The Coherence of US cities

Simone Daniotti, Matte Hartog & Frank Neffke

Papers in Evolutionary Economic Geography

25.22

Utrecht University Human Geography and Planning

The Coherence of US cities

Simone Daniotti[a,b,d], **Matté Hartog**[c], **and Frank Neffke**[a]

[a]Complexity Science Hub Vienna, Vienna, 1080, Austria
[b]Vienna University of Technology, Informatics, Vienna, 1040, Austria
[c]Growth Lab, Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA
[d]Utrecht University, Copernicus Institute of Sustainable Development

ABSTRACT

Diversified economies are critical for cities to sustain their growth and development, but they are also costly because diversification often requires expanding a city's capability base. We analyze how cities manage this trade-off by measuring the coherence of the economic activities they support, defined as the technological distance between randomly sampled productive units in a city. We use this framework to study how the US urban system developed over almost two centuries, from 1850 to today. To do so, we rely on historical census data, covering over 600M individual records to describe the economic activities of cities between 1850 and 1940, as well as 8 million patent records and detailed occupational and industrial profiles of cities for more recent decades. Despite massive shifts in the economic geography of the U.S. over this 170-year period, average coherence in its urban system remains unchanged. Moreover, across different time periods, datasets and relatedness measures, coherence falls with city size at the exact same rate, pointing to constraints to diversification that are governed by a city's size in universal ways.

1 Introduction

Diversification is pivotal to economic development and cities' capacity to generate prosperity for their inhabitants[1–3]. Diversified economies are less exposed to idiosyncratic sector-specific shocks[4], have a broader capability base from which to develop new economic activities[5,6] and are better positioned to innovate[7–10]. These and other consequences of diversification have been studied extensively. However, we know much less about how much diversity cities can manage, how this has changed over time, and how cities' capacity to do so is affected by their size. Here, we address these questions by leveraging large-scale micro-datasets that allow us to describe the evolution of the US urban system and the distribution of technological and economic activities across its cities almost from its inception to the present day. In particular, taking inspiration from the literature on economic complexity[5,11,12] we study how *coherent* the activity mix of a city is, in terms of the expected relatedness or technological proximity between two randomly sampled productive units in the city. The less coherent a city is, the broader the—latent—underlying capability base required to support its activity mix will be. To do so, we develop a measure of coherence that is insensitive to economic classification systems and the exact measurement of relatedness. Studying the long-term evolution of the US urban system through this lens reveals that coherence falls with city size at a rate that remains constant across

Contribution Statement
S.D. and F.N.: Conceptualization, Methodology, Writing – Original Draft.
S.D.: Investigation, Formal analysis, Visualization, Software.
F.N.: Supervision.
M.H. and S.D.: Data Curation.
All authors: Writing – Review and Editing.

decades and datasets, suggesting the existence of universal constraints to diversification related to a city's size.

The breadth of a city's capability base is closely related to notions of complexity and diversity. Diversity metrics originated in the field of ecology[13,14]. They provide generalized species counts that account for different aspects of diversity, such as variety (the number of different species), balance (their frequency distributions) and disparity (dissimilarities among species). These approaches inspired similar measures in Evolutionary Economic Geography (EEG[15]). Examples are related and unrelated variety[7], which aim to overcome the dichotomy between specialized and diversified cities by positing that configurations with a wide variety of closely related activities combine elements from both specialized and diversified cities, facilitating knowledge spillovers and fostering innovation. Although generalized species counts and related metrics are useful when comparing across ecosystems or when thinking about innovation in terms of Schumpeterian new combinations[16,17], their exact relation to economies and the capability bases that support their productive structures is less clear. Moreover, measures that take inspiration from the diversity metrics in ecology to describe the breadth of activities in cities tend to mechanically rise with city size[18], such that larger cities will, by construction, be more diversified. In the SI, sec. S4, we describe the mathematical connections between our approach and the broader field of diversity metrics.

In contrast, economic complexity indices[11,12,19] aim to assess how many *capabilities* an economy has. They build on a theory of economic production that postulates the existence of such capabilities and that proposes that economies will produce only those products and services for which they possess all required capabilities. However, there is substantial debate about, if not the usefulness, at least the meaning of complexity measures[12,20–22]. Moreover, the units in which complexity is expressed typically have no straightforward, i.e., economically meaningful, interpretation and neither complexity nor diversity metrics have well defined confidence intervals.

Instead, we build on the branch of the economic complexity literature that focuses on economic transformation[5]. This allows us to start from the notion of *relatedness*. Relatedness is regarded as core to understanding diversification dynamics, with a consensus emerging around the so-called *principle of relatedness*[23]: cities tend to diversify preferentially into activities that are closely related to their current activities. Moreover, unlike economic complexity indices, which have mostly been used to predict a country's productivity, we are mainly interested in what is feasible for a city, balancing costs and benefits. Our coherence metric therefore attempts to assess how *compact* a city's capability base is by analyzing how related different productive units in the city are to one another. Relatedness is here meant to capture the degree to which two activities share the same capability requirements. Therefore, the less related two randomly sampled units on average are, the less coherent the city and the broader its capability base will be.

Methodologically, we take inspiration from information-theory-based metrics of diversity[13,14,24], but instead define a measure of coherence. This approach has several advantages. First, it connects to a well established notion of relatedness as a proxy for overlapping capability requirements that governs cities' diversification patterns. From this perspective, the expected distance between two random productive units conveys a sense of the "volume" of the city's capability base; coherence — the expected relatedness — expresses the opposite: the compactness of this volume. Second, because of its formulation as an expected value, there is a clear way to define not only coherence, but also quantify uncertainty in its measurement, allowing for the construction of confidence intervals. Third, a major challenge to studying the long-run transformation of urban economies is that economic activities and the classification systems to describe them change drastically

over longer periods. Our measure of coherence is insensitive to changes in classification systems and *a priori* unrelated to city size.

We use this coherence to study the long-run evolution of the US urban system between 1850 and today. To do so, we combine several large-scale micro-datasets, from historical census data that cover hundreds of millions of individuals in the 19th and 20th century to millions of patent records that describe the technologies used by US inventors between 1975 and 2020. These datasets cover different periods, different definitions of the US urban system – which grew roughly from 500 cities in 1850 to 900 cities today – and different types of activities (occupations and technologies).

Set against this heterogeneity, our analysis yields two surprising findings. First, the average coherence of cities in the US urban system remained constant in datasets that stretched over 170 years and across activity types. This stability is remarkable, given the drastic economic transformation that took place in this period, including the transition from agriculture to first manufacturing and then services, the rise and fall of the Rustbelt, and the emergence of today's technology hubs. Moreover, individual cities do undergo drastic structural change, as in Detroit's economic ascent and collapse with the fortunes of its automotive industry and Boston's transformation from a port city to a city of higher education[25]. The fact that, in spite of such transformation, coherence remains unchanged therefore suggests that, as cities transform, they, on average, do so in a way that preserves the coherence among their (changing) activities along the way. Second, we uncover a universal relation between coherence and city size: the elasticity of coherence with city size is constant, at about -4%, across time periods and activity types. That is, moving from smaller to larger cities, the mix of activities broadens in a predictable way, such that doubling the size of a city translates roughly into a 4% decrease in coherence. This holds not just true for the urban system as a whole, but also for the urban system on the West Coast of the U.S., which remained relatively disconnected from the eastern U.S. until the early 20th century and whose development we can trace in its entirety, starting from the mid 19th century.

To help understand these patterns, we postulate that larger cities can maintain wider capability bases, which should allow them to develop less coherent activity portfolios. We develop this logic more formally in a model of collective learning in which cities' workers balance imitation and innovation. The model shows how the number of capabilities grows with the size of a city, as well as how this growth is reflected in a decrease in the expected capability overlap between randomly sampled workers, i.e., in the city's coherence.

Conceptually, our work relates to the framework of economic complexity[5,11,26]. The literature on economic complexity assumes that economies mobilize capabilities to produce output. Although capabilities are treated as hidden, unobserved variables, different products and services are generally assumed to require different capabilities. This makes it costly to produce a wide variety of outputs because this will require a broad set of capabilities. Economies can save on the number of capabilities by focusing on sets of closely related activities.

The notion of coherence itself has been widely studied at the firm level in strategic management[27,28] as reviewed in[29], and more recently in research on economic complexity[30,31]. Scholars in EEG have taken the concept of coherence to the regional level. Our work relates most closely to papers in this latter tradition[6,32]. Compared to these prior studies, we offer a principled approach to quantifying coherence as well as to constructing confidence intervals around this estimate. Moreover, most papers in this literature analyze cross-sections or short panels of cities in which activities are recorded in stable classification systems. As a consequence, they study cities and urban systems over years, instead of the decades and centuries over which their transformation typically unfolds.

Finally, our study relates to a variety of academic fields that have documented striking relationships between diversification and the economic dynamism of cities. First, economic geographers who study regional and urban diversity highlight its role in innovation[8], agglomeration externalities[33–35] and the path-dependent nature of regional diversification[7,36–38]. Second, literature on urban scaling[1,39–42] explores how economic activities scale with city size, concentrating in large cities[43–46], and how occupational diversity relates to economic productivity and social network structures[39,47]. In focusing on the expected relatedness between randomly sampled productive units, our approach shares an implicit reliance on the notion of serendipitous encounters—common in both fields—when assessing the relevant radius for individual productive units in a city. However, these bodies of research tend to focus on the consequences of cities' activity mixes, not on what determines the breadth of the activity mix that cities can sustain in the first place.

2 Results

2.1 Data

Our analysis draws from three different data sources. First, we use decennial US censuses for the period 1850-1940. This dataset covers over 600M individual records and allows us to analyze changes in the occupational composition of between 550 and 900 cities in the U.S. at the level of 250 detailed different occupations. Second, for the period between 2002 and 2022, we use data from the US Bureau of Labor Statistics (BLS) on employment by occupation for over 800 occupations in 350 metropolitan areas. For both data sources, we concentrate on occupations that are likely to produce tradable output whose geographic distribution is driven by the availability of relevant capabilities, ignoring occupations that mainly cater to the demand of the local population, such as bakers, teachers, and nurses (see SI, sec. S1). Third, we aggregate data from the US Patent and Trademark Office on over 8M patents between 1980 and 2020 to city-technology cells, distinguishing between 650 technological areas and 900 cities. Together these datasets describe the occupational and/or technological composition of US cities between 1850 and today, except for the decade of 1890 for which a fire destroyed census records and of 1950, 1960 and 1970, for which neither comprehensive employment nor patenting information exists at the city level. Details on cleaning and geocoding are provided in the *Methods* section.

2.2 Defining Coherence

We define our metric of economic coherence in terms of the *relatedness* between the economic activities in which productive entities in a city, such as workers, inventors or firms, are active. Relatedness has been measured in various ways and is often interpreted as a measure of cognitive or technological proximity. To show that our findings are not dependent on the exact definition of relatedness, we explore various relatedness metrics (see SI, sec. S2). First, using matched census records, we construct a measure of skill-relatedness[48] that connects occupations with exceptionally large labor flows between them. Second, in the BLS data, we derive measures of relatedness that express the extent to which two occupations are found in the same cities or industries. Third, in the patent data, relatedness expresses the degree to which two technologies co-occur on the same patents.

Fig. 1 illustrates how we estimate coherence. We first collect relatedness estimates for all pairs of activities in matrices $\mathbf{P}$. Next, we define coherence as the expected relatedness between two randomly sampled units, such as workers or patents, conditional on both units being sampled from the same city $c_1 = c_2 = c$. This involves two steps. The first step assesses how related an activity is to the rest of the urban economy. This measure is known as the activity's *density* in the city[5,49]. The second step averages this density across all activities.

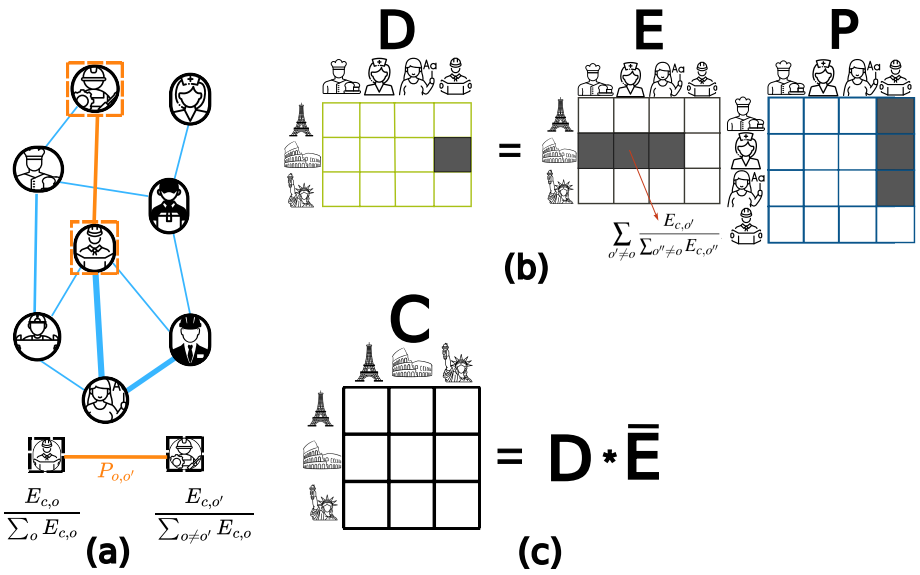


Figure 1. Measuring coherence. **(a)** Proximity (**P**). Stylized depiction of $(n_o \times n_o)$ proximity matrix as a network of n_o activities (here: occupations) connected by edges that reflect their relatedness. **(b)** Density (**D**). The $(n_c \times n_o)$ density matrix reflects occupations' "fit" with each of the n_c cities in our data. It is defined as the expected relatedness to all other occupations in the city, when the other occupations are sampled with probability $p = \frac{E_{c,o}}{\sum_{l \neq o} E_{cl}}$, where $E_{c,o}$ captures the employment of city c in occupation o. Dropping own-occupation contributions, element $D_{c,o}$ is calculated by multiplying a normalized row c of the $(n_c \times n_o)$ employment matrix **E** with column o of matrix **P**, while omitting element o from both vectors. **(c)** Coherence. Estimated as the mean relatedness between workers in different occupations in the same city, coherence is calculated as $D\bar{E}$, where $\bar{\cdot}$ indicates row-normalization by dividing all elements by corresponding row-sums.

To make matters concrete, we focus on the case of workers and their occupations. Elements $P_{o,o'}$ now denote the proximity between occupations o and o'. Furthermore, because the relatedness of an occupation to itself is undefined – and to avoid coherence from picking up individual occupations' own geographical concentration – we condition this expectation on $o \neq o'$, where o and o' denote the first and second sampled worker's occupations. This yields the following expression:

$$C_{c,c'} = \mathbb{E}(P_{o,o'}|c_1 = c, c_2 = c', o_1 = o, o_2 = o' \neq o) \quad (1)$$

Elements of matrix **C** refer to pairs of cities. Coherence estimates are on **C**'s diagonal, where the randomly sampled workers come from the same city. The off-diagonal elements contain the expected proximity between randomly sampled workers from different cities. These elements express how similar these cities are in terms of their activity mix. Expanding matrix **E** with a time dimension, such that it records the employment for a city in a given year in its rows, allows quantifying urban *transformation* as the expected proximity between randomly sampled workers from the same city, but at different points in time. The lower this expected proximity, the more radically a city transforms.

Note that coherence is expressed in the same units as relatedness. Because relatedness definitions may differ across datasets and over time, we normalize coherence estimates by dividing them by a baseline that reflects the coherence of the US as a whole, i.e., the expected relatedness between workers randomly sampled from the entire US economy. The resulting ratio is independent of the unit in which relatedness is expressed and captures how much more or less related workers in the same city are than workers in the US economy as whole. In the SI sec. S4 we explore the relation between coherence and existing diversity indicators.

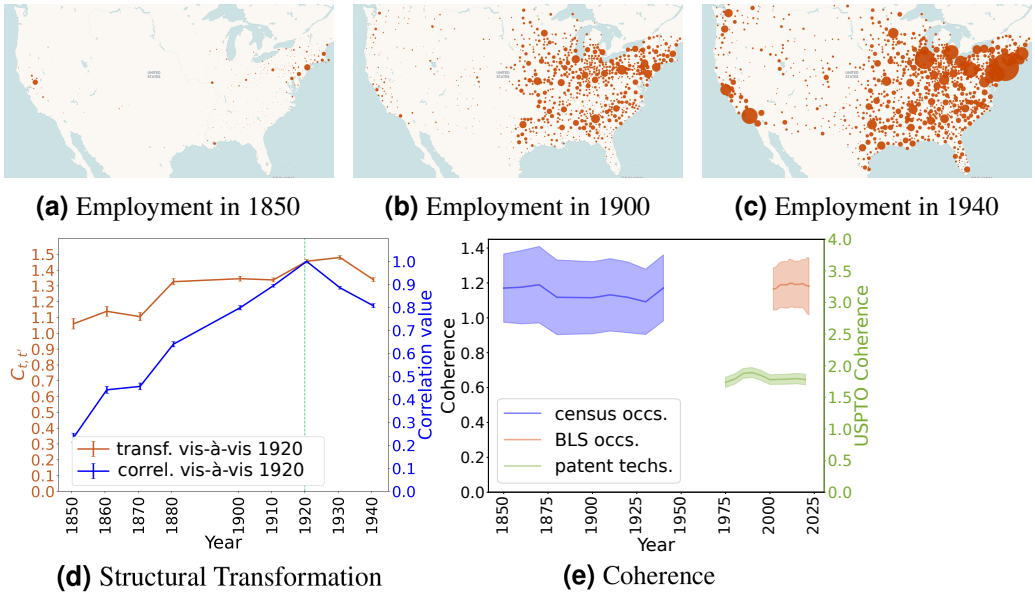
2.3 Structural Transformation of the US urban system

The US population grew from about 23 million inhabitants in 1850 to 132 million inhabitants in 1940 and 332 million inhabitants in 2022. Figure 2a-c show the imprint that this growth left on the human geography of the U.S.. With the growth in population, the economic structure of the U.S. underwent drastic transformation. Whereas in 1850, only about 15% of the population was urban and about 60% of the working population was employed in agriculture, by 1940, 56% of the US population lived in cities and around 25% of workers worked in manufacturing. Nowadays, 80% of Americans live in urban areas and services have become the dominant sector, accounting for 70% of all jobs. These changes are also reflected in the economic structure of US cities, shown in Fig. 2d as shifts in the correlation of their occupational vectors over time and as the average structural transformation as defined in eq. (9) of the Methods section.

Despite such rapid transformation, Fig. 2e shows that cities' average coherence has remained constant over the course of almost two centuries. The graph depicts the average coherence across cities in census data for the period 1850-1940 (blue line), BLS data for 2002-2022 (orange line), and patent data for 1980-2020 (green line). Across all datasets, coherence does not change in a statistically significant way. Moreover, in terms of occupational coherence (blue and orange lines), there is no manifest change in average coherence between the towns and cities of 1850 and the urban areas and metropolises of today. This implies that, although cities substantially transformed their economies, they did so while maintaining a constant level of coherence. That is, as cities moved away from their past activities, they, on average, retained a constant "compactness".

2.4 Coherence and city size

Although the urban system's average coherence remains constant, this does not necessarily hold true for individual cities. In general, diversity rises with city size[50]. Similarly,



(a) Employment in 1850 **(b)** Employment in 1900 **(c)** Employment in 1940

(d) Structural Transformation **(e)** Coherence

Figure 2. Structural Transformation at Constant Coherence.
Urban employment in the U.S. in **(a)** 1850, **(b)** 1900 and **(c)** 1940. Marker area is proportional to the number of employees in each city. The 1850-1940 period covers an important part of the formation and development of the US urban system in which the US went from a mostly rural to an advanced manufacturing and services economy. **(d)** Structural Transformation. Mean correlation between cities' occupational portfolios in different decades with their portfolio in 1920 in blue, and mean structural transformation vis-à vis 1920 in orange. Calculations are limited to cities that exist in both decades and correlations are calculated as $\rho^c_{t,1920} = \text{corr}_{o \in O}\left(E_{c,o,t}, E_{c,o,1920}\right)$, where $E_{c,o,t}$ denotes the employment in occupation o, city c and year t with O the set of occupations. Correlations are subsequently normalized by their country-level counterparts: $\rho_{t,1920} = \text{corr}_{o \in O}\left(E_{o,t}, E_{o,1920}\right)$, where $E_{o,t} = \sum_c E_{c,o,t}$. Ratios $\frac{\rho^c_{t,1920}}{\rho_{t,1920}}$ are depicted along the vertical axis. Structural transformation is calculated as in eq. (9) and normalized by overall US-level transformation between t and 1920. **(e)** Coherence. Mean coherence across US cities. Coherence is calculated as in eq. (1) using occupation data from the census (1850-1940, blue line) and BLS (2002-2022, orange line), and patent data from the USPTO (1980-2020, green line). Values are normalized by dividing by system-level coherence. Shaded areas indicate 95% confidence intervals.

coherence falls with the city size. To study the relationship between coherence and city size, we regress the logarithm of coherence on the logarithm of the total number of workers in the city. Fig. 3a shows that the relation between coherence and city size is downward sloping. More surprisingly, the elasticity of coherence with respect to city size – the slopes $\frac{\partial \log C_{cc}}{\partial \log E_c}$ in Fig. 3a-3c – are statistically indistinguishable across datasets and time periods (a Wald test for equality of slopes yields a p-statistic of 0.8) and close to -4% (95% confidence interval: [-4.4%, -3.4%]). That is, when city size doubles, coherence falls by approximately 4%. In the SI, section S3, we show that these findings also hold with alternative estimators that are better suited to address problems of heteroskedastic errors[51,52]. Moreover, as shown in Fig. S4.1 of the SI, this contrasts with traditional measures of urban diversity, whose elasticities with respect to city size change markedly over the course of a century.

To help understand these findings, we develop an economic complexity inspired

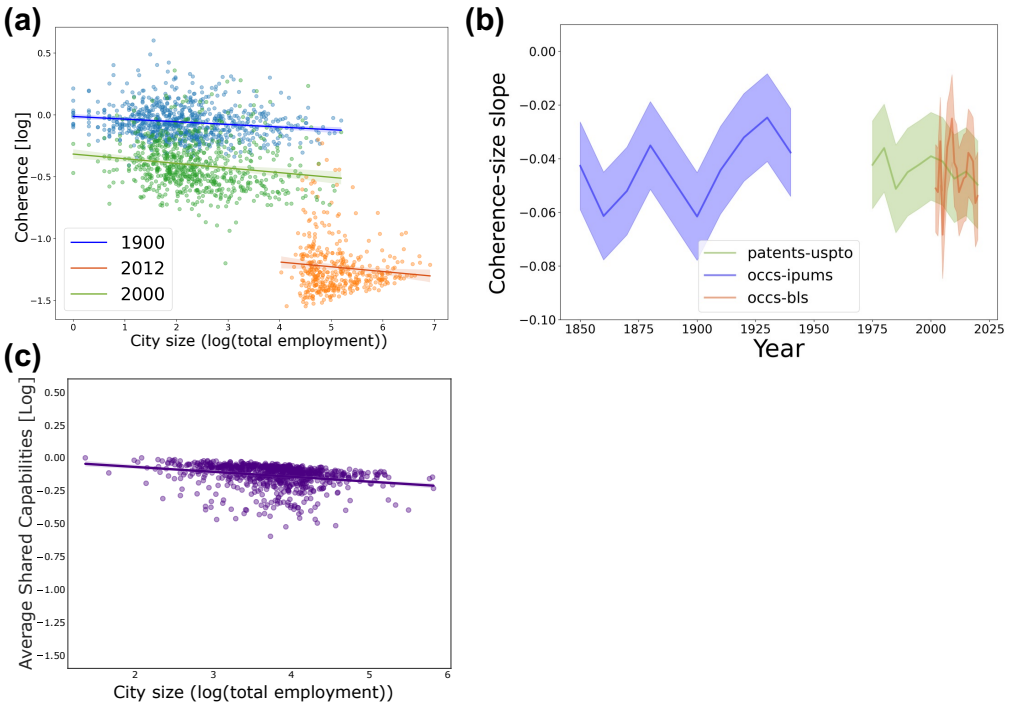


Figure 3. Coherence versus Size. **(a)** Scatter plot of coherence versus city size. Markers refer to cities in a specific dataset, blue: 1900 census, orange: 2012 BLS, green: 2000 patent data. Lines represent best linear fits in each dataset. **(b)** Estimated elasticity of coherence with respect to city size. Colors are as in panel (a). Shaded areas reflect 95% two-sided confidence intervals, based on robust standard errors. The null hypothesis of equal slopes cannot be rejected at any conventional level ($p = 0.8$) in an equality-of-slopes Wald test. **(c)** Simulated coherence in a micro-simulation that balances innovation with imitation (see SI, sec. S5). City sizes are taken from the urban system of 1900 to mimic the blue scatter of panel a. The expected capability overlap, i.e., coherence, drops with an elasticity that depends on the parameter that governs workers' propensity to innovate (see Fig. S5.1), which is calibrated to the observed elasticity of -4%, corresponding to an innovation propensity of 3%.

micro-simulation, where workers either imitate existing workers or innovate and develop new capabilities (see sec. S5 of the SI). As the city grows, so does the number of available capabilities, i.e., the city's capability base. In this context, a natural way to define coherence is the expected frequency with which two randomly drawn workers in a city will have the same capabilities. Fig. 3b shows that by covering just two essential aspects of collective learning – imitation and innovation – this simple model is able to reproduce the functional form of the relation between coherence and city size.

2.5 West-Coast

So far, our analysis has focused on the US urban system as a whole. While still relatively small, by 1850 the eastern part of this urban system had already become somewhat developed. The same is not true for the US West Coast (Fig. 4 (a-c)). The population west of the Rocky Mountains amounted to just 300,000 people in 1850. In 1848, San Francisco had no more than 1,000 inhabitants, and the largest city in the region, Los Angeles, had 1,610 inhabitants, quickly surpassed in the following years by Sacramento. Moreover, in most of the 19th century, the West Coast remained isolated from the rest of the US.

Before the construction of the Panama Canal in 1914, ships had to round Cape Horn to travel between the Atlantic and Pacific Oceans and over land, mountain ranges acted as a significant barrier, until the completion of the Transcontinental Railroad in 1869, which connected the East and West Coast and fueled rapid urban growth and economic development along its route. The US West Coast therefore offers a unique opportunity to study how an urban system develops from scratch and then integrates into a larger, existing urban system.

The structural transformation of the West Coast unfolded very fast. Fig. 4d shows how quickly its cities diversified: whereas, in 1850, they hosted just about 40% of existing occupations related to tradable activities, by 1900, this number had risen to close to 90%. Even in this period of rapid diversification, average coherence of West Coast cities remained remarkably constant and, notably, at levels indistinguishable from those in the eastern US (Fig. 4e). Moreover, although the elasticity of coherence with respect to city size (Fig. 4f) oscillates in the first 30–40 years – possibly due to imprecise measurement in small populations – it thereafter converges rapidly to the same levels as observed in the rest of the U.S.

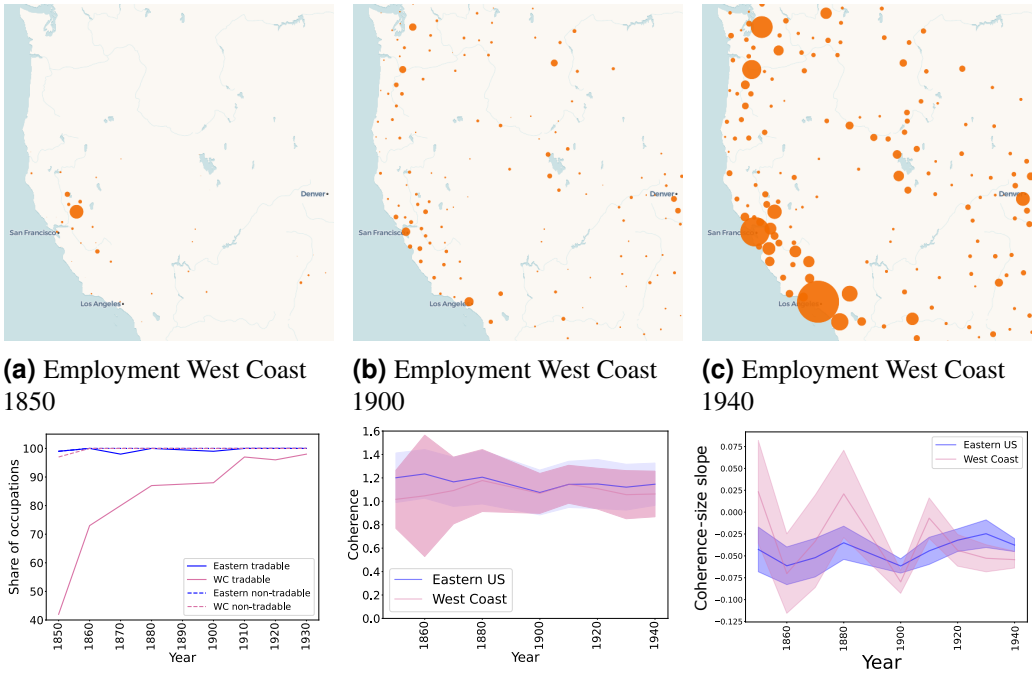
3 Discussion

A diverse economy is of great importance for a city's capacity to innovate, grow and absorb adverse shocks. However, diverse economies require a range of different capabilities[50], which are often expensive to acquire and maintain. This begs the question of how broad a range of activities a city can sustain. To address this question, we have defined a city's coherence as the expected relatedness between randomly sampled productive units, e.g., workers or inventors, from the same city, while controlling for a nation-wide benchmark. Building on the literature on economic complexity, we interpret this coherence as a proxy for the breadth of the city's capability base. This allowed us to address challenges inherent in long-term analyses of economic structures of cities, such as changing classification systems and distinguishing city-level change from broader economy-wide trends, while also focusing on fundamental changes as opposed to superficial shifts between closely related activities in a city's activity mix.

Applying this framework to data sets that describe the mix of economic activities in US cities over a 170-year time period uncovered important regularities. First, although the US urban system has undergone substantial structural change, the coherence of cities within this system has, on average, remained remarkably stable. This suggests that cities' development trajectories are constrained: as cities transition from old activities to new ones, they, on average, maintain a constant level of internal coherence.

Second, coherence decreases with city size at a universal rate. Across different time periods, relatedness measures and activity types the elasticity of coherence with respect to city size is constant at about -4%. That is, coherence decreases by about 4% with each doubling of a city's size, implying that larger cities are able to support a broader set of activities. The constancy of this point estimate across periods and contexts suggests there may exist universal constraints that govern urban diversification. Interestingly, the estimated elasticity closely aligns with leading estimates of the urban wage premium in the U.S., according to which wages rise by around 5% with each doubling of city size[53]. Whether this is a coincidence or due to a connection between coherence and labor productivity is an interesting question for future research.

Third, after a turbulent initial period, cities on the West Coast settle into the same regularities as eastern US cities. The West Coast is an interesting case study, because our data describe its development more or less from the birth of its urban system, when geographical barriers still isolated it from the wider U.S.. In spite of this isolation and



(a) Employment West Coast 1850

(b) Employment West Coast 1900

(c) Employment West Coast 1940

(d) Occupational diversity

(e) Coherence

(f) Elasticity

Figure 4. US West Coast. **(a-c)** Employment in West Coast cities in 1850, 1900 and 1940. In 1850, the largest city on the West Coast is Sacramento. After the gold rush, population growth shifts to other cities, such as Portland, Seattle, San Francisco and Los Angeles. **(d)** Number of existing occupations as a share of all potential occupations in Eastern US (blue) and on the West Coast (pink). Dashed lines refer to occupations in nontradable activities, solid lines in tradable activities. Whereas occupations in nontradables were already abundant on the West Coast in 1850, in tradable activities, less than half of all potential occupations had been developed by then. **(e)** Coherence. Mean coherence across US cities on the West Coast (pink) and in the Eastern US (blue). Coherence is calculated as in eq. (1) and normalized by overall subsystem-level coherence. Shaded areas refer to 95% confidence intervals. **(f)** Estimated elasticity of coherence with respect to city size. Shaded areas reflect 95% two-sided confidence intervals, based on robust standard errors. By 1920, after some initial fluctuations, West Coast cities exhibit the same level of coherence and elasticity of coherence with respect to city size as the remainder of the urban system.

the rapid structural transformation it underwent, cities on the West Coast come to rapidly exhibit the same (constant) coherence and elasticity of coherence with respect to city size as its counterparts east of the Rocky Mountains. This suggests that our findings may generalize to other urban systems.

Our study has several limitations that can be tackled in future research. The first involves theoretical explanations for the functional form and observed elasticity of -4% for the relation between coherence and city size. In the SI, Fig. S5.1, we show that simple probabilistic models of imitation and innovation can reproduce our findings. This points to universality in collective learning as studied in the field of cultural accumulation[54,55]. However, other explanations may exist. One example is regularities in the way division of labor deepens with city size and gives rise to new specializations[39]. Given the invariance across contexts and time of average coherence, as well as of its relation with city size, plausible candidates should be independent of technology and other aspects of societies that change on relatively short time horizons.

A second limitation is the macro-level focus of our work on *average* proximity between a city's workers. This is similar to the notion of serendipitous encounters invoked by the literature on Jacobs externalities and urban scaling. However, if workers direct their search for collaboration partners to connect to, what may matter more is the extent to which they can find a critical mass of closely related workers. Large cities often consist of clusters of highly related activities[56], such that workers may find a sizable number of proximate workers even at low levels of coherence. Our approach can be easily adjusted to this by looking at relatedness quantiles. For instance, one could calculate for each worker the 90th percentile of relatedness to other workers in the city. Cities with low levels of mean coherence but high levels of 90th percentile coherence would consist of disparate clusters of tightly related activities. Such configurations would explain why coherence falls with city size: although workers in large cities often find many workers in related activities, the wide variety of clusters they host lowers the average relatedness captured by our current coherence metric.

A third limitation is that our focus has remained on the urban system as a whole. However, the measures of coherence and structural transformation can also be used to identify outliers: cities that are particularly coherent or incoherent for their size or that have transformed very rapidly. In future research, such analysis may shed light on important aspects of transformational change and urban renewal that help understand why some cities decline and others manage to consistently transform their economy to diversify onto new growth paths.

Finally, our findings have important implications for the broader discourse on regional development and growth policy[57]. Such policy often focuses on fostering diversity in cities[7] and helping a city move into new economic activities to provide opportunities for growth and to avoid lock-in[38]. From the academic literature on related diversification, we know that cities preferentially enter activities related to their existing activity mix and exit unrelated activities. However, our analysis suggests that the breadth of activities a city can sustain is constrained by its size in a precise way. Similarly, the fact that, as cities transform, they maintain a constant level of coherence implies there are structural constraints limiting the speed and trajectory of diversification. This suggests that diversification strategies should benchmark a city's coherence against its size and analyze transformation trajectories that would allow the city to maintain its internal coherence.

4 Methods

4.1 Data

Our analysis is based on three different datasets: US census records between 1850 and 1940, Bureau of Labor Statistics (BLS) data between 2002 and 2022 and United States Patent and Trademark Office (USPTO) data between 1980 and 2020.

US census records. Census data are provided by the Integrated Public Use Microdata Series (IPUMS)[58]. This dataset has approximately 650 million records of responses to Census inquiries for every resident in the United States in the years 1850, 1860, 1870, 1880, 1900, 1910, 1920, 1930, and 1940 (the 1890 records were destroyed in a fire). These records include essential details for our analysis such as each individual's name, occupation, industry, year and state of birth and place of residence. We focus on the working population, which we define as individuals aged 15 to 65 with a known occupation of employment. We exclude individuals in occupational categories "Unknown" or related to agriculture, given that the latter are not part of the urban economy.

Individuals were geocoded and linked across census waves by[59]. We aggregate these data to the level of cities using point-in-polygon merges, where polygons are the metropolitan and micropolitan areas defined by the US Census bureau in its TIGER/Line Shapefiles (https://www.data.gov/). We use the urban shapes for the year 2020, projecting them backward in time to maintain the same spatial definitions of cities. The result is a dataset with employment for approximately 250 harmonized occupations (occ1950) in between 550 US cities in 1850 and 900 cities in 1940. At this level of aggregation, spatial units represent comprehensive economically meaningful functional areas that internalize most local economic interactions.

BLS data. The BLS data are taken from the BLS Occupational Employment Statistics (OES) tables, available at https://www.bls.gov/oes/tables.htm. These tables record the number of employees in approximately 800 (Standard Occupational Classification, 6 digit) occupations across about 350 US metropolitan areas. Furthermore, we use the BLS' industry-occupation matrix, which records the number of employees in occupation-industry cells to calculate relatedness between occupations from their co-occurrence across industries. This data counts around 300 distinct industry codes and 1000 occupations.

Patent data. The USPTO dataset is obtained from PatentsView, https://www.uspto.gov/ip-policy/economic-research/patentsview. We focus on patents granted by the USPTO between 1980 and 2020, geocoding US inventors based on their places of residence through point-in-polygon merges to TIGER/Line Shapefiles. We aggregate patents to the city-technology level, distributing each patent proportionally to the share of its inventors in each cell. This yields a dataset of (fractional) patent counts for approximately 650 technologies in about 900 cities. Although more detailed technology classes are in principle available, this would lead to a very low cell fill in the co-occurrence matrices that are used to estimate our proximity matrices. To strike a balance between detail and cell fill, we opt to use 4-digit technology classes (CPC codes) to define the technological profiles of cities.

Tradable and nontradable occupations. An important distinction exists between economic activities that cater to the need of a city's own population and those whose output is traded with other cities. The former are called "nontradable" and include occupations such as bakers, school teachers, doctors, retail workers, etc.. Demand for nontradable activities is driven mostly by the size of the local population and its purchasing

power, such that employment shares in nontradable occupations are very similar across cities. In contrast, tradable activities depend on the city's productivity in these activities. Examples include manufacturing activities, but also services sold to inhabitants of other cities, such as investment banking, research and development or higher education.

Unlike a city's nontradable activities, the tradable activities a city can develop depend on its capability base. Consequently, the capability base of a city is best reflected in its tradable activities, which is why we drop nontradable activities from our analysis of a city's coherence. In patent data, we consider all activities tradable, given that patents protect inventions on the entire US market. When analyzing occupations, we leverage the fact that nontradable activities essentially follow population and calculate for each occupation how closely its distribution across cities follows the distribution of the US population. That is, we calculate the correlation between two vectors, $\vec{e}_o$, whose elements, $E_{c,o}$, contain the employment of occupation o in city c and $\vec{e}$, whose elements, $E_{c.} = \sum_o E_{c,o}$, describe city c's overall employment as a proxy for its population. This yields the following nontradability score: $NT_o = corr(\vec{e}_o, \vec{e})$. This mimics Krugman's locational Gini index of spatial concentration[60]. In the SI, sec. S1 we show that our tradability ranking is close to one based on the former metric, with a Spearman rank correlation of 0.9 between the two measures.

Fig. S1.1 in the SI shows that coherence estimates rise the more we limit the analysis to tradable occupations in census and BLS data. In the census data, we observe a sharp transition after removing 70% of the least tradable occupations.This shows that coherence is mostly driven by tradable occupations. Although in the BLS data, we do not find a specific transition point, we observe the same strong relation between tradability and coherence. We therefore define tradable occupations in both census and BLS data as occupations with $NT < 0.7$.

4.2 Proximity

The concept of *proximity* between economic activities is central to research on Economic Complexity. In this field, economies are represented as networks of related activities[5,6,23], where relatedness captures the degree to which different activities require similar capabilities. When cities develop related activities, they can support a wide variety of such activities with a limited set of capabilities. In this context, our coherence measure can be viewed as a way to quantify a city's (lack of) diversity, not in terms of its activities, but of its capabilities. This approach resonates with research on diversity in ecosystems, which often distinguishes between variety, balance and disparity[14]. We discuss the relation between coherence and the metrics in this literature in the SI, section S4.

Relatedness can be measured in various ways[49]. For the census data, we base relatedness on labor flows, i.e., on a count of how many individuals move from one occupation to another between two consecutive census waves. To be precise, we assess to what extent the labor flows between occupation o and o' are surprisingly large, using Pointwise Mutual Information as a metric of surprise:

$$\text{PMI}\left(p_{o,o'}\right) = \log\left(\frac{p_{o,o'}}{p_o p_{o'}}\right), \tag{2}$$

where $p_{o,o'}$ is the joint probability that an individual moves from occupation o to occupation o' and p_o and $p_{o'}$ are the marginal probabilities of moving out of occupation o and into occupation o'. We estimate these probabilities by the observed relative frequencies. For instance, we estimate $p_{o,o'}$ as $\hat{p}_{o,o'} = \frac{F_{o,o'}}{\sum_{k,l} F_{k,l}}$, where $F_{o,o'}$ is the observed labor flow from occupation o to o'.

Often, authors draw a sharp distinction between relatedness and unrelatedness[49,61]. Following recommendations in this literature, we define proximity as:

$$P_{oo'} = \begin{cases} \widehat{PMI}(\frac{F_{o,o'}}{\sum_{k,l} F_{k,l}}) & \text{if } \widehat{PMI}(\frac{F_{o,o'}}{\sum_{k,l} F_{k,l}}) > 0, \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $F_{o,o'}$ is the mean of the labor flow between two occupations o and o' across all pairs of consecutive census waves and $\widehat{PMI}$ is estimated using the Bayesian approach in[62]. This sets all negative elements of matrix $\mathbf{P}$ to zero. Furthermore, because the relatedness of an activity to itself is ill-defined, diagonal elements of proximity matrices are ignored in the definition of coherence (see below).

Replacing flows by co-occurrences, we can also derive estimates of proximity from the frequency with which two occupations co-occur in the same industry or city. We use city-level co-occurrences as an alternative proximity metric in our census data and city and industry-level co-occurrences to produce two different proximity metrics in the BLS data. In patent data, we calculate proximity from the frequency with which technology codes co-occur on the same patent. Our results prove remarkably robust to changes in the definition of proximity (sec. S2 of the SI).

4.3 Defining Coherence

We define coherence as the expected proximity between two randomly sampled workers, conditional on the workers being from the same city and employed in different occupations. Breaking down the calculation of coherence into two steps helps connect the coherence metric to the literature on economic complexity[5,26,63]. This is illustrated in Fig. 2. First, we calculate the weighted average proximity of a given occupation, o to all other occupations in the city c. This quantity is closely related to what the economic complexity literature[49] refers to as o's *density*, $D_{c,o}$, in city c:

$$D_{c,o} = \sum_{o' \neq o} Pr(o_2 = o' | c_1 = c, c_2 = c, o' \neq o) P_{o,o'}, \tag{4}$$

which can be estimated as:

$$\hat{D}_{c,o} = \sum_{o' \neq o} \frac{E_{c,o'}}{\sum_{o'' \neq o} E_{c,o''}} P_{o,o'} \tag{5}$$

Coherence, a city-level variable, is now simply the expected density across all occupations. To calculate this, we construct the following matrix:

$$\begin{aligned} C_{c,c'} &= \mathbb{E}(P_{o,o'} | c_1 = c, c_2 = c', o_1 = o, o_2 = o' \neq o) \\ &= \sum_o Pr(o_1 = o | c_1 = c, c_2 = c') \sum_{o'} Pr(o_2 = o' | c_1 = c, c_2 = c', o_1 = o, o_2 = o' \neq o) P_{o,o'}, \end{aligned} \tag{6}$$

which can be estimated as the employment-weighted average density:

$$\begin{aligned} \hat{C}_{c,c'} &= \sum_o \frac{E_{c,o}}{\sum_{o''} E_{c,o''}} \sum_{o' \neq o} \frac{E_{c',o'}}{\sum_{o'' \neq o} E_{c',o''}} P_{o,o'} \\ &= \sum_o \frac{E_{co}}{\sum_{o''} E_{c,o''}} \hat{D}_{c',o}, \end{aligned} \tag{7}$$

where E_{co} denotes the number of workers employed in occupation o and city c. A city's coherence is found on the diagonal of matrix $\hat{\mathbf{C}}$, which contains estimates of the expected proximity between workers that were sampled from the same city. The average coherence across the urban system, plotted in Fig. 2e, is calculated as the weighted average of the diagonal elements of $\hat{\mathbf{C}}$, using cities' overall employment as weights. Next, we rescale this estimate by dividing by the analogous quantity for the US economy as a whole. That is we combine all cities other than c into one unit such that the calculations in eq. (6) yield a scalar. Confidence intervals are based on estimates of the standard errors of eq. (7), which are calculated as follows:

$$\sigma(\hat{C}_{c,c'}) = \sqrt{\sum_{o}\sum_{o'\neq o}\left[\frac{E_{c,o}}{\sum_{o''}E_{c,o''}}\frac{E_{c',o'}}{\sum_{o''\neq o}E_{c',o''}}\right]^2\sigma(P_{o,o'})^2}, \tag{8}$$

where $\sigma(P_{o,o'})$ is the Bayesian estimate of the standard deviation of the proximity between occupations o and o'[62].

The off-diagonal elements of $\hat{\mathbf{C}}$ also have a useful interpretation: they can be regarded as estimates of the proximity between two cities. We use this to quantify the amount of *structural transformation* a city undergoes. To do so, we change matrix $\hat{\mathbf{C}}$ such that its elements refer to observations for the same city at different points in time instead of to different cities in the same period. This yields elements that estimate the expected proximity between two workers that were sampled from the same city, but in different years:

$$\hat{C}_{c_t,c_{t+\tau}} = \sum_{o}\bar{E}_{c_t,o}\sum_{o'\neq o}\frac{E_{c_{t+\tau},o'}}{\sum_{o''\neq o}E_{c_{t+\tau},o''}}P_{oo'}, \tag{9}$$

where $\bar{\cdot}$ indicates row-normalization by dividing by row-sums. Values $\hat{C}_{c_t,c_{t,t+\tau}}$, normalized by the estimated average coherence in 1920, are shown in orange in Fig. 2d.

References

1. Youn, H. *et al.* Scaling and universality in urban economic diversification. *J. The Royal Soc. Interface* **13**, 20150937, DOI: 10.1098/rsif.2015.0937 (2016).
2. Chong, S. K. *et al.* Economic outcomes predicted by diversity in cities. *EPJ Data Sci.* **9**, 17, DOI: 10.1140/epjds/s13688-020-00234-x (2020).
3. Mazzarisi, O., de Azevedo-Lopes, A., Arenzon, J. J. & Corberi, F. Maximal Diversity and Zipf's Law. *Phys. Rev. Lett.* **127**, 128301, DOI: 10.1103/PhysRevLett.127.128301 (2021).
4. Rosenthal, S. S. & Strange, W. C. Chapter 49 - Evidence on the Nature and Sources of Agglomeration Economies. In Henderson, J. V. & Thisse, J.-F. (eds.) *Handbook of Regional and Urban Economics*, vol. 4 of *Cities and Geography*, 2119–2171, DOI: 10.1016/S1574-0080(04)80006-3 (Elsevier, 2004).
5. Hidalgo, C. A., Klinger, B., Barabási, A.-L. & Hausmann, R. The Product Space Conditions the Development of Nations. *Science* **317**, 482–487, DOI: 10.1126/science.1144581 (2007).
6. Neffke, F., Henning, M. & Boschma, R. How Do Regions Diversify over Time? Industry Relatedness and the Development of New Growth Paths in Regions: ECONOMIC GEOGRAPHY. *Econ. Geogr.* **87**, 237–265, DOI: 10.1111/j.1944-8287.2011.01121.x (2011).

7. Frenken, K., Van Oort, F. & Verburg, T. Related Variety, Unrelated Variety and Regional Economic Growth. *Reg. Stud.* **41**, 685–697, DOI: 10.1080/00343400601120296 (2007).

8. Jacobs, J. *The Economy of Cities* (Random House, 1969).

9. Feldman, M. P. & Audretsch, D. B. Innovation in cities: Science-based diversity, specialization and localized competition. *Eur. Econ. Rev.* (1999).

10. Moore, M. F., Ryan. *An Analysis of Jane Jacobs's The Death and Life of Great American Cities* (Macat Library, London, 2017).

11. Hidalgo, C. A. & Hausmann, R. The building blocks of economic complexity. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 10570–10575, DOI: 10.1073/pnas.0900943106 (2009).

12. Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A. & Pietronero, L. A New Metrics for Countries' Fitness and Products' Complexity. *Sci Rep* **2**, 723, DOI: 10.1038/srep00723 (2012).

13. Stirling, A. A general framework for analysing diversity in science, technology and society. *J. R. Soc. Interface.* **4**, 707–719, DOI: 10.1098/rsif.2007.0213 (2007).

14. van Dam, A. Diversity and its decomposition into variety, balance and disparity (2019). 1902.09167.

15. Boschma, R. A. & Frenken, K. Why is economic geography not an evolutionary science? Towards an evolutionary economic geography. *J. Econ. Geogr.* **6**, 273–302, DOI: 10.1093/jeg/lbi022 (2006).

16. Schumpeter, J. A. *The Theory of Economic Development: An Inquiry Into Profits, Capital, Credit, Interest, and the Business Cycle* (Transaction Publishers, 1911).

17. Weitzman, M. L. Recombinant growth. *The quarterly journal economics* **113**, 331–360 (1998).

18. Bathelt, H. Related Variety and Regional Development. *Hum. Geogr. Plan.* (2022).

19. Sciarra, C., Chiarotti, G., Ridolfi, L. & Laio, F. Reconciling contrasting views on economic complexity. *Nat. communications* **11**, 3352 (2020).

20. McNerney, J., Li, Y., Gomez-Lievano, A. & Neffke, F. Bridging the short-term and long-term dynamics of economic structural change (2021). 2110.09673.

21. van Dam, A. *et al.* Correspondence analysis, spectral clustering and graph embedding: applications to ecology and economic complexity. *Sci. reports* **11**, 8926 (2021).

22. Nomaler, Ö. & Verspagen, B. Reinterpreting economic complexity in multiple dimensions. *arXiv preprint arXiv:2409.01830* (2024).

23. Hidalgo, C. A. *et al.* The Principle of Relatedness. In Morales, A. J., Gershenson, C., Braha, D., Minai, A. A. & Bar-Yam, Y. (eds.) *Unifying Themes in Complex Systems IX*, 451–457, DOI: 10.1007/978-3-319-96661-8_46 (Springer International Publishing, Cham, 2018).

24. Rao, C. R. Diversity and dissimilarity coefficients: A unified approach. *Theor. Popul. Biol.* **21**, 24–43, DOI: 10.1016/0040-5809(82)90004-1 (1982).

25. Glaeser, E. L. Reinventing boston: 1630–2003. *J. Econ. Geogr.* **5**, 119–153 (2005).

26. Balland, P.-A. *et al.* The new paradigm of economic complexity. *Res. Policy* **51**, 104450 (2022).

27. Teece, D. J., Rumelt, R., Dosi, G. & Winter, S. Understanding corporate coherence: Theory and evidence. *J. Econ. Behav. & Organ.* **23**, 1–30, DOI: 10.1016/0167-2681(94)90094-9 (1994).

28. Palich, L. E., Cardinal, L. B. & Miller, C. C. Curvilinearity in the diversification–performance linkage: an examination of over three decades of research. *Strateg. management journal* **21**, 155–174 (2000).

29. Robins, J. A. & Wiersema, M. F. The measurement of corporate portfolio strategy: Analysis of the content validity of related diversification indexes. *Strateg. Manag. J.* **24**, 39–59 (2003).

30. Pugliese, E., Napolitano, L., Zaccaria, A. & Pietronero, L. Coherent diversification in corporate technological portfolios. *PloS one* **14**, e0223403 (2019).

31. Aufiero, S., De Marzo, G., Sbardella, A. & Zaccaria, A. Mapping job fitness and skill coherence into wages: an economic complexity analysis. *Sci. Reports* **14**, 11752 (2024).

32. Essletzbichler, J. Relatedness, industrial branching and technological cohesion in us metropolitan areas. In *Evolutionary Economic Geography*, 48–62 (Routledge, 2017).

33. Glaeser, E. L., Kallal, H. D., Scheinkman, J. A. & Shleifer, A. Growth in Cities. *J. Polit. Econ.* **100**, 1126–1152, DOI: 10.1086/261856 (1992).

34. Marshall, A. *Principles of Economics* (Palgrave Macmillan UK, London, 1890).

35. Porter, M. E. *The Competitive Advantage of Nations* (simon and schuster, 1990), simon and schuster edn.

36. Boschma, R. A. & Lambooy, J. G. Evolutionary economics and economic geography. *J. Evol. Econ.* **9**, 411–429, DOI: 10.1007/s001910050089 (1999).

37. Martin, R. & Sunley, P. Path dependence and regional economic evolution. *J. Econ. Geogr.* **6**, 395–437, DOI: 10.1093/jeg/lbl012 (2006).

38. G, G. The weakness of strong ties ; The lock-in of regional development in Ruhr area. *The embedded firm ; On socioeconomics industrial networks* 255–277 (1993).

39. Bettencourt, L. M. A., Samaniego, H. & Youn, H. Professional diversity and the productivity of cities. *Sci Rep* **4**, 5393, DOI: 10.1038/srep05393 (2014).

40. Bettencourt, L. M. A. The Origins of Scaling in Cities. *Science* **340**, 1438–1441, DOI: 10.1126/science.1235823 (2013).

41. Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C. & West, G. B. Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl. Acad. Sci.* **104**, 7301–7306, DOI: 10.1073/pnas.0610172104 (2007).

42. Bettencourt, L. M. A. & Zünd, D. Demography and the emergence of universal patterns in urban systems. *Nat Commun* **11**, 4584, DOI: 10.1038/s41467-020-18205-1 (2020).

43. Bettencourt, L. M. A., Lobo, J., Strumsky, D. & West, G. B. Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities. *PLOS ONE* **5**, e13541, DOI: 10.1371/journal.pone.0013541 (2010).

44. Schläpfer, M. *et al.* The scaling of human interactions with city size. *J. R. Soc. Interface.* **11**, 20130789, DOI: 10.1098/rsif.2013.0789 (2014).

45. Balland, P.-A. *et al.* Complex economic activities concentrate in large cities. *Nat Hum Behav* **4**, 248–254, DOI: 10.1038/s41562-019-0803-3 (2020).

46. Duranton, G. & Puga, D. Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products. *Am. Econ. Rev.* **91**, 1454–1477, DOI: 10.1257/aer.91.5.1454 (2001).

47. Youn, H. *et al.* The systematic structure and predictability of urban business diversity. *J. R. Soc. Interface.* **13**, 20150937, DOI: 10.1098/rsif.2015.0937 (2016). 1405.3202.

48. Neffke, F. & Henning, M. Skill relatedness and firm diversification: Skill Relatedness and Firm Diversification. *Strat. Mgmt. J.* **34**, 297–316, DOI: 10.1002/smj.2014 (2013).

49. Li, Y. & Neffke, F. Evaluating the principle of relatedness: Estimation, drivers and implications for policy, DOI: 10.48550/arXiv.2205.02942 (2023). 2205.02942.

50. Gomez-Lievano, A., Patterson-Lomba, O. & Hausmann, R. Explaining the Prevalence, Scaling and Variance of Urban Phenomena. *Nat Hum Behav* **1**, 0012, DOI: 10.1038/s41562-016-0012 (2016). 1604.07876.

51. Silva, J. S. & Tenreyro, S. The log of gravity. *The Rev. Econ. statistics* 641–658 (2006).

52. Leitao, J. C., Miotto, J. M., Gerlach, M. & Altmann, E. G. Is this scaling nonlinear? *Royal Soc. open science* **3**, 150649 (2016).

53. Chauvin, J. P., Glaeser, E., Ma, Y. & Tobio, K. What is different about urbanization in rich and poor countries? cities in brazil, china, india and the united states. *J. Urban Econ.* **98**, 17–49 (2017).

54. Kempe, M. & Mesoudi, A. An experimental demonstration of the effect of group size on cultural accumulation. *Evol. Hum. Behav.* **35**, 285–290, DOI: 10.1016/j.evolhumbehav.2014.02.009 (2014).

55. Kline, M. A. & Boyd, R. Population size predicts technological complexity in Oceania. *Proc. Royal Soc. B: Biol. Sci.* **277**, 2559–2564, DOI: 10.1098/rspb.2010.0452 (2010).

56. Porter, M. The economic performance of regions. *Reg. Stud.* **37**, 549–578 (2003).

57. Boschma, R. Evolutionary economic geography and its implications for regional innovation policy. *Pap. Evol. Econ. Geogr.* **9**, 1–33 (2009).

58. Ruggles, S. *et al.* *IPUMS USA: Version 11.0 [Dataset]. Minneapolis* (IPUMS Minnsesota: MN, 2021).

59. Protzer, E. S., Orazbayev, S., Gomez-Lievano, A., Hartog, M. & Neffke, F. A new algorithm to efficiently match us census records and balance representativity with match quality. Tech. Rep., Harvard's Growth Lab (2024).

60. Krugman, P. *Geography and trade* (MIT press, 1992).

61. Muneepeerakul, R., Lobo, J., Shutters, S. T., Goméz-Liévano, A. & Qubbaj, M. R. Urban economies and occupation space: Can they get "there" from "here"? *PloS one* **8**, e73676 (2013).

62. van Dam, A., Gomez-Lievano, A., Neffke, F. & Frenken, K. An information-theoretic approach to the analysis of location and colocation patterns. *J. Reg. Sci.* **63**, 173–213, DOI: 10.1111/jors.12621 (2023).

63. Hidalgo, C. A. Economic complexity theory and applications. *Nat Rev Phys* **3**, 92–113, DOI: 10.1038/s42254-020-00275-1 (2021).

64. Macarthur, R. H. Patterns of Species Diversity. *Biol. Rev.* **40**, 510–533, DOI: 10.1111/j.1469-185X.1965.tb00815.x (1965).

65. Balassa, B. Trade liberalisation and "revealed" comparative advantage 1. *The manchester school* **33**, 99–123 (1965).

Acknowledgements

The authors thank the following persons for their helpful comments on earlier drafts of this paper: Andrea Musso, Andrés Gómez Liévano, Johannes Wachs, Sándor Juhász, Xiangnan Feng, James McNerney, Hillary Vipond.

S.D. and F.N. receive financial support from the Austrian Research Promotion Agency (FFG), project #873927 (ESSENCSE).

Author contributions statement

All authors reviewed the manuscript.

Supplementary information

This document provides supplementary information to the paper "The Coherence of US Cities." Sec. S1 classifies jobs into tradable and nontradable occupations using a nontradability index. Sec. S2 explores different methods to calculate occupational relatedness and replicates the paper's main results using the resulting alternative measures. Sec. S3 reports regression results using alternative, likelihood-based estimators when estimating elasticities of coherence with respect to city size. Sec. S4 discusses the relation between our coherence measure and the notions of variety, balance and disparity as developed in research on diversity in ecosystems. Sec. S5 explores the relation between city size and coherence in a micro-simulation where productive units choose between innovating new activities or imitating existing activities.

S1 Tradable versus nontradable occupations

We have focused our analysis on occupations in tradable activities ("tradable occupations"). The rationale is that, if coherence is to assess the breadth of a city's capability base, demand for the occupations we use to measure coherence should be driven by a city's capabilities. This is not necessarily the case for occupations that cater to local demand, whose size mainly reflects the size, income and preferences of a city's population, not its capabilities. The dependence on local customers also constrains the growth of these occupations to the local demand. In contrast, occupations that are used in activities that can be exported to other cities ("tradable occupations") can grow independently of the size of the local market. Their growth will instead be determined by the city's productivity in the activity, in other words, by the nature of the city's capability base. Tradable occupations therefore capture what a city is good at, making them more informative of a city's coherence than nontradable occupations.

To distinguish nontradable from tradable occupations we leverage the fact that nontradable occupations follow local demand. Consequently, their spatial distribution across cities will closely mimic the spatial distribution of population. Following this logic, in the main text, we defined a nontradability index as $NT_o = corr(\vec{e}_o, \vec{e})$, where $\vec{e}_o$ is a vector that describes the distribution of workers in occupation o across cities and $\vec{e}$ a vector describing the distribution of all workers across cities. The higher this correlation, the more the occupation's spatial footprint mimics the spatial footprint of the US population.

Fig. S1.1 shows how average coherence in the US urban system changes as we restrict the sample of occupations used in the calculations to increasingly less nontrable (i.e., more tradable) occupations. Moving from left to right on the horizontal axis, we progressively drop more occupations, by decreasing the maximum acceptable NT score. The figures plot coherence averaged across the time windows considered in the main text against the share of dropped workers for occupations in census data (left panel) and in BLS data (right panel).

The census data manifest a sharp transition in coherence after removing the 70% workers in the most nontradable occupations. We therefore use this threshold to distinguish between tradable and nontradable occupations. In the BLS data, no such sharp transition point exists, but the inflection stretches over an interval around the same 70% value. Therefore, we use the same cut-off in the BLS data.

An alternative way of ranking occupations by their tradability refines Krugman's locational Gini index of geographic concentration[60], G_o. The Spearman rank correlation between $NT_o|$ to G_o ranges from 0.9 to 0.94 (Kendall's tau between 0.73 and 0.84) across our datasets, showing that both measures lead to very similar tradability scores. We include the tradability rankings of occupations in the replication data.

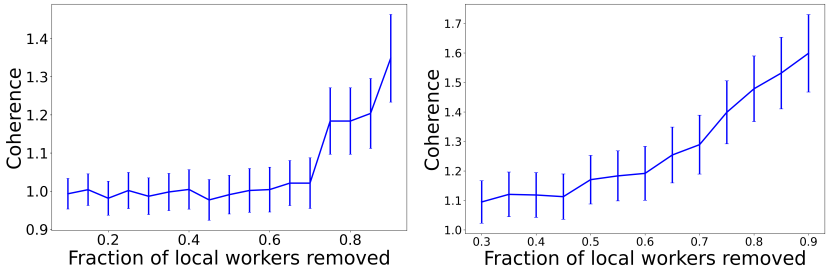


Figure S1.1. Tradable versus nontradable occupations. Estimated average coherence across cities in the US for progressively more tradable occupations. Left panel: census data. Right panel: BLS data. The data has been aggregated taking the time average over all available years. In all panels coherence rises as occupations become more tradable.

S2 Alternative relatedness metrics

In this section we test the robustness of the paper's main results to changes in the definition of occupational relatedness. We explore three different types of relatedness metrics that use different types of data as an input.

Labor flows. We create labor flows between occupations using linkages between records of individuals across different census waves developed by Protzer and colleagues[59]. This allows us to follow a large number of individuals across census waves and to observe how many individuals change occupations between two waves. To do so, for each pair of sequential decades[1], we count the number of individuals who were employed in occupation o in one decade and subsequently in occupation o' in the next. We collect these labor flows in matrix F, with elements $F_{oo'}$. The resulting metric is used to derive the census results in the main text.

Industrial co-occurrences. Industrial co-occurrences measure the extent to which different occupations appear in the same industries. To be precise, we count the number of industries in which both occupations of a pair of occupations (o, o') are present to obtain a co-occurrence matrix, $K^{ind}_{o,o'}$. We use this matrix to calculate a relatedness matrix, by replacing $F_{o,o'}$ in eq. (3) by $K^{ind}_{oo'}$. The resulting occupational relatedness metric is used in the analysis of the BLS data in the main text. The level of definition of the industry codes is 4-digit NAICS industry.

Geographical co-occurrences (co-agglomeration). Following the original approach introduced by Hidalgo and colleagues[5], we also explore geographical co-agglomeration patterns to measure relatedness. That is, we count how often two occupations are present in the same cities and collect such co-occurrences in matrix $K^{geo}_{o,o'}$. Co-occurrences are then converted into a relatedness matrix by replacing $F_{o,o'}$ with $K^{geo}_{o,o'}$ in eq. (3).

These different measures are likely to emphasize different types of relatedness. Labor flows most likely reveal similarities in skill requirements[48], whereas co-occurrence measures will capture different types of economies of scope. For instance, co-occurrences in industries reflect benefits of combining different types of skills in production processes, whereas co-occurrences at the city level will in addition shed light on agglomeration benefits, such as those derived from labor market pooling effects.

[1] Note that due to the missing 1890 census, 1880 labor flows are taken between 1880 and 1900, crossing two full decades instead of one.

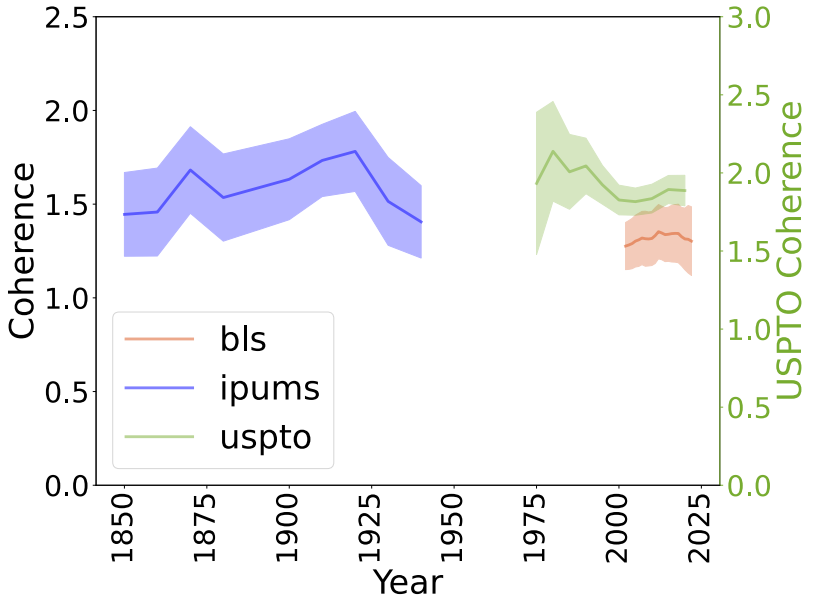


Figure S2.1. Coherence. Mean coherence across US cities with 95% confidence intervals. Coherence is calculated as in eq. (1) using employment data from the US census (1850-1940, blue line) and BLS (2002-2022, orange line), and patent data from the USPTO (1980-2020, green line). Values are normalized by dividing by system-level coherence. The proximity matrices are based on city-level co-occurrences.

Fig. S2.1 replicates the results in the main text with the alternative relatedness metrics based on co-agglomeration patterns. The upper panel shows that coherence remains more or less unchanged across time periods. The lower panel shows that the elasticity of coherence with respect to city size is constant across time periods and datasets. Moreover, a Wald test that tests whether the estimated elasticities are the same in each year and data set and equal to the average elasticity observed in Fig. 3 of the main text cannot be rejected at any conventional level (p-value: 0.6).

S3 Alternative estimators of elasticities

In the main text, we log transform coherence and city size (measured in terms of the total number of employees in the city). Log-transforming dependent variables has been criticized by[51] and[52] for biases that arise when errors are heteroskedastic. In Fig. S3.1, we therefore repeat the estimation of the elasticity of coherence with respect to city size using the PPML estimator proposed by[51] and the likelihood based approach in[52]. Both estimators corroborate the finding in the main text that the estimated elasticity is constant across time and datasets, with a point estimate of around -4%.

S4 Variety, Coherence and Capabilities

Diversity is a key concept in research in ecology that describes ecosystems in terms of their composition of species[64]. Methodological work in this area[13,14,24] distinguishes among three fundamental aspects of diversity:

Variety: This describes the number of distinct species in the ecosystem. A typical metric would be a simple count of the different species in the ecosystem.

Balance: This refers to the distribution or allocation of specimens across the ecosystem's species. Metrics include the Herfindahl-Hirschman Index and entropy.

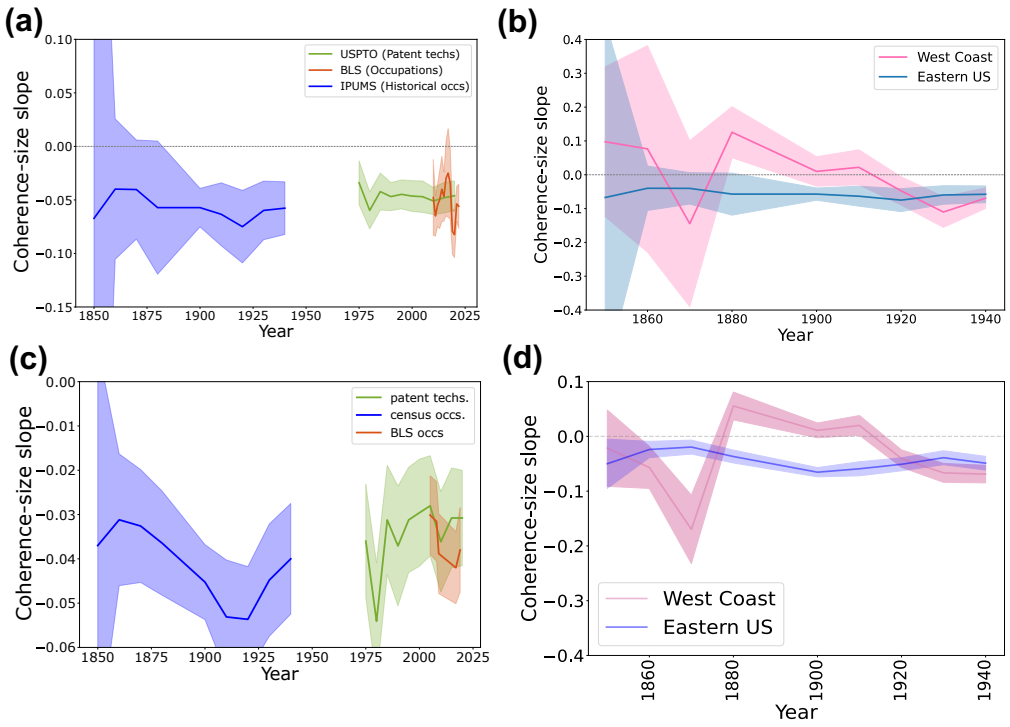


Figure S3.1. Coherence versus Size. Estimated elasticity of coherence with respect to city size using the method by Leitao and colleagues[52](Panel **a** for the main datasets, Panel **b** for the West Coast) and using PPML estimation (Panel **c** for the main datasets, Panel **d** for the West Coast). Shaded areas reflect 95% two-sided confidence intervals, based on heteroskedasticity robust standard errors.

Metric	Variety	Balance	Disparity	Original Equation	Imposed Conditions on Coherence
Species Count / Richness	x			Richness = Number of nonzero E_{oc}	Set all weights $E_{oc} = 1$ if present; ignore proximity matrix $P_{oo'}$
Shannon Entropy	x	x		$H = -\sum_i p_i \log p_i$	$P_{oo'} = \mathbb{I}$; $E_{co} = \sqrt{E_{co} log(E_{co})}$; assume $c = c'$ (same group)
Simpson's Index	x	x		$D = \sum_i p_i^2$	Set $P_{oo'} = \mathbb{I}$; assume $c = c'$
Kullback-Leibler Divergence		x		$D_{KL}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$	Set $P_{oo'} = \log \frac{P(o')}{Q(o')}$; compare distributions across classes
Density Metric	x		x	Density $= \frac{\sum_j \phi_{ij} M_j}{\sum_j \phi_{ij}}$	Set $E_{oc} = 1$ if active; use $P_{oo'}$ = relatedness; ignore frequency

Table S2.1. Comparison of Diversity Metrics: Variety, Balance, Disparity, Original Equations, and Conditions to Recover from Coherence

Disparity: This involves assessing the dissimilarity among species. It examines the question of "how distinct or dissimilar are different species from each other?".

Each of these properties – variety, balance, and disparity – addresses a specific facet of diversity and different metrics typically focus on one or a combination of these facets. For instance, entropy-based related variety measures[7] incorporate variety and balance, whereas the density metrics that are often used in economic complexity research[5,49] incorporate variety and disparity.

In comparison, coherence as defined in eq. (6) is similar in spirit to the Rao-Stirling Entropy[24], which incorporates all three aspects of diversity, but requires symmetric relations and global structure, which may not fit the directional and local nature of occupational flows or relatedness. In fact, slight modifications allow us to turn on or off any of these aspects in the coherence metric to let it mimic specific existing measures. For instance, imposing $P_{o,o'} = \mathbb{I}$ such that the proximity is the identity matrix eliminates disparity. If we furthermore redefine the employment weights in eq. (7) as $E_{c,o} = \sqrt{E_{c,o} log(E_{c,o})}$ we obtain an entropy-based related variety measure[7]. If we instead let $E_{c,o} = \frac{M_{c,o}}{\sum_o M_{c,o}}$, where $M_{c,o} = 1\,(PMI(E_{c,o}) > 0)$, we obtain a density-based variety metric[5].

Our coherence metric definition generalizes these approaches by incorporating a flexible proximity kernel between occupations with weighted, potentially asymmetric relations. It simplifies to classical measures under some assumptions, but provides a richer structure capable of capturing the complexity of economic structures. Entropic measures, as Shannon Entropy, and Simpson's index do not take into account any measure of proximity between the activity units. These measures can be obtained from our coherence definition by imposing the proximity to be the identity matrix. Density-derived metrics, instead, do not take into account the balance characteristics. In table S2.1 we describe characteristics of various well-known metrics used in the literature, with the associated characteristics and how to obtain them restricting our coherence definition.

S4.1 Elasticity of diversity measures with respect to population

In the main text, we show that the relation between coherence and city size is unchanging across time periods and data sets. Here, we show that this is not the case for existing measures of diversity. In particular, Fig. S4.1 plots the logarithm of diversity against the logarithm of total employment in a city, using as measures of diversity Frenken et al.'s[7] related variety, the entropy of the activity distribution across classes and a simple count of activities ("RCA-variety") with $PMI > 0$.[2] The plots show that the relation between different diversity measures and population size does not stay constant, but becomes flatter over time.

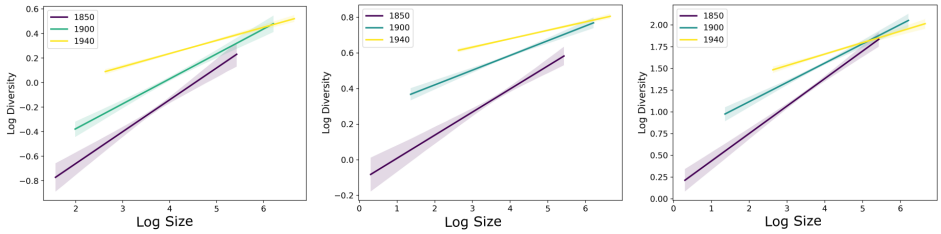


Figure S4.1. Diversity versus Size. Left panel: Related variety. Middle panel: entropy, Right panel: RCA-variety, i.e, number of activities with $RCA > 1$, where $RCA = \frac{E_{c,o}/\sum_l E_{c,l}}{\sum_k E_{k,o}/\sum_{k',l'} E_{k',l'}}$. Colors refer to different years selected.

[2]Note that this condition is equivalent to requiring that the Revealed Comparative Advantage (RCA)[65] in the activity exceeds 1.

S5 Micro simulation

To explore mechanisms that could generate the observed relation between coherence and city size, we develop a computational generative model that simulates how productive units select their capabilities. We assume that productive units have two options: either they copy existing capabilities, or they innovate and create a new capability. The model then simulates how the capability base of an urban economy develops with the following key parameters:

- **Capability assignment:** Each new productive unit picks an economic capability through one of two mechanisms:
  - With probability π, the unit adopts an existing capability from to those that are already present in the city.
  - With probability $1-\pi$, the individual introduces a new capability.
- **Selecting from among the pre-existing capabilities:** As in preferential attachment models, when imitating an existing capability, each available capability is selected with a probability that is proportional to its current prevalence in the city.

As a measure of coherence in this simulation, we estimate the likelihood that two randomly chosen productive units have the same capability. The simulation proceeds as described in the pseudo code described in 1.

Algorithm 1: Simulate Capability Assignment and Compute Average Skill Overlap

Input: City size N, attachment probability π
Output: Average skill overlap $\mathbb{E}_{i\neq j}[S_i \cdot S_j]$

Initialization:
Create a matrix $C \in \{0,1\}^{N\times N}$ initialized to zero ; `// Skill matrix`
Create a vector $P \in \mathbb{N}^N$ initialized to zero ; `// Skill popularity`
Set $next\_skill\_id \leftarrow 0$;
Assign skill 0 to first person:
$C[0,0] \leftarrow 1$, $P[0] \leftarrow 1$, $next\_skill\_id \leftarrow 1$;
for $i \leftarrow 1$ **to** $N-1$ **do**
    Generate a random number $r \in [0,1]$;
    if $r < \pi$ **then**
        Compute total popularity: $Z \leftarrow \sum_{k=0}^{next\_skill\_id-1} P[k]$;
        Compute probability vector: $p_k \leftarrow \frac{P[k]}{Z}$ for all k ;
        Sample a skill s from $\{0,\ldots,next\_skill\_id-1\}$ with probability p_k ;
    else
        Assign new skill: $s \leftarrow next\_skill\_id$, $next\_skill\_id \leftarrow next\_skill\_id+1$;
    end
    Assign skill to person i: $C[i,s] \leftarrow 1$, $P[s] \leftarrow P[s]+1$;
end
Compute Expected Skill Overlap:
Set $total \leftarrow 0$, $num\_pairs \leftarrow \binom{N}{2} = \frac{N(N-1)}{2}$;
for $i \leftarrow 1$ **to** $N-1$ **do**
    for $j \leftarrow i+1$ **to** N **do**
        $total \leftarrow total + C[i]\cdot C[j]$; `// Dot product = shared skills`
    end
end
return $\frac{total}{num\_pairs}$; `// Expected number of shared skills`

Fig. S5.1 now presents the simulation results under varying innovation probabilities. The left panel shows how the slope of the size-coherence relationship evolves as π increases. The right panel plots the resulting elasticity against π. It highlights that an innovation probability of $\pi = 0.03$ (red dotted line) reproduces the size-coherence slope matching our empirical observations. This analysis shows that a relatively simple generative model of local imitation and innovation can offer a mechanistic explanation for the observed relation between coherence and city size.

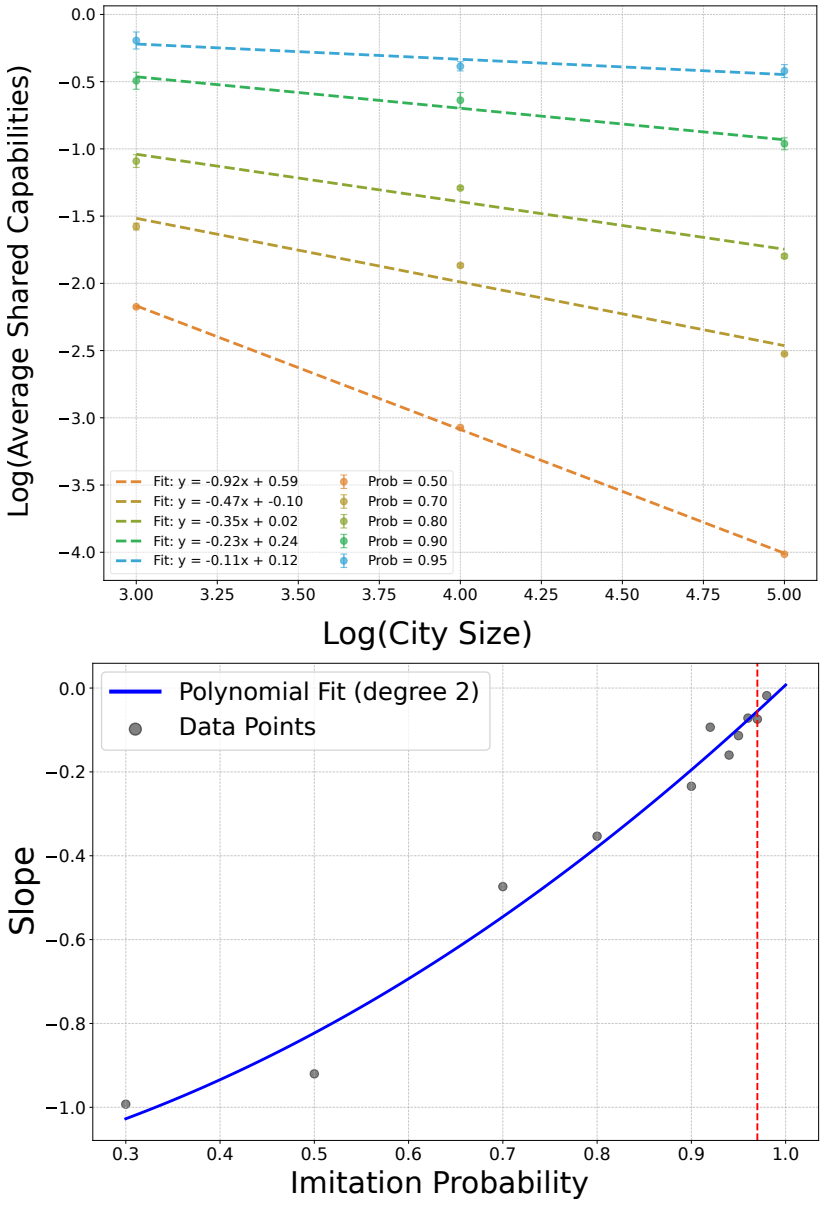


Figure S5.1. Capability simulation. The top panel shows how the slope of the size-coherence relationship changes as the imitation probability (*Prob*) increases. The bottom panel plots the estimated elasticities against the probability that units choose to imitate existing capabilities in the city. It shows that an imitation probability of 0.97 (red dotted line) produces a size-coherence slope statistically consistent with our empirical observations.