Center: occupational composition. The left scale (blue diamonds) is the average effective number of occupations per industry, while the right scale (red triangles) is the average correlation to the US occupational profile. Right: Intermediates share. Average share of intermediates for all industries.

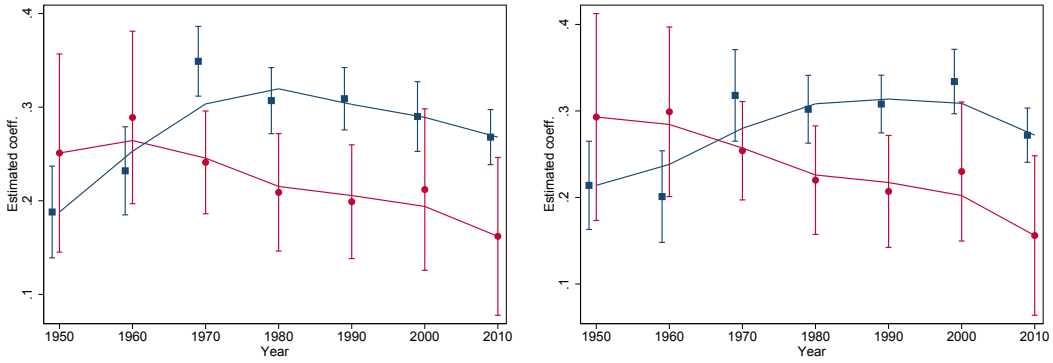
5.2 Marshallian externalities

Have the changes described in Figure 5 left a mark on agglomeration forces? In particular, we'd expect that increased outsourcing would diminish the importance of geographical proximity to value chain partners. Moreover, the documented rise in the division of labor, together with the fact that workers can be shared more widely across industries, may have raised the benefits of opportunities to share skilled workers.

Figure 6 shows the evolution of how labor and value-chain proximity have affected coagglomeration patterns in the period 1950-2010. The graph on the left shows parameter estimates using OLS regressions, while the graph on the right is based on IV estimates. Note, however, that we do not have data on the Mexican economy for this period. Therefore, we cannot construct the same instruments as in section 3. Instead, we instrument the proximities in one decade with the proximities in another decade.[28] This does not mitigate reverse causality problems, which, given our earlier estimates, we believe are relatively unimportant. However, doing so does help address the errors-in-variables bias that may result from using mismeasured proximities.

[28]For all decades but the first, we use as instruments the proximities measured in the preceding decade. In the first decade, 1950, we use the matrices for 1960 as instruments.

Figure 6. *Trends in Marshallian agglomeration forces - OLS (left) and IV (right)*



Estimated impact of input-output (red circles) and labor (blue squares) linkages on EG-coagglomeration in the period 1950-2010, with 95% confidence intervals based on robust standard errors. The lines represent LOWESS smooths.

Both graphs display clear downward trends in the effect of value-chain linkages on coagglomeration. This confirms the conjecture that locating close to value-chain partners has become less relevant with improving transportation and communication technologies. Instead, the labor-pooling effects has sharply increased in the early decades but than stabilized or even mildly declined towards the end of the period. Overall, however, in line with the notion that workers have become more specialized, labor pooling is more important nowadays than in the 1950s.

To test the existence of these trends more formally, we pool all data across decades and interact the labor and value-chain proximities with a linear time-trend. We express time in units of 100 years and – for reasons of consistency with the subsequent section – set the year 1910 to zero. Consequently, the level effects ($t = 0$) reflect the (extrapolated) situation in 1910, whereas the sum of level and interaction effect (i.e., $t = 1$) refers to the decade of 2010. As shown in Table 5, both trends are statistically significant. The estimated parameters imply that, in 1950, a one-standard-deviation increase in labor similarities increased industries coagglomeration by 0.259 ($0.205+0.5\times0.108$) standard deviations. This effect rises to 0.324 ($0.205 + 1.1 \times 0.108$) in 2010. In contrast, the effect of a one-standard-deviation stronger input-output linkages drops from a 0.256 to a 0.154 standard-deviations increase in the (EG-)coagglomeration of two industries.[29]

[29]In Appendix F, Table F.5 we add 2-way fixed effects to the estimation to control for any industry idiosyncrasies. This leads to even more pronounced time trends. Moreover, we find similar trends if we exclude services, as shown in Figure F.1 which uses a sample of only manufacturing industries. Consequently, the changes in labor-pooling and input-output effects are not simply the result of the US economy shifting towards more and more services.

Table 5. *Trends in Marshallian agglomeration forces*

EG	OLS (1)	OLS (2)	OLS (3)	IV (4)	IV (5)	IV (6)
Labor	0.205		0.188	0.210		0.193
	(0.025)		(0.025)	(0.026)		(0.026)
Labor × year	0.108		0.100	0.116		0.110
	(0.034)		(0.034)	(0.036)		(0.036)
Input-Output		0.341	0.315		0.384	0.364
		(0.032)	(0.031)		(0.034)	(0.033)
Input-Output × year		-0.170	-0.194		-0.211	-0.254
		(0.042)	(0.041)		(0.044)	(0.044)
R^2	0.05	0.02	0.07	0.05	0.02	0.07
N	31,757	31,757	31,757	31,757	31,757	31,757

Time is expressed in units of 100 years. Regression includes dummies for each decade. Robust standard errors in parentheses.

In principle, the IPUMS data allow us to measure coagglomeration and labor-similarity matrices for industries from 1910 onwards, albeit using a different industry classification system.[30] The only constraint is that we have no input-output tables for years before 1947. However, it turns out that the results of Table 5 are not driven by changes in the value-chain matrices. In fact, as shown in Appendix F, Table F.5, fixing the input-output matrices to the 1950 version, we find the same results as when using contemporaneous matrices. That is, input-output linkages have been constant enough to reliably estimate their impact on coagglomeration using a time-invariant matrix. We exploit this fact to extend our historical analysis backwards to 1910.

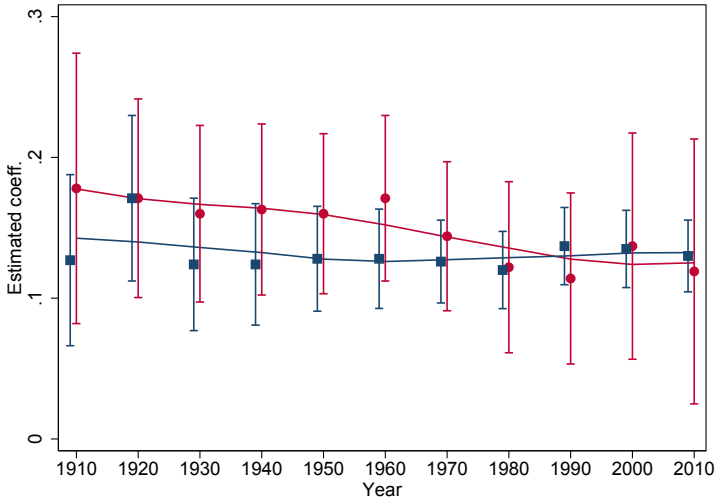
The results of this exercise are depicted in Figure 7.[31] They show that the decline of input-output linkages as a driver of agglomeration set in long before 1950, whereas, viewed over the entire period, there is no significant increase (nor decrease) in the effect of labor linkages.[32] Overall, these analyses confirm that Marshallian agglomeration forces have changed significantly over the course of a century. Moreover, if these estimated trends were to continue, value chains will lose their potency as an agglomerating force, leaving the skills of the workforce as the principal remaining driver of agglomeration in the cities of the future.

[30] Before 1910, the US census only recorded an individual's occupation, not her industry. The industry codes recorded in IPUMS in that period are imputed from these occupation codes, making them unusable for our purposes.

[31] Table F.5 of Appendix F shows the results of a regression that interacts externality channels with a time trend, analogous to Table 5.

[32] Closer inspection reveals that the mild positive trend between 1950 and 2010 that was reported in Table 5 is balanced by a negative trend between 1910 and 1950. This U-shaped pattern is weakly corroborated when we include an interaction of labor-linkages with a squared time-trend, which locates the minimum of the labor-pooling effect in 1967. However, the interacted coefficients are only jointly significant with the trend variables themselves.

Figure 7. *Trends in Marshallian agglomeration forces, long-run perspective (OLS)*



Estimated impact of input-output (red circles) and labor (blue squares) linkages on EG-coagglomeration in the period 1910-2010, with 95% confidence intervals based on robust standard errors. The lines represent LOWESS smooths.

6 Conclusion

Extending the work by Ellison et al. (2010), we find substantial variation in why industries coagglomerate. Whereas in some industries, firms tend to locate close to their value-chain partners, in others, they tend to locate near industries that share their labor requirements. In fact, the largest labor pooling effects are found in the coagglomeration patterns of services. Extreme examples are industries in *architecture & engineering*, *media* and *knowledge-intensive business services*. Value-chain effects are generally weaker and comparatively strong in manufacturing. However, overall we find that human capital similarities and, although less so, input-output linkages, have a stronger effect on the coagglomeration of services than of manufacturing industries.

This heterogeneity not only affects coagglomeration, but also helps improve predictions of industries' local growth rates: whereas there are ample studies that show that local industries tend to grow faster in regions with substantial employment in related industries, we are able to show that this relatedness matters more whenever the corresponding inter-industry proximity is more strongly expressed in an industry's coagglomeration patterns.

Finally, we find that the relative importance of different Marshallian agglomeration channels has changed over time. Using coagglomeration patterns in the period 1910-2010, we find that value-chain linkages were dominant in determining which industries coagglomerated at the beginning of the

twentieth century. However, their influence has weakened significantly since. At the same time, labor pooling effects have, on the whole, maintained their importance in coagglomeration patterns. Overall, we, therefore, conclude that whereas value chains were the main organizing principle of the spatial configuration of industries at the start of the twentieth century, their importance has dropped to a point where they have been overtaken by local pools of specialized labor as the main driver of industrial coagglomeration.

The fact that Marshallian forces operate more strongly on the locational patterns of services than of manufacturing has important implications for the future of cities. In spite of tremendous improvements in transportation and communication technologies – which should have allowed economic activities to become more dispersed – the increasing dominance of services in modern economies means that agglomeration externalities are likely to become more, not less, important. What is more, not only do services benefit particularly strongly from labor pooling externalities, labor pooling has also overtaken access to value chain partners as the dominant reason for why industries coagglomerate. As a result, our findings suggest that large cities increasingly derive their strength from the ease with which skills circulate in their economies.

References

Abdel-Rahman, Hesham M (1996), "When do cities specialize in production?" *Regional Science and Urban Economics*, 26, 1–22.

Autor, David and David Dorn (2013), "The growth of low-skill service jobs and the polarization of the US labor market." *The American Economic Review*, 103, 1553–1597.

Beaudry, Catherine and Andrea Schiffauerova (2009), "Who's right, Marshall or Jacobs? The localization versus urbanization debate." *Research Policy*, 38, 318–337.

Behrens, Kristian (2016), "Agglomeration and clusters: Tools and insights from coagglomeration patterns." *Canadian Journal of Economics/Revue canadienne d'économique*, 49, 1293–1339.

Black, Duncan and Vernon Henderson (2003), "Urban evolution in the USA." *Journal of Economic Geography*, 3, 343–372.

Cockburn, Iain and Zvi Griliches (1988), "The estimation and measurement of spillover effects of R&D investment-industry effects and appropriability measures in the stock market's valuation of R&D and patents." *The American Economic Review*, 78, 419–423.

Dauth, Wolfgang (2010), "Agglomeration and regional employment growth." Technical report, IAB discussion paper.

Delgado, Mercedes, Michael E Porter, and Scott Stern (2010), "Clusters and entrepreneurship." *Journal of Economic Geography*, 10, 495–518.

Duranton, Gilles and Diego Puga (2004), "Micro-foundations of urban agglomeration economies." *Handbook of Regional and Urban economics*, 4, 2063–2117.

Ellison, Glenn and Edward L Glaeser (1997), "Geographic concentration in US manufacturing industries: a dartboard approach." *Journal of Political Economy*, 105, 889–927.

Ellison, Glenn and Edward L Glaeser (1999), "The geographic concentration of industry: does natural advantage explain agglomeration?" *The American Economic Review*, 89, 311–316.

Ellison, Glenn, Edward L. Glaeser, and William R. Kerr (2010), "What causes industry agglomeration? Evidence from coagglomeration patterns." *The American Economic Review*, 100, 1195–1213.

Faggio, Giulia, Olmo Silva, and William C Strange (2017), "Heterogeneous agglomeration." *Review of Economics and Statistics*, 99, 80–94.

Glaeser, Edward, Hedi D Kallal, Jose Scheinkman, and Andrei Shleifer (1992), "Growth in cities." *Journal of Political Economy*, 100, 1126–52.

Glaeser, Edward L and William R Kerr (2009), "Local industrial conditions and entrepreneurship: how much of the spatial distribution can we explain?" *Journal of Economics & Management Strategy*, 18, 623–663.

Goldin, Claudia Dale and Lawrence F Katz (2009), *The race between education and technology*. Harvard University Press.

Goldschlag, Nathan, Travis J Lybbert, and Nikolas Jason Zolas (2016), "An algorithmic links with probabilities crosswalk for USPC and CPC patent classifications with an application towards industrial technology composition."

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti (2008), "Identifying agglomeration spillovers: evidence from million dollar plants." Technical report, National Bureau of Economic Research.

Groot, Henri LF, Jacques Poot, and Martijn J Smit (2015), "Which agglomeration externalities matter most and why?" *Journal of Economic Surveys*.

Hall, Bronwyn H, Adam B Jaffe, and Manuel Trajtenberg (2001), "The NBER patent citation data file: Lessons, insights and methodological tools." NBER Working Papers 8498, National Bureau of Economic Research.

Hanlon, W Walker and Antonio Miscio (2017), "Agglomeration: A long-run panel data approach." *Journal of Urban Economics*, 99, 1–14.

Hausmann, Ricardo, Cesar Hidalgo, Daniel P Stock, and Muhammed Ali Yildirim (2014), "Implied comparative advantage." *CID Working Paper*.

Helsley, Robert W and William C Strange (1990), "Matching and agglomeration economies in a system of cities." *Regional Science and Urban Economics*, 20, 189–212.

Henderson, Vernon, Ari Kuncoro, and Matt Turner (1995), "Industrial development in cities." *Journal of Political Economy*, 103, 1067–1090.

Holmes, Thomas J. and John J. Stevens (2004), "Spatial distribution of economic activities in North America." In *Handbook of Regional and Urban Economics* (J. V. Henderson and J. F. Thisse, eds.), volume 4 of *Handbook of Regional and Urban Economics*, chapter 63, 2797–2843, Elsevier.

Howard, Emma, Carol Newman, and Finn Tarp (2015), "Measuring industry coagglomeration and identifying the driving forces." *Journal of Economic Geography*, 16, 1055–1078.

Jofre-Monseny, Jordi, Raquel Marín-López, and Elisabet Viladecans-Marsal (2011), "The mechanisms of agglomeration: Evidence from the effect of inter-industry relations on the location of new firms." *Journal of Urban Economics*, 70, 61–74.

Kolko, Jed (1999), "Can I get some service here? Information technology, service industries, and the future of cities." *SSRN Working Paper Series*.

Marshall, Alfred (1920), *Principles of Economics*. Eighth Edition, 1982 reprint. MacMillan, London.

Morgan, Kevin and Philip Cooke (1998), "The associational economy: firms, regions, and innovation." *University of Illinois at Urbana-Champaign's Academy for Entrepreneurial Leadership Historical Research Reference in Entrepreneurship*.

Neffke, Frank, Martin Henning, and Ron Boschma (2011), "How do regions diversify over time? Industry relatedness and the development of new growth paths in regions." *Economic Geography*, 87, 237–265.

O'Sullivan, Arthur and William C Strange (2017), "The emergence of coagglomeration." *Journal of Economic Geography (forthcoming)*.

Porter, Michael (2003), "The economic performance of regions." *Regional Studies*, 37, 545–546.

Porter, Michael E (1998), "Cluster and the new economics of competition." *Harvard Business Review*, 77–90.

Richardson, George B (1972), "The organisation of industry." *The Economic Journal*, 82, 883–896.

Rosenthal, Stuart S and William C Strange (2001), "The determinants of agglomeration." *Journal of Urban Economics*, 50, 191–229.

Ruggles, Steven, Ronald Goeken Katie Genadek, Josiah Grover, and Matthew Sobek (2017), "Integrated public use microdata series: Version 7.0." *Integrated Public Use Microdata Series: Version 7.0 [dataset]. Minneapolis: University of Minnesota, 2017. https://doi.org/10.18128/D010.V7.0.*

Scherer, Frederic M (1984), "Using linked patent and r & d data to measure interindustry technology flows." In *R&D, Patents, and Productivity* (Zvi Griliches, ed.), 417–61, University of Chicago Press.

Zucker, Lynne G, Michael R Darby, and Marilynn B Brewer (1994), "Intellectual capital and the birth of US biotechnology enterprises." Technical report, National Bureau of Economic Research.

Appendix

A Descriptives

Table A.1. *Top-10 industry pairs, EG coagglomeration index.*

Top-10 co-location (EG index)		
industry i	industry j	value
Software Publishers	Aerospace Product Manufacturing	0.0276
Textile furnishings mills	Fiber, Yarn, and Thread Mills	0.0233
Motor Vehicle Manufacturing	Metalworking Machinery Manufacturing	0.0231
Tobacco Manufacturing	Fiber, Yarn, and Thread Mills	0.0225
Motor Vehicle Parts Manufacturing	Motor Vehicle Manufacturing	0.0212
Seafood Preparation and Packaging	Aerospace Product Manufacturing	0.0178
Motor Vehicle Manufacturing	Leather and Hide Tanning	0.0170
Software Publishers	Communications Equipment Manufacturing	0.0167
Software Publishers	Computer Equipment manufacturing	0.0165
Software Publishers	Seafood Preparation and Packaging	0.0163

Top-10 most coagglomerated industry pairs in terms of EG index (equation 1), using city-industry employment data for the US (County Business Patterns, 2003). The industry classification is based on NAICS 4-digit industries and includes 120 distinct industry codes containing both services and manufacturing. The top-10 cities have been excluded from the computation to reduce noise.

Table A.1 and shows the top 10 most coagglomerated industry pairs. When calculating coagglomeration scores, we exclude the ten largest cities to avoid that this top 10 gets cluttered by patterns generated by industries that locate in extremely large cities. Although all our estimations use coagglomeration scores calculated with the full set of cities, reported results (available on request) are robust to excluding the largest cities.

Table A.2. *Top-10 industry pairs, occupational similarity.*

Top-10 labor links		
industry i	industry j	value
Other Textile Product Mills	Other Apparel Manufacturing	0.9919
Other Apparel Manufacturing	Cut and Sew Apparel Manufacturing	0.9885
Other Textile Product Mills	Cut and Sew Apparel Manufacturing	0.9805
Other Transportation Equipment	Agri/Construction/Mining Machinery	0.9802
Motor Vehicle Parts Manufacturing	Hardware Manufacturing	0.9766
Other Transportation Equipment	Motor Vehicle Body Manufacturing	0.9760
Motor Vehicle Body Manufacturing	Heating/Cooling Equipment	0.9756
Other Machinery Manufacturing	Agri/Construction/Mining Machinery	0.9742
Household Appliance Manufacturing	Hardware Manufacturing	0.9737
Heating/Cooling Equipment	Hardware Manufacturing	0.9715

Top-10 industry pairs in terms of occupational similarity, computed as the correlation of industries' occupational employment shares (equation 4), using industry-occupation data for the year 2002 as reported in the US Occupational Employment Statistics (OES).

Table A.3. *Top-10 industry pairs, input-output links.*

Top-10 input-output links		
industry i	industry j	value
Petroleum and Coal Manufacturing	Oil and Gas Extraction	0.6776
Other Apparel Manufacturing	Cut and Sew Apparel Manufacturing	0.6001
Leather and Hide Tanning	Animal slaughtering and processing	0.5811
Motor Vehicle Parts Manufacturing	Motor Vehicle Manufacturing	0.5720
Motor Vehicle Manufacturing	Motor Vehicle Body Manufacturing	0.5351
Pulp, Paper, and Paperboard Mills	Paper Product Manufacturing	0.4617
Support Activities for Mining	Oil and Gas Extraction	0.4575
Motor Vehicle Manufacturing	Audio-Video Equipment Manufacturing	0.4343
Spectator Sports	Radio and Television Broadcasting	0.4269
Motor Vehicle Parts Manufacturing	Leather and Hide Tanning	0.4227

Top-10 industry pairs in terms of input-output linkages, defined as the maximum relative importance of one industry as a customer or as a supplier of the other and vice versa (equation 3), based on make-and-use tables provided by the US Bureau of Economic Analysis (BEA) for the year 2002.

Table A.4. *Top-10 industry pairs, technology links.*

Top-10 technology links		
industry i	industry j	value
Finance and Insurance	Computer Equipment manufacturing	1.2346
Finance and Insurance	Communications Equipment Manufacturing	0.8123
Other telecommunications	Communications Equipment Manufacturing	0.4905
Other telecommunications	Computer Equipment manufacturing	0.4622
Radio and Television Broadcasting	Communications Equipment Manufacturing	0.3605
Service Industry Machinery Manufacturing	Other telecommunications	0.3207
Management Consulting Services	Computer Equipment manufacturing	0.2913
Manufacturing of Magnetic and Optical Media	Finance and Insurance	0.2892
Management Consulting Services	Communications Equipment Manufacturing	0.2742
Computer Equipment manufacturing	Communications Equipment Manufacturing	0.2595

Top-10 industry pairs in terms technological linkages, defined as the maximum overrepresentation of citation flows between industries' patents (equation 5), using USPTO data for the years 1975-1999.

Table A.5. *Top-10 industry pairs, natural advantage similarities.*

Top-10 natural advantage (EG)		
industry i	industry j	value
Lime and Gypsum Product Manufacturing	Iron and Steel Mills	0.0084
Wood product Manufacturing	Sawmills and Wood Preservation	0.0078
Pulp, Paper, and Paperboard Mills	Iron and Steel Mills	0.0070
Sawmills and Wood Preservation	Other Wood Product Manufacturing	0.0067
Steel Product Manufacturing	Iron and Steel Mills	0.0067
Iron and Steel Mills	Fiber, Yarn, and Thread Mills	0.0062
Iron and Steel Mills	Alumina and Aluminum Production	0.0061
Sawmills and Wood Preservation	Iron and Steel Mills	0.0059
Petroleum and Coal Manufacturing	Iron and Steel Mills	0.0054
Wood product Manufacturing	Iron and Steel Mills	0.0053

Top-10 industry pairs in terms of shared use of natural advantages (equation B.3).

Table A.6. *Correlation coefficients.*

	EG (US)	LC (US)	IO (US)	Labor (US)	Tech (US)	N-EG (US)	N-LC (US)	EG (MX)	LC (MX)	IO (MX)	Labor (MX)	Tech (IV)
EG (US)	1.00											
LC (US)	0.27	1.00										
IO (US)	0.20	0.15	1.00									
Labor (US)	0.23	0.04	0.24	1.00								
Tech (US)	0.04	0.05	0.13	0.15	1.00							
NA EG (new)	0.31	0.03	0.15	0.37	0.11	1.00						
NA LC (new)	0.14	0.21	0.06	0.25	0.09	0.45	1.00					
EG (MX)	0.19	0.04	0.10	0.16	-0.02	0.16	0.07	1.00				
LC (MX)	0.04	0.49	0.06	0.05	-0.02	-0.00	0.23	0.20	1.00			
IO (MX)	0.13	0.10	0.53	0.21	0.11	0.15	0.03	0.14	0.05	1.00		
Labor (MX)	0.16	-0.02	0.29	0.54	0.13	0.17	0.07	0.19	-0.02	0.28	1.00	
Tech (IV)	0.04	0.05	0.14	0.15	0.99	0.11	0.09	-0.02	-0.02	0.11	0.14	1.00

Correlation table of pairwise inter-industry linkages. Variables are created using US (US) or Mexican (MX) data sources. NA EG (new) and NA LC (new) refer to the EG- and LC-based measures of shared natural advantages using the authors' own calculations (see section B). Tech (US) refers to technological similarity between industries measured by patent citations among patents by US inventors. Tech (IV) is an analogous measure using patents by non-US inventors.

B Variables construction

Industry classification

In section 3, US industries are classified according to the North American Industry Classification System (NAICS). Whereas data for 2003 use the NAICS2002 classification, 2008 data – which are used in section 4 to calculate 5-year growth rates – are recorded using NAICS2007 classes. Moreover, although the Mexican classification systems are based on NAICS as well, they are not fully harmonized with those in the US. Therefore, we aggregate industries where necessary and create a new composite industry classification, reducing the original 317 4-digit NAICS codes to 215 harmonized industry classes. After dropping industries with missing data, we are left with 184 industries. In our main analysis, we restrict this number further to 120 traded industries (see Appendix D)

To construct the coagglomeration measure for the historical analysis, we use industry-state employment matrices by aggregating US census samples provided by IPUMS (Ruggles et al., 2017). IPUMS provides harmonized industry codes across decades in two classification systems: IND1950, available for all decades and IND1990, which is available between 1950 and 2010. When studying the period 1950-2010, we use the IND1990 classification, with some small corrections proposed by Autor and Dorn (2013). When extending the analysis back to 1910, we use the somewhat coarser IND1950 classification.

Input-output

We construct inter-industry value-chain matrices using three different data sources. For the US, input-output linkages are based on make and use tables provided by the Bureau of Economic Analysis (BEA) for the year 2002. To merge these data to our coagglomeration data, we create a concordance between the 337 IO codes used by the BEA and our (harmonized) NAICS codes. Whenever one IO code corresponds to multiple NAICS codes, we distribute the flow of intermediates to or from the IO industry across the corresponding NAICS codes in proportion to each NAICS code's employment as reported in the CBP. Next, we use the resulting aggregated supply (make) matrix, S, and demand (use) matrix, U, to construct an IO-matrix using $IO = SD_S^{-1}U$, where D_S is a matrix with the column sums of S on its diagonal and all off-diagonal elements equal to zero.

Similar calculations are carried out for Mexico, using input-output data for the year 2008 provided by the Mexican statistical office, Instituto Nacional de Estadística y Geografía (INEGI). The ten strongest input-output linkages (based on US data) are shown in table A.3.

For the analyses in section 5, we use BEA historical IO tables from 1947 to 2012. Throughout this period, BEA uses a consistent industry classification that is closely linked to NAICS. Data are divided into three separated waves (1947-1962, 1963-1996, and 1997-2012). The level of detail of the classification system differs by wave. Therefore, we aggregate all data to the industrial classification

used in the wave of 1947 to 1962. We convert this classification to the IPUMS IND1990 and IND1950 classifications.[33] Finally, because the census data in IPUMS are collected once a decade, we compute 5-year averages of these input-output tables, centered on the years 1950, 1960, 1970, 1980, 1990, 2000 and 2010.

Labor linkages

The labor similarity variable used in section 3 is derived from the US Occupational Employment Statistics (OES) of the year 2002 as published by the Bureau of Labor Statistics. To instrument this variable, we aggregate the Mexican Encuesta Nacional de Ocupación y Empleo (ENOE) for the year 2005 at the industry-occupation level. The fact that the US and Mexico use different occupational classification systems does not complicate the calculation of occupational similarity matrices, because industries are recorded in NAICS codes in both data sources. However, because the ENOE uses a mix of 3- and 4-digit industry classes (consisting of 68 3-digit and 113 4-digit codes), we have to split some of the Mexican 3-digit industries into 4-digit codes. As a result, some industry pairs are mechanically attributed the same P^L values, and somewhat under 3% of off-diagonal elements are equal to 1.

For the historical analysis in section 5, we aggregate the IPUMS census samples to the industry-occupation level. For the period 1950-2010, we use the harmonized OCC1990 codes, whereas in the 1910-2010 period, we use the OCC1950 classification. With up to over 380 different codes, the OCC1990 classification is very detailed. Due to the limited size of the IPUMS samples, this risks imprecise estimates of industries' occupational structures, especially if industries are small. Therefore, we aggregate these codes into 82 occupational segments. In contrast, the OCC1950 industries are grouped by only a dozen occupational segments. Moreover, because, with about 200 different occupation codes in a typical decade, the OCC1950 classification itself is not as detailed as the OCC1990 classification, we do not aggregate occupations further in the 1910-2010 period.

Technological similarity

The technological similarity matrices are based on patent citations extracted from the NBER database (Hall et al., 2001). This data set contains 16 million patents filed between 1975 and 1999. We divide these patents into two samples, depending on whether they were filed by inventors residing inside or outside the US, dropping patents that list both, US and non-US inventors. After grouping the patents by their main patent class, we compute the total number of citations among these patent classes. The result is a 428×428 matrix of cross-patent-class citations. Using a concordance developed by Goldschlag et al. (2016), we map patent classes to the NAICS industry classification. Next, following EGK, we estimate the degree of industries' technological similarity as described in equation 5. The

[33] The concordance was constructed by hand and can be found in the code provided in the supplementary materials.

instrument for this metric is created by repeating this process with the sample of patents filed by non-US inventors.

Natural advantages

The estimate of natural advantages in EGK is based on the following nonlinear estimation:

$$s_{ir} = \frac{E_r^{\alpha_0} E_{r(m)}^{\alpha_1} exp(\mathbf{Xb})}{\sum_r E_r^{\alpha_0} E_{r(m)}^{\alpha_1} exp(\mathbf{Xb})} \tag{B.1}$$

where s_{ir} is the employment share of industry i in region r, while E_r and $E_{r(m)}$ are the total employment and manufacturing employment in region r.[34] The term $\mathbf{Xb}$ is a linear combination of natural advantage factors $k \in K$, such that $\mathbf{Xb} = \sum_k \beta_k y_{kr} z_{ki}$. In this expression, y_{kr} measures the abundance of k in region r, while z_{ki} quantifies how strongly industry i relies on factor k.

The estimated employment shares, $\hat{s}_{ir}$, are subsequently substituted for their actual counterparts in the EG-coagglomeration index:

$$EG_{ij} = \frac{\sum_{r=1}^{R} (\hat{s}_{ir} - x_r)(\hat{s}_{jr} - x_r)}{1 - \sum_{r=1}^{R} x_r^2}. \tag{B.2}$$

This expression can be seen as the *expected* coagglomeration of two industries based on their dependence on the natural advantages in K.

To replicate the natural-advantage estimates in EGK, we plug two different estimates of s_{ir} into (B.2). First, we retrieve EGK's own nonlinear estimates of $\hat{s}_{ir}$ from their supplementary material. Next, we convert these shares to the NAICS classification using the following formula

$$\tilde{s}_{i'r} = \frac{\sum_i (E_r(m)\hat{s}_{ir} \times T_{i,i'})}{\sum_{ir} (E_r(m)\hat{s}_{ir} \times T_{i,i'})} \tag{B.3}$$

where $E_r(m)$ is the manufacturing employment recorded in the CBP and $T_{i,i'}$ is a correspondence matrix that converts employment from SIC industry i to NAICS industry i'. That is, we calculate the employment in region r and industry i that EGK would have predicted in our data by computing $E_r(m)\hat{s}_{ir}$. This predicted employment, $\hat{E}_{ir}$, is still recorded in the SIC classification. We, thus, convert it to NAICS using the formula $\sum_i (\hat{E}_{ir} \times T_{i,i'})$. Next, we express this updated employment as a share of an industry's total US employment. This share, $\tilde{s}_{ir}$, is then used in equations 1 and 2 to compute the predicted EG and LC coagglomeration indices which feature as measures of similarity in natural advantages.

[34] Ellison and Glaeser (1999) use population, but we find total employment to be more appropriate.

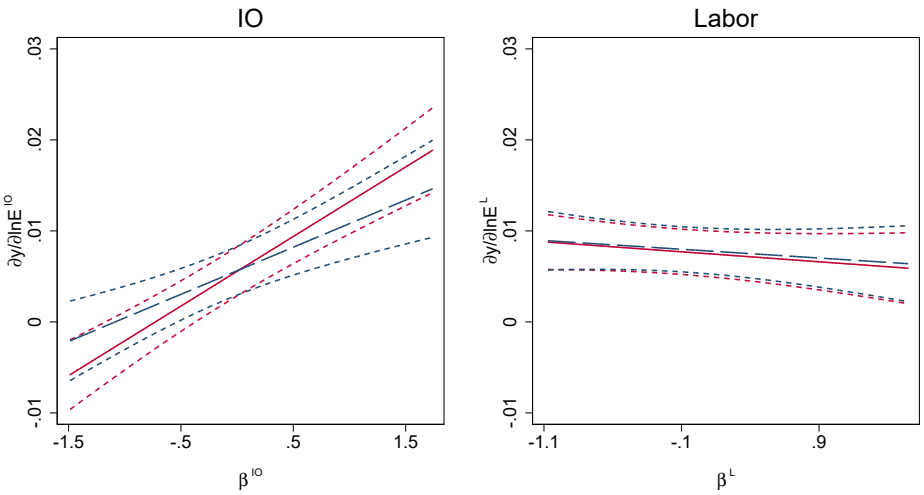
Since this procedure relies on the original estimates in EGK, services are excluded. To arrive at natural-advantage estimates for services, we would need the price information used in EGK, which we do not have. Instead, we follow a procedure that is close in spirit to the calculations in EGK, albeit with some differences. First, we use a different list of production factors. In particular, we use a combination of extractive industries, whose location is pinned down by a natural resource,[35] and unskilled labor categories.[36] Second, we change the way we measure y_{kr} and z_{ki}. For y_{kr}, which proxies the availability of a "natural" resource in a location, we either use the employment share of extractive industry k in region r, or, to proxy access to unskilled labor, the employment share of occupation k in region r. To proxy the intensity with which industry i relies on natural advantage k, z_{ki}, we use i's input share sourced from extractive industry k or the share of occupation k in the industry's overall employment.

[35]These five NAICS industries are 1141 (Fishing), 2111 (Oil and Gas Extraction), 2121 (Coal Mining), 2122 (Metal Ore Mining), 2123 (Nonmetallic Mineral Mining and Quarrying)

[36]We select six unskilled occupations (which we aggregate into one 'unskilled' factor): 35 (Food Preparation and Serving Related Occupations), 37 (Building and Grounds Cleaning and Maintenance Occupations), 47 (Construction and Extraction Occupations), 49 (Installation, Maintenance, and Repair Occupations), 51 (Production Occupations), 53 (Transportation and Material Moving Occupations)

C Growth at the extensive margin

Figure C.1. *Marginal effects of related employment – Extensive margin.*



Marginal effects of related employment with 95% confidence intervals, computed as the partial derivative of entry probability from Linear Probability Models (LPMs) with respect to related employment. The left panel depicts IO effects, the right panel labor-pooling effects. The horizontal axis plots β_{IO} and β_L, respectively, on a range that reflects the distribution of coefficient estimates obtained in section 3.3. Blue dashed lines were constructed with estimates from coagglomeration regressions using the EG-index, red solid lines using the LC-index.

Table C.1. *Growth in local industries.*

	City-industry appearances								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
lnE^{IO}_{ir03}	0.0054		0.0056		0.0061		0.0039	0.0041	0.0047
	(0.0014)		(0.0014)		(0.0014)		(0.0014)	(0.0014)	(0.0014)
lnE^{L}_{ir03}		0.0080		0.0079		0.0078	0.0072	0.0072	0.0069
		(0.0013)		(0.0013)		(0.0013)	(0.0013)	(0.0013)	(0.0013)
$lnE^{IO}_{ir03}\hat{\beta}^{IO}_{i,EG}$			0.0053					0.0053	
			(0.0013)					(0.0013)	
$lnE^{L}_{ir03}\hat{\beta}^{L}_{i,EG}$				-0.0010				-0.0004	
				(0.0011)				(0.0011)	
$lnE^{IO}_{ir03}\hat{\beta}^{IO}_{i,LC}$					0.0078				0.0077
					(0.0011)				(0.0011)
$lnE^{L}_{ir03}\hat{\beta}^{L}_{i,LC}$						-0.0011			-0.0009
						(0.0009)			(0.0009)
Obs.	43082	43082	43082	43082	43082	43082	43082	43082	43082
$Adj.R^2$	0.1162	0.1166	0.1166	0.1166	0.1175	0.1166	0.1167	0.1171	0.1180

Idem table 4, but the dependent variable is an indicator variable that takes value one if a local industry came into existence between 2003 and 2008. All independent variables are log-transformed.

D Industry lists

Table D.1. *Industry list with IO and labor coefficients (EG).*

	name	group	data avail.	traded	IO coef.	labor coef.
1	Crop production	.	0	0	.	.
2	Animal production	.	0	0	.	.
3	Aviculture	.	0	0	.	.
4	Mixed agriculture	.	0	0	.	.
5	Silviculture	.	0	0	.	.
6	Forestry	.	0	0	.	.
7	Logging	.	0	0	.	.
8	Fishing	.	0	0	.	.
9	Hunting and Trapping	.	0	0	.	.
10	Support Activities for Crop Production	.	0	0	.	.
11	Support Activities for Animal Production	.	0	0	.	.
12	Support Activities for Forestry	.	0	0	.	.
13	Oil and Gas Extraction	1	1	1	0.27	-0.26
14	Coal Mining	1	1	1	0.06	0.97
15	Metal Ore Mining	1	1	1	-0.02	0.82
16	Nonmetallic Mineral Mining and Quarrying	1	1	1	-0.04	0.47
17	Support Activities for Mining	1	1	1	0.36	0.95
18	Utilities	.	1	0	.	.
19	Utility System Construction	.	1	0	.	.
20	Land Subdivision	.	1	0	.	.
21	Highway, Street, and Bridge Construction	.	1	0	.	.
22	Heavy and Civil Engineering Construction	.	1	0	.	.
23	Building Exterior Contractors	.	1	0	.	.
24	Building Equipment Contractors	.	1	0	.	.
25	Building Finishing Contractors	.	1	0	.	.
26	Other Specialty Trade Contractors	.	1	0	.	.
27	Animal Food Manufacturing	2	1	1	0.01	0.24
28	Grain and Oilseed Milling	2	1	1	-0.09	0.35
29	Sugar Product Manufacturing	2	1	1	0.01	0.19
30	Fruit and Vegetable Preserving	2	1	1	-0.00	0.23
31	Dairy Product Manufacturing	2	1	1	-0.16	0.16
32	Animal slaughtering and processing	2	1	1	0.02	0.54
33	Seafood Preparation and Packaging	2	1	1	0.14	-0.06
34	Bakeries and Tortilla Manufacturing	2	1	1	0.08	-0.10
35	Other Food Manufacturing	2	1	1	0.01	0.07
36	Beverage Manufacturing	2	1	1	0.02	0.13
37	Tobacco Manufacturing	3	1	1	0.26	0.35
38	Fiber, Yarn, and Thread Mills	4	1	1	-0.11	2.10
39	Textile and Fabric Mills	4	1	1	0.28	-0.18
40	Textile furnishings mills	4	1	1	0.16	0.21
41	Other Textile Product Mills	4	1	1	-0.01	0.12
42	Apparel Knitting Mills	4	1	1	-0.04	0.69
43	Cut and Sew Apparel Manufacturing	4	1	1	0.31	0.21
44	Other Apparel Manufacturing	4	1	1	0.37	-0.09
45	Leather and Hide Tanning	4	1	1	0.10	0.38
46	Footwear Manufacturing	4	1	1	0.13	0.31
47	Other Leather Manufacturing	4	1	1	0.00	0.11
48	Sawmills and Wood Preservation	5	1	1	0.05	0.64
49	Wood product Manufacturing	5	1	1	0.11	0.31
50	Other Wood Product Manufacturing	5	1	1	0.10	0.22

	name	group	data avail.	traded	IO coef.	labor coef.
51	Pulp, Paper, and Paperboard Mills	6	1	1	-0.13	0.74
52	Paper Product Manufacturing	6	1	1	-0.05	0.28
53	Printing and Related Support Activities	7	1	1	0.02	0.24
54	Petroleum and Coal Manufacturing	8	1	1	0.14	0.31
55	Basic Chemical Manufacturing	9	1	1	0.04	0.36
56	Synthetic Fibers Manufacturing	9	1	1	0.04	0.43
57	Agricultural Chemical Manufacturing	9	1	1	0.01	0.35
58	Pharmaceutical and Medicine Manufacturing	10	1	1	0.30	-0.33
59	Paint, Coating, Adhesive Manufacturing	9	1	1	0.09	0.09
60	Soap, Cleaning Compound Manufacturing	9	1	1	0.17	-0.27
61	Other Chemical Product Manufacturing	9	1	1	0.03	0.16
62	Plastics Product Manufacturing	11	1	1	-0.02	0.20
63	Rubber Product Manufacturing	11	1	1	0.09	0.32
64	Clay Product and Refractory Manufacturing	11	1	1	0.11	0.40
65	Glass and Glass Product Manufacturing	11	1	1	0.06	0.26
66	Lime and Gypsum Product Manufacturing	11	1	1	0.07	0.11
67	Other Mineral Product Manufacturing	12	1	1	-0.02	0.46
68	Iron and Steel Mills	12	1	1	0.13	0.81
69	Steel Product Manufacturing	12	1	1	0.18	0.35
70	Alumina and Aluminum Production	12	1	1	0.04	0.36
71	Nonferrous Metal Production	12	1	1	0.07	0.33
72	Foundries	12	1	1	0.04	0.55
73	Forging and Stamping	12	1	1	0.19	0.32
74	Cutlery and Handtool Manufacturing	13	1	1	0.22	0.17
75	Structural Metals Manufacturing	13	1	1	0.02	0.10
76	Boiler, Tank, Container Manufacturing	13	1	1	0.09	0.28
77	Hardware Manufacturing	13	1	1	0.21	0.19
78	Spring and Wire Product Manufacturing	13	1	1	0.19	0.21
79	Screw, Nut, and Bolt Manufacturing	13	1	1	0.20	0.21
80	Coating, Engraving, Heat Treating	13	1	1	0.28	0.22
81	Other Metal Product Manufacturing	13	1	1	0.17	0.10
82	Agri/Construction/Mining Machinery	14	1	1	0.46	0.18
83	Industrial Machinery Manufacturing	14	1	1	0.09	0.17
84	Service Industry Machinery Manufacturing	14	1	1	0.18	0.08
85	Heating/Cooling Equipment	14	1	1	0.21	0.14
86	Metalworking Machinery Manufacturing	14	1	1	0.31	0.46
87	Engine, Turbine, Transmission Manufacturing	14	1	1	0.16	0.46
88	Other Machinery Manufacturing	14	1	1	0.28	0.15
89	Computer Equipment manufacturing	15	1	1	0.15	0.06
90	Audio-Video Equipment Manufacturing	15	1	1	-0.04	-0.10
91	Semiconductor Manufacturing	15	1	1	0.04	0.13
92	Communications Equipment Manufacturing	15	1	1	0.23	0.01
93	Manufacturing of Magnetic and Optical Media	15	1	1	0.07	0.00
94	Electric Lighting Equipment Manufacturing	16	1	1	0.10	0.01
95	Household Appliance Manufacturing	16	1	1	0.13	0.23
96	Electrical Equipment Manufacturing	16	1	1	0.20	0.15
97	Other Electrical Equipment Manufacturing	17	1	1	0.04	0.12
98	Motor Vehicle Manufacturing	17	1	1	0.11	0.13
99	Motor Vehicle Body Manufacturing	17	1	1	0.00	0.25
100	Motor Vehicle Parts Manufacturing	17	1	1	0.17	0.22
101	Aerospace Product Manufacturing	17	1	1	0.24	-0.08
102	Railroad Rolling Stock Manufacturing	17	1	1	0.24	0.47
103	Ship and Boat Building	17	1	1	0.01	0.27
104	Other Transportation Equipment	17	1	1	0.12	0.17
105	Household Furniture Manufacturing	18	1	1	-0.02	0.29

	name	group	data avail.	traded	IO coef.	labor coef.
106	Office Furniture Manufacturing	18	1	1	0.05	0.10
107	Other Furniture Related Manufacturing	18	1	1	-0.01	-0.04
108	Medical Supplies Manufacturing	10	1	1	0.13	0.02
109	Other Miscellaneous Manufacturing	99	1	1	0.09	-0.04
110	Wholesale Trade	.	1	0	.	.
111	Retail Trade	.	0	0	.	.
112	Scheduled Air Transportation	.	1	0	.	.
113	Nonscheduled Air Transportation	.	1	0	.	.
114	Rail transportation	.	0	0	.	.
115	Water Transportation	.	1	0	.	.
116	Inland Water Transportation	.	1	0	.	.
117	General Freight Trucking	.	1	0	.	.
118	Specialized Freight Trucking	.	0	0	.	.
119	Urban Transit Systems	.	1	0	.	.
120	Interurban Bus Transportation	.	1	0	.	.
121	School and Employee Bus	.	1	0	.	.
122	Charter Bus Industry	.	1	0	.	.
123	Taxi and Limousine Service	.	1	0	.	.
124	Pipeline Transportation of Crude Oil	.	0	0	.	.
125	Pipeline Transportation of Gas	.	1	0	.	.
126	Other Pipeline Transportation	.	1	0	.	.
127	Sightseeing Transportation, Land	.	1	0	.	.
128	Sightseeing Transportation, Water	.	1	0	.	.
129	Sightseeing Transportation, Other	.	1	0	.	.
130	Support for Air Transportation	.	1	0	.	.
131	Support for Rail Transportation	.	1	0	.	.
132	Support for Water Transportation	.	1	0	.	.
133	Support for Road Transportation	.	1	0	.	.
134	Freight Transportation Arrangement	.	1	0	.	.
135	Support Activities for Transportation	.	1	0	.	.
136	Postal Services	.	0	0	.	.
137	Couriers and express delivery services	.	1	0	.	.
138	Local Messengers and Local Delivery	.	0	0	.	.
139	Warehousing and Storage	.	1	0	.	.
140	Publishers	19	1	1	0.18	0.81
141	Software Publishers	21	1	1	0.17	0.61
142	Sound Recording Industries	25	1	1	1.91	1.71
143	Radio and Television Broadcasting	19	1	1	0.01	1.16
144	Satellite Telecommunications	20	1	1	0.11	0.47
145	Data Processing Services	21	1	1	0.04	0.42
146	Other telecommunications	20	1	1	0.05	0.68
147	Real estate and construction	.	1	0	.	.
148	Offices of Real Estate Agents	.	1	0	.	.
149	Activities Related to Real Estate	.	1	0	.	.
150	Automotive Equipment Rental	.	1	0	.	.
151	Consumer Goods Rental	.	1	0	.	.
152	General Rental Centers	.	0	0	.	.
153	Machinery and Equipment Rental	.	0	0	.	.
154	Lessors of Nonfinancial Intangible Assets	23	1	1	-0.20	1.11
155	Legal Services	23	1	1	0.13	2.58
156	Accounting Services	23	1	1	0.44	0.68
157	Architectural and Engineering Services	22	1	1	0.09	0.59
158	Specialized Design Services	22	1	1	0.05	2.25
159	Computer Systems Design	.	0	0	.	.
160	Scientific and R&D Services	.	0	0	.	.

	name	group	data avail.	traded	IO coef.	labor coef.
161	Advertising and Related Services	.	0	0	.	.
162	Professional and Scientific Services	.	0	0	.	.
163	Management of Companies and Enterprises	23	1	1	0.07	0.46
164	Office Administrative Services	23	1	1	-0.04	0.54
165	Facilities Support Services	23	1	1	0.07	0.71
166	Management Consulting Services	23	1	1	0.12	0.13
167	Business Support Services	24	1	1	0.04	0.21
168	Travel Arrangement Services	.	1	0	.	.
169	Investigation and security services	.	1	0	.	.
170	Services to Buildings and Dwellings	.	0	0	.	.
171	Other Support Services	.	0	0	.	.
172	Waste Treatment and Disposal	.	1	0	.	.
173	Elementary and Secondary Schools	.	1	0	.	.
174	Junior Colleges	24	1	1	0.32	0.57
175	Colleges, Universities, Professional Schools	24	1	1	0.07	0.38
176	Business Schools and Computer Training	24	1	1	0.27	0.39
177	Technical and Trade Schools	24	1	1	0.13	0.10
178	Other Schools and Instruction	.	1	0	.	.
179	Educational Support Services	.	1	0	.	.
180	Offices of Physicians	.	1	0	.	.
181	Offices of Dentists	.	1	0	.	.
182	Offices of Other Health Practitioners	.	0	0	.	.
183	Medical and Diagnostic Laboratories	.	0	0	.	.
184	Home Health Care Services	.	0	0	.	.
185	Other Ambulatory Health Care Services	.	1	0	.	.
186	General Medical and Surgical Hospitals	.	1	0	.	.
187	Psychiatric and Substance Abuse Hospitals	.	1	0	.	.
188	Specialty Hospitals	.	1	0	.	.
189	Nursing Care Facilities	.	1	0	.	.
190	Mental Health and Substance Abuse Facilities	.	1	0	.	.
191	Individual and Family Services	.	1	0	.	.
192	Community Food and Housing	.	1	0	.	.
193	Vocational Rehabilitation Services	.	1	0	.	.
194	Child Day Care Services	.	1	0	.	.
195	Performing Arts Companies	25	1	1	-0.01	3.41
196	Spectator Sports	25	1	1	0.03	0.17
197	Promoters of Performing Arts	25	1	1	0.23	0.42
198	Agents and Managers	25	1	1	0.85	2.96
199	Independent Artists	25	1	1	0.26	2.19
200	Museums and Historical Sites	25	1	1	0.18	0.44
201	Amusement Parks and Recreation Industry	26	1	1	0.16	0.15
202	Traveler Accommodation	26	1	1	-0.00	0.13
203	RV Parks and Recreational Camps	26	1	1	-0.24	-0.17
204	Residential Care Facilities	.	1	0	.	.
205	Restaurants	.	1	0	.	.
206	Special Food Services	.	1	0	.	.
207	Drinking Places (Alcoholic Beverages)	.	0	0	.	.
208	Automotive Repair and Maintenance	.	1	0	.	.
209	Machinery Repair and Maintenance	27	1	1	0.05	0.31
210	Household Goods Repair and Maintenance	.	1	0	.	.
211	Other Personal Services	.	1	0	.	.
212	Associations and Organizations	.	1	0	.	.
213	Household services	.	0	0	.	.
214	Other Public Services	.	1	0	.	.
215	Finance and Insurance	23	1	1	0.07	0.53

Table D.2. *Average point-estimates by subsector.*

group	name	IO coef.	labor coef.
1	Extraction	0.13	0.59
2	Food Manufacturing	0.00	0.18
3	Tobacco Manufacturing	0.26	0.35
4	Textile Manufacturing	0.12	0.39
5	Wood products	0.09	0.39
6	Paper products	-0.09	0.51
7	Printing	0.02	0.24
8	Petroleum and Coal Manufacturing	0.14	0.31
9	Chemical Manufacturing	0.06	0.19
10	Pharmaceutical and medical supply	0.21	-0.15
11	Materials	0.06	0.26
12	Mineral products manufacturing	0.09	0.45
13	Hardware Manufacturing	0.17	0.18
14	Machinery	0.24	0.23
15	Electronics	0.09	0.02
16	Electrical equipment manufacuting	0.14	0.13
17	Transportation Equipment	0.11	0.19
18	Forniture	0.01	0.12
19	Media	0.09	0.99
20	Telecommunication	0.08	0.57
21	IT services	0.10	0.51
22	Architecture and Engineering	0.07	1.42
23	Professional KIBS	0.08	0.84
24	Educational	0.16	0.33
25	Art and culture	0.49	1.61
26	Recreation	-0.03	0.04
27	Machinery repair	0.05	0.31
99	Other	0.09	-0.04

Industries used from IPUMS classifcations

IND1990={40 41 42 50 60 100 101 102 110 111 112 120 121 122 130 132 140 141 142 150 151 152 160 161 162 171 172 180 181 182 190 191 192 200 201 210 211 212 220 221 222 230 231 232 241 242 250 251 252 261 262 270 271 272 280 281 282 290 291 292 300 301 310 311 312 320 321 322 331 332 340 341 342 350 351 352 360 361 362 370 371 372 380 381 390 391 392 700 701 702 710 711 712 721 722 731 732 740 741 762 770 800 802 810 841 850 872 882 890 891 892 893}.

IND1950={206 216 226 236 239 306 307 308 309 316 317 318 319 326 336 337 338 346 347 348 356 357 358 367 376 377 378 379 386 387 388 399 406 407 408 409 416 417 418 419 426 429 436 437 438 439 446 448 449 456 457 458 459 466 467 468 469 476 477 478 487 488 489 499 716 726 736 746 756 806 807 808 856 857 858 859 879}.

E Analysis using Location Correlation

Table E.1. *Top-10 industry pairs by coagglomeration (Locational Correlation index).*

Top-10 co-location (LC index)		
industry i	industry j	value
Legal Services	Accounting Services	0.9712
Management Consulting Services	Accounting Services	0.9684
Legal Services	Finance and Insurance	0.9681
Screw, Nut, and Bolt Manufacturing	Coating, Engraving, Heat Treating	0.9636
Specialized Design Services	Accounting Services	0.9614
Specialized Design Services	Legal Services	0.9611
Publishers	Legal Services	0.9601
Management Consulting Services	Architectural and Engineering Services	0.9565
Finance and Insurance	Accounting Services	0.9565
Architectural and Engineering Services	Accounting Services	0.9552

Idem table A.1, but here, we show the top-10 industry pairs using the LC-based coagglomeration metric (see equation 2) using city-industry employment data for the US (County Business Patterns, 2003). The top-10 cities have been excluded from the computation to reduce noise.

Table E.1 and shows the top 10 most coagglomerated industry pairs. When calculating coagglomeration scores, we exclude the ten largest cities to avoid that this top 10 gets cluttered by patterns generated by industries that locate in extremely large cities. Although all our estimations use coagglomeration scores calculated with the full set of cities, reported results (available on request) are robust to excluding the largest cities.

Table E.2. *OLS and IV univariate regressions, extended sample.*

	(1) OLS state	(2) OLS city	(3) OLS county	(4) IV state	(5) IV city	(6) IV county
			LC index - All traded			
Input-Output	0.132	0.119	0.122	0.162	0.148	0.139
	(0.015)	(0.015)	(0.015)	(0.027)	(0.028)	(0.027)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.024	0.021	0.026	0.023	0.020	0.026
Labor	0.120	0.033	0.044	0.061	-0.025	0.005
	(0.008)	(0.008)	(0.008)	(0.017)	(0.016)	(0.015)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.024	0.002	0.004	0.018	-0.004	0.001
Technology	0.087	0.046	0.061	0.085	0.041	0.056
	(0.023)	(0.018)	(0.020)	(0.022)	(0.016)	(0.018)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.010	0.003	0.006	0.010	0.003	0.006
Nat.Advantages	0.677	0.487	0.477	2.414	2.380	3.058
	(0.039)	(0.033)	(0.031)	(0.207)	(0.219)	(0.237)
Observations	7140	7140	7140	3828	3828	3828
R^2	0.077	0.043	0.048	-0.575	-0.756	-1.342

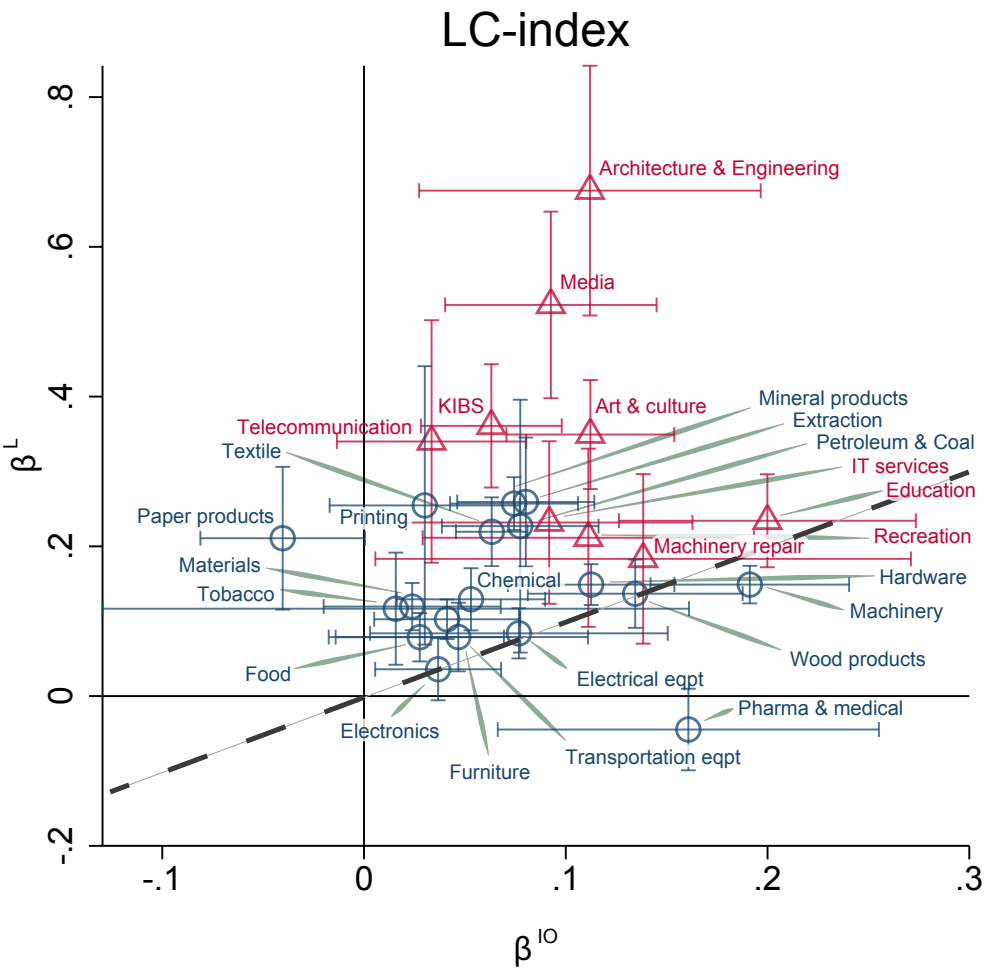
Robust standard errors in parentheses.

Table E.3. *OLS and IV multivariate regressions, extended sample, controlling for all four channels.*

	(1) OLS state	(2) OLS city	(3) OLS county	(4) IV state	(5) IV city	(6) IV county
			LC index - All traded			
Input-Output	0.103	0.115	0.115	0.167	0.201	0.171
	(0.014)	(0.015)	(0.014)	(0.033)	(0.037)	(0.034)
Labor	0.042	-0.034	-0.023	-0.059	-0.151	-0.107
	(0.009)	(0.009)	(0.009)	(0.024)	(0.024)	(0.023)
Technology	0.046	0.019	0.034	0.049	0.019	0.033
	(0.013)	(0.013)	(0.014)	(0.015)	(0.015)	(0.015)
Nat.Advantages	0.614	0.490	0.468	0.684	0.570	0.527
	(0.040)	(0.034)	(0.032)	(0.043)	(0.039)	(0.036)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.103	0.063	0.072	0.085	0.036	0.057

Robust standard errors in parentheses.

Figure E.1. *Coagglomeration effects for 27 subsectors (LC-index).*



Replication of Figure 3 using the LC index. The figure depicts labor pooling effects (vertical axis) and value chain effects (horizontal axis). Estimates for coagglomeration patterns of services are marked with red triangles, for manufacturing with blue circles.

The correlation coefficient between the vectors of estimated labor-similarity coefficients using the LC index, respectively EG index, is 0.85. For the estimated input-output proximity effects, the correlation is 0.44.

F Additional robustness analysis

Table F.1. *OLS and IV univariate regressions by sector (EG-index).*

	EG index					
	(1) OLS state	(2) OLS city	(3) OLS county	(4) IV state	(5) IV city	(6) IV county
			All industries			
Input-output	0.174	0.138	0.112	0.237	0.172	0.132
	(0.023)	(0.015)	(0.015)	(0.027)	(0.020)	(0.019)
Observations	16836	16836	16836	16836	16836	16836
R^2	0.033	0.022	0.015	0.028	0.020	0.015
Labor	0.201	0.175	0.129	0.322	0.282	0.191
	(0.009)	(0.007)	(0.007)	(0.023)	(0.016)	(0.015)
Observations	16836	16836	16836	16836	16836	16836
R^2	0.042	0.034	0.019	0.027	0.021	0.015
			Manufacturing			
Input-output	0.239	0.161	0.115	0.293	0.199	0.135
	(0.028)	(0.018)	(0.016)	(0.036)	(0.024)	(0.019)
Observations	11786	11786	11786	11786	11786	11786
R^2	0.048	0.025	0.017	0.045	0.024	0.017
Labor	0.229	0.191	0.121	0.302	0.254	0.137
	(0.010)	(0.007)	(0.006)	(0.025)	(0.015)	(0.011)
Observations	11786	11786	11786	11786	11786	11786
R^2	0.061	0.049	0.027	0.055	0.044	0.026
			Services			
Input-output	0.127	0.171	0.175	0.209	0.290	0.272
	(0.019)	(0.028)	(0.041)	(0.038)	(0.053)	(0.078)
Observations	5360	5360	5360	5360	5360	5360
R^2	0.015	0.016	0.013	0.009	0.008	0.009
Labor	0.213	0.275	0.294	0.421	0.507	0.515
	(0.024)	(0.037)	(0.053)	(0.057)	(0.075)	(0.100)
Observations	5360	5360	5360	5360	5360	5360
R^2	0.019	0.018	0.016	0.001	0.005	0.007

This table contains robustness checks, using all 184 industries as the "destination" industry. Robust standard errors in parentheses.

Table F.2. *OLS and IV univariate regressions, extended sample.*

	(1) OLS state	(2) OLS city	(3) OLS county	(4) IV state	(5) IV city	(6) IV county
			EG index - All traded			
Input-Output	0.295	0.215	0.176	0.372	0.267	0.201
	(0.041)	(0.025)	(0.028)	(0.045)	(0.032)	(0.032)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.069	0.038	0.025	0.064	0.036	0.024
Labor	0.261	0.229	0.178	0.347	0.296	0.194
	(0.012)	(0.011)	(0.011)	(0.031)	(0.020)	(0.018)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.065	0.052	0.030	0.058	0.048	0.030
Technology	0.077	0.041	0.030	0.083	0.046	0.031
	(0.029)	(0.021)	(0.015)	(0.030)	(0.022)	(0.015)
Observations	7140	7140	7140	7140	7140	7140
R^2	0.005	0.001	0.001	0.005	0.001	0.001
Nat.Advantages	0.348	0.398	0.328	1.053	0.833	0.584
	(0.017)	(0.022)	(0.030)	(0.153)	(0.120)	(0.082)
Observations	7140	7140	7140	3828	3828	3828
R^2	0.069	0.094	0.061	-0.030	0.006	0.003

Robust standard errors in parentheses.

Table F.3. *OLS and IV multivariate regressions, manufacturing.*

	(1) OLS state	(2) OLS city	(3) OLS county	(4) IV state	(5) IV city	(6) IV county
			EG index - Manufacturing			
Input-Output	0.239	0.142	0.098	0.315	0.171	0.135
	(0.039)	(0.023)	(0.021)	(0.056)	(0.034)	(0.028)
Labor	0.210	0.184	0.141	0.188	0.194	0.103
	(0.011)	(0.009)	(0.008)	(0.032)	(0.020)	(0.015)
Observations	6474	6474	6474	6474	6474	6474
R^2	0.110	0.073	0.053	0.106	0.072	0.050

Robust standard errors in parentheses.

Table F.4. *OLS and IV multivariate regressions, services.*

	(1) OLS state	(2) OLS city	(3) OLS county	(4) IV state	(5) IV city	(6) IV county
			EG index - Services			
Input-Output	0.128	0.157	0.185	-0.050	0.296	0.221
	(0.031)	(0.044)	(0.069)	(0.186)	(0.258)	(0.420)
Labor	0.252	0.355	0.315	0.554	0.323	0.388
	(0.034)	(0.052)	(0.074)	(0.248)	(0.338)	(0.526)
Observations	3312	3312	3312	3312	3312	3312
R^2	0.043	0.042	0.030	0.009	0.036	0.029

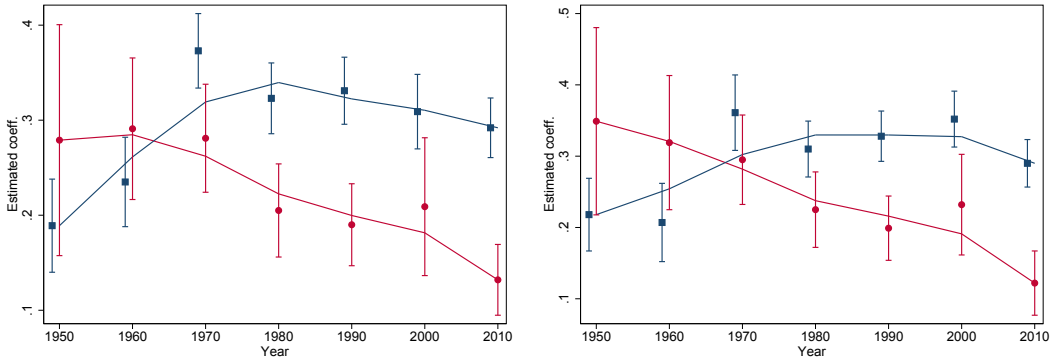
Robust standard errors in parentheses.

Table F.5. *Historical analysis of labor and value-chain linkages in coagglomeration patterns.*

Historical regressions										
	[1] standard 1950-2010		[2] 2-way FE 1950-2010		[3] IO fixed year 1950-2010		[4] standard 1910-2010		[5] 2-way FE 1910-2010	
	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV
Labor	0.188	0.193	0.193	0.225	0.188	0.183	0.115	0.108	0.163	0.171
	(0.025)	(0.026)	(0.032)	(0.034)	(0.031)	(0.046)	(0.016)	(0.025)	(0.021)	(0.049)
Labor × year	0.100	0.110	0.178	0.193	0.101	0.101	-0.001	-0.003	0.005	-0.004
	(0.034)	(0.036)	(0.040)	(0.041)	(0.041)	(0.057)	(0.023)	(0.038)	(0.023)	(0.036)
Input-Output	0.315	0.364	0.362	0.415	0.332	0.424	0.209	0.274	0.189	0.228
	(0.031)	(0.033)	(0.068)	(0.073)	(0.063)	(0.122)	(0.027)	(0.036)	(0.028)	(0.039)
Input-Output × year	-0.194	-0.254	-0.272	-0.331	-0.217	-0.268	-0.097	-0.117	-0.091	-0.121
	(0.041)	(0.044)	(0.095)	(0.101)	(0.084)	(0.156)	(0.047)	(0.065)	(0.050)	(0.071)
R^2	0.067	0.067	0.185	0.185	0.067	0.068	0.029	0.030	0.119	0.122
N	31757	31757	31757	31757	31757	26401	30356	29403	30356	29403

Year is divided by 100. Time dummies for each decade are included, but not reported in the table. Robust standard errors in parentheses. The first block (column 1 and 2) replicates columns 3 and 6 of table 5. The second block controls for (2-way) industry fixed effects. In the third block, the input-output table used to construct value-chain proximity does not change over time but is fixed to the 1950 table (IV: 2010 table). Block 4 and 5 extend the estimates to the period 1910-2010.

Figure F.1. *Trends in Marshallian agglomeration forces for manufacturing - OLS (left) and IV (right)*



Estimated impact of input-output (red circles) and labor (blue squares) linkages on EG-coagglomeration in the period 1950-2010 for manufacturing, with 95% confidence intervals based on robust standard errors. The lines represent LOWESS smooths.

G Correction to CBP censoring

To avoid disclosure of sensitive information, the Census Bureau witholds precise data in the CBP, whenever the information in a cell (by geography and by industry) reveals information about a firm. Instead, in such cases, the CBP report employment in terms of size categories. To translate these categories into single employment values, we follow Holmes and Stevens (2004), who provide the following estimate[37] of the mean employment by CBP class in the year 2000.

[37]Estimates for categories below 1,000 employees are directly computed from national totals, dividing total employment by the number of establishments (within each category). For categories with over 1,000 employees, the average employment in each category is imputed estimating the parameters of a log-normal distribution of employment across establishments. See Appendix A.2 in Holmes and Stevens (2004).

Table G.1. *Mean employment by size category in 2000 CBP, as in Holmes and Stevens (2004).*

Employment range	Average employment
1-4	1.7
5-9	6.6
10-19	13.5
20-49	30.2
50-99	68.8
100-249	150.1
250-499	340.7
500-999	681.3
1,000-1,499	1,208.8
1,500-2,499	1,892.9
2,500-4,999	3,374.7
5,000 or more	9592.0

The CBP categories have changed slightly since Holmes and Stevens (2004). We adapt their estimates to the new categories as follows: categories C, E, F and H directly follow the estimates of Holmes and Stevens. Categories A, B and G use weighted averages from the subcategories in Holmes and Stevens (2004). Categories I, J, K and L use the average ratio between the estimated employment of classes A to G and the mid-point of employment classes A to G (0.7974). This ratio is then multiplied by the midpoint of I, J, K and L to obtain an estimate of the average employment. For category M, we use an average employment of 125,000, which is obtained by multiplying 100,000 by the average of the ratio between the minimum and estimated mean of classes I through L. The resulting scheme is shown below.

Table G.2. *Mean employment by size class in 2003 and 2008 CBP, as used in this paper.*

Class	Employment range	Average employment
A	0-19	5
B	20-99	40
C	100-249	150
E	250-499	341
F	500-999	681
G	1,000-2,499	1,488
H	2,500-4,999	3,375
I	5,000-9,999	5,980
J	10,000-24,999	13,955
K	25,000-49,999	29,904
L	50,000-99,999	59,804
M	100,000 or More	125,000