Agglomeration and Economic Growth: Does Size Really Matter?*

Matthew A. Turner† and David N. Weil‡

February 2024

Abstract: Agglomeration is often described as an engine of economic growth. We quantitatively assess this statement, focusing in particular on economies of agglomeration in two dimensions: total factor productivity and the productivity of invention. The former is a static effect that makes production in cities more efficient. The latter works dynamically, slowing the rate of productivity growth in if there is less agglomeration. We use MSA-level patent and population data since 1900 to ask how much lower output would be in the US if agglomerations had been limited in size to populations of one million, one hundred thousand, or fifty thousand. Overall, we find that such limitations would have had a surprisingly small effects on output today.

JEL: O40, R10
Keywords: Agglomeration economies, Economic growth

*The authors are grateful to Enrico Berkes for generously sharing the CUSP patent database.

†Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: Matthew_Turner@Brown.edu. Also affiliated with PERC, IGC, NBER, PSTC, S4.

‡Brown University, Department of Economics, Box B, Brown University, Providence, RI 02912. email: David_Weil@Brown.edu. Also affiliated with NBER

1. Introduction

Cities are economic dynamos. They are hubs of innovation and breeding grounds for new industries. Highly skilled workers, entrepreneurs, and scientists congregate in cities, to take advantage of the efficiencies of thick markets and the externalities associated with agglomerations. For these reasons, cities are often referred to as "engines of economic growth."

In this paper we quantitatively evaluate this idea. Our approach follows Fogel's (Fogel, 1964) analysis of the role of railroads in US economic growth. Prior to Fogel's work it had widely been noted that by the late 19th century, railroads were carrying the vast bulk of inter-regional trade, which was in turn a vital driver of growth. The natural conclusion was that railroads were thus a necessary contributor to that growth. Fogel's innovation was to note that even though railroads were in practice the dominant carrier or freight, in a world where railroads did not exist, it would have been possible for that same freight traffic to flow, only at higher cost – which he showed by constructing the transit network that could have existed in such a case. Analogously, we would like to ask how much slower US economic growth would have been, or how much poorer the country would be today, if the large cities in in which so much economic activity takes place today had not existed. This question cannot be answered simply by observing how much output is produced in cities or how much inventive activity takes place there. Had the cities not existed, much of the benefit of agglomeration would have been lost, but there still would have been skilled workers, entrepreneurs, new ideas waiting to be discovered, and so on.

Our main tools for pursuing this agenda will be explicit estimates of agglomeration effects in two specific dimension: total factor productivity and the productivity of invention. The former is a static effect that makes production in cities more efficient. The latter works dynamically, slowing the rate of productivity growth if there is less agglomeration. In both cases, we rely on estimates from existing research on the magnitudes of these effects. Using a straightforward growth model, we consider counterfactual scenarios where the degree of agglomeration – specifically, the size of the largest cities – differs from the historically observed path. The gap between income (or growth) in the counterfactual relative to the historical baseline is our measure of the growth effects of cities.

The approach that we take in this paper follows the literature on growth accounting, going back to Solow (1957). The accounting approach takes as given growth in population, human capital, and physical capital as well as, in our case, the size distribution of cities. In comparison, the full general equilibrium approach that is more common in the urban literature, there are both advantages and disadvantages. The general equilibrium approach is more explicit about the drivers of the size distribution of cities, but at the same time requires the deployment of a good deal of modeling machinery, whereas the accounting approach allows for a much more direct mapping from data to results. For example, the analysis in Duranton and Puga (2019) is based on a general equilibrium model that features agglomeration effects on both labor productivity and human capital accumulation, an urban rent gradient, commuting costs, and politically-determined restrictions on development. Counterfactual city size distribution can then be generated by considering exogenous alternations to city planning restrictions. However, in this paper, the major driver of long-run growth, which is technological change, is completely exogenous. By contrast, our analysis considers the effect of the city size distribution and technological change.

The rest of this paper is organized as follows. Section 2 introduces our counterfactual approach to assessing the importance of agglomeration and applies it to study the effect of agglomeration statically on total factor productivity in the cross section of cities. We specifically consider counterfactual scenarios in which agglomerations in the US are limited in population to one million, one hundred thousand, or fifty thousand individuals. Section 3 then takes the same approach to study technological progress, specifically using data on MSA-level patents to assess the impact of agglomeration on inventive activity at a point in time. In Section 4, we then cumulate differences in inventive activity between our counterfactual and the baseline of the actual development of the US, to calculate the reduction in TFP that would have resulted from limitations on city sizes.

2. Agglomeration and TFP in a Cross Section

There are N cities, $i = 1,...,N$. Y_{it} is city output, L_{it} is the city population, and x_{it} is a vector of *per capita* inputs into production. x_{it} can include physical capital as well as various types of human capital.

In each city, outputs are produced from labor and inputs according to

$$Y_{it} = A_{it}F_t(L_{it},x_{it}). \tag{1}$$

where A_{it} is TFP in city i decade t. We assume that $F()$ is CRS, so the production function can be rewritten,

$$Y_{it} = A_{it}L_{it}f(x_{it}). \tag{2}$$

We decompose TFP into three components: a time specific national component common to all cities, $\bar{A}_t$, a city specific agglomeration effect that depends on population, $\widetilde{A}_{it}$, and city-decade specific idiosyncratic term, $\widehat{A}_{it}$:

$$A_{it} = \widehat{A}_{it}\bar{A}_t\widetilde{A}_{it} \tag{3}$$

We assume agglomeration economies in the production of output depend on city population according to,

$$\widetilde{A}_{it} = L_{it}^{\sigma_A}. \tag{4}$$

Substituting into (2) we have

$$Y_i = \widehat{A}_{it}\bar{A}_t L_{it}^{1+\sigma_A} f(x_{it}). \tag{5}$$

This production technology nests those commonly used in to study systems of cities, e.g., Desmet and Rossi-Hansberg (2013), Duranton and Puga (2019).

Because congestion effects are not part of equation (1) and because production in the absence of the agglomeration effect is CRS, perfect mobility of all factors of production would lead to an equilibrium in which all production took place in the city with the highest value of $\widehat{A}_{it}$ We are implicitly considering equilibrium population levels that are partly determined by an unspecified congestion process, e.g., land scarcity.

Before moving on it is worth noting that in principle, if one knew the production function $f()$ and the vector of productive inputs x_{it}, and one similarly knew the productive gains from agglomeration σ_A, then it would be possible to back out relative productivity of city locations, that is $\widehat{A}_{it}$. While these would be highly desirable to know, we don't think that such a path is feasible. Among other problems, measurement error in the vector of productive inputs $x_{i,t}$ including, for example, the quality of human capital, would be reflected in the values of $\widehat{A}_{it}$ calculated this way, as would errors in the specification of the production function. As will be seen, our alternative approach allows us to avoid this measurement problem.

With this notation in place, we can begin to investigate the importance of cities to the level and growth of output. To isolate the role of agglomeration economies, we focus attention on particular counterfactual systems of cities that result when reduce the size of city i by dividing its population and all other factors of production across (possibly fractional) copies of itself. In this way, we can consider the implications of decreases in agglomeration but do not change number of people exposed to each city's idiosyncratic productivity.

Let L_{it}^{base} denote the observed population, and L_{it}^{alt} be the populations of each of the $\frac{L_{it}^{\text{alt}}}{L_{it}^{\text{alt}}}$ daughter cities in the alternative scenario. From (5) output in the base case is,

$$Y_{it}^{\text{base}} = \widehat{A}_{it}\bar{A}_t (L_{it}^{\text{base}})^{1+\sigma_A} f(x_{it}), \tag{6}$$

while total output of for set of the daughter cities is

$$Y_{it}^{\text{alt}} = \frac{L_{it}^{\text{base}}}{L_{it}^{\text{alt}}} \widehat{A}_{it}\bar{A}_t (L_{it}^{\text{alt}})^{1+\sigma_A} f(x_{it}). \tag{7}$$

Therefore, total output from the group of daughter cities relative to output of the city if had not been broken up is thus

$$\frac{Y_{it}^{\text{alt}}}{Y_{it}^{\text{base}}} = \left(\frac{L_{it}^{\text{alt}}}{L_{it}^{\text{base}}}\right)^{\sigma_A} \tag{8}$$

We assume that physical and human capital are similarly and evenly divided among these daughter cities. Thus, daughter cities have the same values of per-capita factors of production x_{it} and the same idiosyncratic productivity $\widehat{A}_i$ as the mother city, but smaller populations, and therefore less benefit from agglomeration.

To evaluate the static productivity effects of agglomeration, we consider a counterfactual in which city sizes were limited to some population $L_{\max}$. Specifically, if a city has population L_i greater than $L_{\max}$, we divide it into $\left(\frac{L_i}{L_{\max}}\right)$ "daughter cities," each with population $L_{\max}$.

We can evaluate the aggregate effect of this reduction in agglomeration economies if we sum equation (8) across all cities, noting that if the parent city population is smaller than $L_{\max}$, it will not be broken up in the alternative case.

$$\frac{Y_t^{\text{alt}}}{Y_t^{\text{base}}} = \sum_i \frac{Y_{it}^{\text{base}}}{Y_t^{\text{base}}} \min\left(\left(\frac{L_i}{L_{\max}}\right)^{-\sigma_A}, 1\right) \tag{9}$$

An advantage of this approach is that it does not require measures of city-specific physical and human capital, and similarly does not require the measurement of the city-specific productivity term $\widehat{A}_i$. Rather, all that is required is data only on city population and output.[1]

The key parameter in this analysis is σ_A, which measures agglomeration benefits in productivity. Duranton and Puga (2019) use as a baseline a value of 0.04 for the direct productivity effects of agglomeration. However, in their model there is a second benefit of agglomeration, which is the increase in human capital from experience, which is larger in larger cities. They also estimate this elasticity as 0.04. because we don't account for this effect elsewhere, like Duranton and Puga (2019), we include it as part of the benefit of agglomeration. We thus take as our baseline value $\sigma_A = .08$.

To evaluate equation (9) we also require data describing city level output and population. For our cities, we consider a set of 275 constant boundary MSAs defined to the same boundaries as Duranton and Puga (2019), along with a single non-metropolitan area that aggregates all non-metropolitan counties. We measure output using the county level

[1] Were we to try to estimate city-specific productivity in producing output, we would immediately run into serious problems of measurement error. Does a particular city produce high output relative to its measured human capital because it has a high idiosyncratic productivity due to location or institutions, or because we have failed to properly measure the quality of human capital? As will be seen in the next section, this problem would be even more severe if we tried to measure city productivity in R&D.

Table 1: Output in 2010 for three counterfactual size caps and values of σ_A.

σ_A	$L^{max} = 1m$	$L^{max} = 100k$	$L^{max} = 50k$
0.04	0.96	0.89	0.87
0.08	0.92	0.80	0.77
0.12	0.88	0.72	0.68

Note: Each cell reports the share of total output relative totals reported in the 2010 BEA data, for a a particular cap on city size and value of σ_A. For the purpose of this calculation,the rural population is treated as an extra MSA whose output is constant across scenarios.

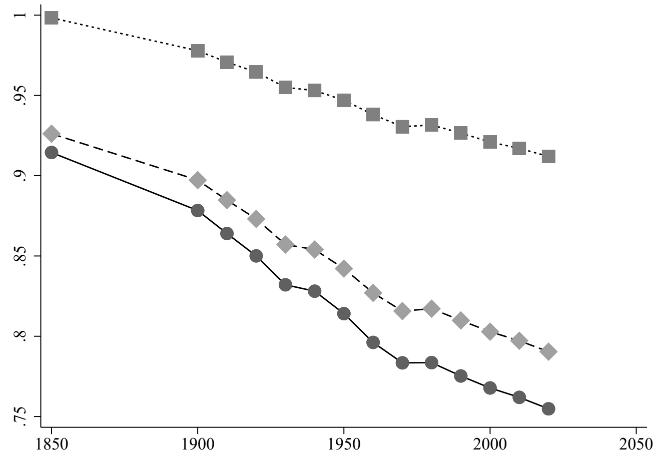
output data from the BEA (USDOC/BEA/RD, 2023), and aggregate counties to MSAs. We construct decadal population data by combining population data in replication files from Duranton and Puga (2019) with 1900-1990 county population data Forstall and NBER (1995). This results in an MSA by decade panel of MSA population stretching from 1850 to 2010, with 1860-1890 missing.

We first evaluate the static effect of agglomeration for data from the year 2010. In addition to the baseline value of $\sigma_A = 0.08$, we consider two other values: 0.04 and 0.12. Together, these values span the most of the range of agglomeration effects the literature (e.g., Combes et al. (2008), Ciccone and Hall (1996)). We similarly consider three possible values of maximum city size, L_{max}: 1,000,000, 100,000, and 50,000. The first two values are binding for 52 and 261 cities, while the cap of 50,000 requires that we split every MSA into a set of daughter cities. Non-metropolitan output does not change between realized and counterfactual cases. Table 1 presents results.

The rows of table 1 report the value of equation (9), the ratio of counterfactual to realized aggregate output, as the strength of static agglomeration economies increase. Columns describe different counterfactual systems of cities. Moving from column 1 to 3, we consider systems of cities in which cities are constrained to be smaller and agglomeration economies are less important. We see that counterfactual output is 96% of realized output when cities are allowed to be as large as 1m and agglomeration economies take their smallest value, and this share declines to 0.68 when the strength of agglomeration economies is largest and we consider a counterfactual system of cities in which no city is larger than 50,000. Table 1 suggests two conclusions. First, that in every case, agglomeration effects make an important contribution to economic output. Second, noting that the US economy has increased by about a factor of 13 since 1900, agglomeration economies are at most a moderately important contributor to the overall increase in income over this period.

These results can be extended historically, although with some difficulty. BEA data

Figure 1: Share of total output produced under three hypothetical city networks, imputed baseline output



Note: Counterfactual output as a fraction of actual outcome when city sizes are capped at 1m (squares), 100k (diamonds), and 50k (circle). Calculations are based on city output imputed from the BEA data for 2000 and population data, and assume $\sigma_A = 0.08$.

on MSA output only goes back to 2000. To look at earlier periods, we have to impute historical output. For each city i we can impute it's output in decade t from its output in 2000 and its population in decade t. In particular, let

$$Y_{it}^{0,2000} \equiv = \widehat{A}_{i,2000} \bar{A}_{2000} L_{it}^{1+\sigma_A} f_{2000}. \tag{10}$$

That is, $Y_{it}^{0,2000}$ is the output we would observe in decade t if all of the structural parameters that determine output in 2000 are held constant, but population is allowed to change to its value at decade t. We perform this calculation using the BEA data for 2000 and our population data from 1850 to 2020. Given these imputed levels of output, we can then calculate the hypothetical output that would result from imposing a cap on city size. Figure 1 presents these results.

Overall, the static productivity effects in Table 1 and Figure 1 strike us as modest. Even in the most extreme counterfactual case, in which $L_{\max} = 50{,}000$, so that there are no large, or even medium size cities, the effect on output is only 20% percent, for our baseline value of $\sigma_A = 0.08$. In the case where city size is limited to one million, which would still require a massive restructuring of the urban landscape, the reduction in output even with our high-end value of the agglomeration parameter is only 12%. Going back to 1900, when in fact the largest MSA in the country already had a population of 5.5m, all of these effects are reduced in magnitude by roughly half.

3. Productivity Growth

The analysis of the previous section takes the city invariant component of TFP, $\bar{A}_t$ as given. We now turn to this component, and in particular to the effect of agglomeration on the speed of technological progress. This channel obviously has the potential to produce a much larger effect on current output than the channel of static productivity from agglomeration, because improvements in technology accumulate over time.

We proceed in parallel with our approach in the previous section, although as will be seen we have to make adjustments to deal with the dynamics of technological change.

Define R_t as research output at time t. The use of this new terminology is required because research output will not map directly into the speed of technological progress. Specifically, as will be seen in Section 4, the speed of technological progress depends on both research output and the level of technology itself. However, at a given point in time, we assume that cross-sectional variation in research output among cities will produce proportional variation in patents per city.

Research output in a city is taken as being a function of a vector of research inputs Z_{it} as well a set of productivity terms in the production of research: a time specific national component common to all cities, $\bar{B}_t$, a city specific agglomeration effect that depends on population, $\widetilde{B}_{it}$, and city-decade specific idiosyncratic term, $\widehat{B}_{it}$. These three elements combine to give the productivity of a city in creating new ideas:

$$B_{it} = \widehat{B}_{it}\bar{B}_t\widetilde{B}_{it} \tag{11}$$

To be clear, we do not impose any particular relationship between city-specific productivity in producing output and city-specific productivity in producing research. Some places might be good for one but not the other. Similarly, we do not impose any relationship between vectors of inputs used to produce output X_{it}, on the one hand, and the inputs used to produce research Z_{it}, on the other. For example, two cities might have the same numbers of Ph.D.s working in production, while they have radically different numbers of Ph.D.s working in research.

The production function for research is

$$R_{it} = B_{it}G(Z_{it}) \tag{12}$$

where $G()$ has constant returns to scale. As above, we can rewrite this equation in terms of the vector of *per capita* inputs into R&D, z_{it} :

$$R_{it} = B_{it}L_{it}g(z_{it}). \tag{13}$$

Finally, we model the agglomeration effect in producing research output in the same way that we did for producing output, but with a different value of agglomeration economies:

$$\widetilde{B}_{it} = L_{it}^{\sigma_B}. \tag{14}$$

As in the previous section, we can examine the effect restricting city size on total national research output at a point in time:

$$\frac{R_t^{\text{alt}}}{R_t^{\text{base}}} = \sum_i \frac{R_{it}^{\text{base}}}{R_t^{\text{base}}} \min\left(\left(\frac{L_i}{L_{\max}}\right)^{-\sigma_B}, 1\right) \tag{15}$$

Given the assumption that we have made regarding the relationship between research output and patents, this equation can we rewritten as

$$\frac{R_t^{\text{alt}}}{R_t^{\text{base}}} = \sum_i \frac{P_{it}}{P_t} \min\left(\left(\frac{L_i}{L_{\max}}\right)^{-\sigma_B}, 1\right) \tag{16}$$

where P_{it} is the number of patents in MSA i in period t, and P_t is the total number in the country.

A *Research Output in a Cross Section*

We begin by considering the effect of limiting city sizes at a point in time, holding constant the time-specific component of research output, $\bar{B}$ constant. As our measure of research output, we use patents at the MSA level.

It is worth pointing out that this approach to examining the effect of limiting city sizes on research productivity represents something of an extreme case. To see why, consider the case in which there is a city of two million people, of whom 20,000 are engaged in R&D. Agglomeration effects in research presumably depend on the number of other researchers in a city, rather than the number of people overall. Thus one could imagine splitting the parent city into two daughter cities, each with one million people, but with one daughter city containing all 20,000 researchers. In that case, research output would not fall at all. By contrast, in dividing up the resources devoted to R&D proportionally with population, we have made the assumption that maximizes the effect of limiting agglomeration in research productivity.

The key parameter required for our calculation is σ_B, the effect of agglomeration on research productivity. Moretti (2021) estimates exactly this parameter at about 0.06. We use this for our baseline value and consider 0.03 and 0.09 as upper and lower bounds. This range includes the result of almost every one of the various specifications that Moretti reports, and is much wider than the 95% confidence interval around his preferred estimate of $\sigma_B = 0.0676$. We rely on the CUSP data (Berkes, 2018), to measure patents. These data report on all patents issued by the US Patent office from 1836 to 2015 along with the year of issue and county of residence for all listed inventors. Using these data,

Table 2: Patents during 2000-9 for three counterfactual size caps and values of σ_B.

σ_B	$L^{max} = 1m$	$L^{max} = 100k$	$L^{max} = 50k$
0.03	0.96	0.91	0.89
0.06	0.93	0.83	0.80
0.09	0.90	0.75	0.71

Note: Each cell reports the share of total patents during 2000-2009 relative totals reported in the CUSP data Berkes (2018), for a particular cap on city size and value of σ_B. For the purpose of this calculation,the rural population is treated as an extra MSA whose patents are constant across scenarios.

and pro-rating patents with multiple inventors, we construct county-by-year counts of patents. Because MSAs are defined as collections of counties, we can easily aggregate to counts of patents produced in each MSA during each decade, e.g. 1900-1909, from 1850 to 2010. Together with our population data, these data allow us to evaluate the aggregate production of patents for counterfactual systems of cities given by equation (16)

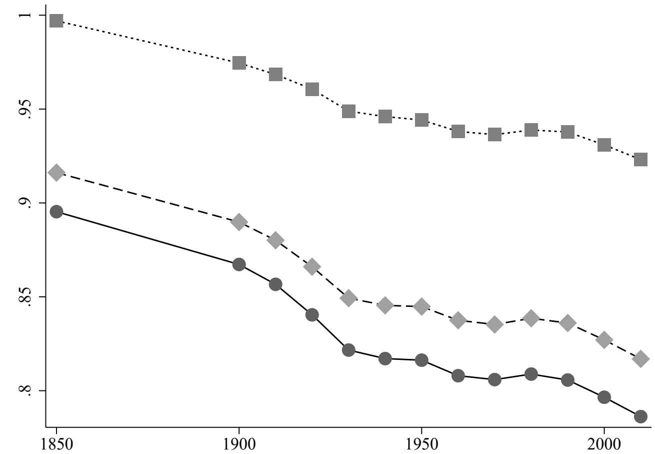
Table 2 shows national research output for 2020 in the alternative case where cities are limited in size relative to the base case of observed research output. We consider a range of values of L_{max} as in table 1 and allow σ_B to take the values 0.03, 0.06, and 0.09. When agglomeration economies in patenting are weakest and we cap city size at 1m, the output of patents fall by only 7%. In the most extreme case, where we consider the largest plausible value of $sigma_B$ and cap city size at 50k, the output of patents falls by about 29%. Magnitudes are similar to those we see in table 1. We find this surprising. If, as we had expected, patenting were more concentrated in larger cities than output, then reducing the opportunity for agglomeration economies to operate by restricting city size would be more harmful to patenting than output.

We can also examine the evolution of research output over time. For this purpose, we restrict attention to our baseline value of $\sigma_B = 0.06$ and calculate how patenting changes in each of our three counterfactual systems of cities in each decade for which we have both population and patent data. Figure 2 presents our results. This figure corresponds closely to figure 1, which reports changes in aggregate output for different networks over time, except that figure 1 is based on imputed output levels for the observed system of cities, while figure 2 does require this imputation.

4. From Research Output to the Speed of Technological Progress

Our goal is to calculate how the speed of technological progress would differ if the US were less urbanized. An immediate issue that arises is how to think about the rest of

Figure 2: Share of total patents produced under three hypothetical city networks



Note: Counterfactual patents as a fraction of actual patents reported in CUSP when city sizes are capped at 1m (squares), 100k (diamonds), and 50k(circles). Calculations assume $\sigma_B = 0.06$.

the world. In practice, ideas easily cross borders, and so if there were less effective R&D taking place in the US, the decline in research productivity would to a large extent simply be made up by a larger fraction of innovation taking place abroad. (maybe cite recent Bloom, Jones, etc. paper here).

One way of dealing with this issue would be to assume that the same restriction on agglomeration that we impose on the US was imposed on the world as a whole. Using global data on city populations and research output, we could then perform an analysis of the effect of this size restriction. Unfortunately, data on city research output at the global level is not available. As an alternative, we make the assumption technological progress in the US results only from R&D in the US. Another set of assumptions that would produce the same result would be that new technologies flow freely across borders and that the reduction in R&D input that take place in the rest of the world is of the same magnitude as that in the US. We view this as a reasonable approximation.

Bloom et al. (2020) empirically explore the relationship between productivity growth and aggregate R&D in the US over the period 1930-2015. Following Jones (2002), they formulate the relationship as follows,

$$\dot{A}_t / A_t = \alpha S_t^{\lambda} A_t^{-\beta} \tag{17}$$

where S is the number of researchers and in our setting will be equivalent to research output. The parameter λ captures the “stepping on toes" effect, whereby a the speed of technological progress may not scale linearly with research output. The parameter β captures the extent to which ideas become harder to find as more of them have been discovered. They take as their baseline assumption $\lambda = 1$ (no stepping on toes effect), and under this assumption estimate that $\beta = 3.1$ As alternative they consider $\lambda = 3/4$, in which case they estimate that $\beta = 2.4$ In the analysis that follows, we use both pairs of parameterizations. (Along balanced growth paths, the ratio of the growth rate of productivity to the growth rate of research output is determined solely by the ratio of λ to β. However, we will be looking along transition paths where the both parameter values matter independently.

We want to derive the time path of the city-invariant component of productivity $\bar{A}_t$ under the assumption that city sizes were limited.

Consider a set of observed values of $\bar{A}^{\text{base}}$ in the baseline case where city sizes where not restricted – that is, what actually happened. Call these $\bar{A}_1^{\text{base}}, \bar{A}_2^{\text{base}}, ...$ We want to derive an alternative pathway of this component of productivity, $\bar{A}^{\text{alt}}$ in the case where city sizes were restricted. These are $\bar{A}_1^{\text{alt}}, \bar{A}_2^{\text{alt}}, ...$ We assume that in period one, the level of $\bar{A}$ in the two scenarios were equal.

The time periods that we examine will be decades. We take the discrete version of the equation for technological progress, and further substitute our variable for city-invariant productivity, $\bar{A}$, as the measure of technology:

$$\Delta\bar{A}_t/\bar{A}_t = \alpha S_t^{\lambda}\bar{A}_t^{-\beta} \tag{18}$$

Rewriting this separately for the base and alternative cases,

$$\bar{A}_t^{\text{base}} = \bar{A}_{t-1}^{\text{base}} + \alpha(S_{t-1}^{\text{base}})^{\lambda}(\bar{A}_{t-1}^{\text{base}})^{1-\beta} \tag{19}$$

$$\bar{A}_t^{\text{alt}} = \bar{A}_{t-1}^{\text{alt}} + (\bar{A}_{t-1}^{\text{alt}})^{1-\beta} \tag{20}$$

To calculate the path of $\bar{A}$ in the alternative case, we proceed as follows. In the base case, given a series of values for $\bar{A}_t^{\text{base}}$ we can back out a series for S_t^{base}. Equation (15) then gives us the ratio of research output (which we take as equivalent to S) in the alternative case where city sizes are restricted relative to the base case where they are not. [2] This allows us to produce a series for S_t^{alt}. Under the assumption that $A_{base,0} = A_{alt,0}$, we can then generate a full time series for $\bar{A}$ in the alternative case relative to the base case, by forward iteration of equation (20).

Figure 3 shows the result of this calculation. In addition to the two sets of parameters considered by Bloom et al. (2020), ($\lambda = 1, \beta = 3.1$) and ($\lambda = .75, \beta = 2.4$), we also consider a "naive" parameterization of $\lambda = 1, \beta = 0$, which would imply that both the

[2]Formally, what we back out is the series for $\alpha(S_{t-1}^{\text{base}})^{\lambda}$ and what we then construct for the alternative case is the series for $\alpha(S_{t-1}^{\text{alt}})^{\lambda}$. Because we are interested only in the ratio of these two objects, the value of α is irrelevant.

Table 3: $\bar{A}_{alt}/\bar{A}_{base}$ for L_{max} = 1,000,000

Parameters	$\sigma_B = .03$	$\sigma_B = .06$	$\sigma_B = .09$
$\gamma = 1, \beta = 3.1$	0.989	0.978	0.968
$\gamma = .75, \beta = 2.4$	0.990	0.980	0.970
$\gamma = 1, \beta = 0$	0.958	0.921	0.886

Note: Each cell reports the ratio of the time-specific component of aggregate productivity, $\bar{A}$, for the year 2020 in the case where maximum city size is limited to one million, relative to the base case in which city size is not limited. The agglomeration parameter $\sigma_B = .06$.

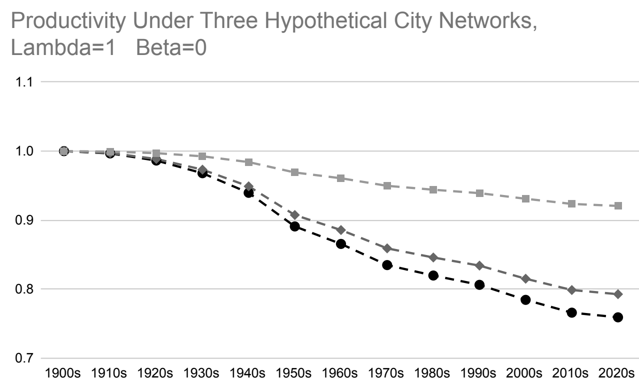
stepping on toes effect and the negative effect of current technology on the ease of finding new technologies are absent.

As the figure shows, the cumulative effect of reduced research output turns out to have a remarkably small effect on the level of productivity in the year 2020 under either of the parameterizations used by Bloom et al. (2020). When city size is limited to one million, productivity in the year 2020 is only two percent lower in the alternative case than in the baseline. Even when city size is limited to 50 thousand, the impact on the level of productivity is on the order of seven percent. This seems somewhat puzzling, given that Figure 2 shows that research output in the counterfactual cases is between 5% and 20% lower than in the base case, depending on which scenario one is looking at, for all of the decades of the twentieth century.

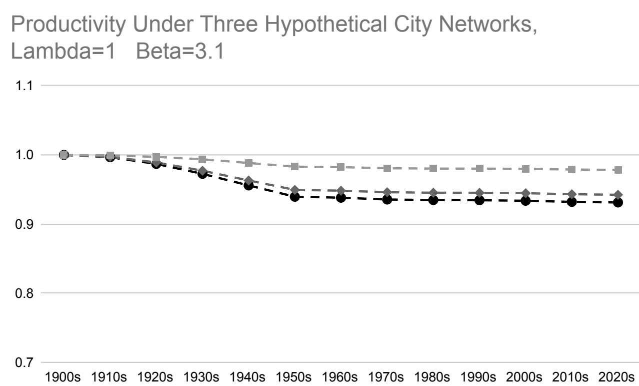
The resolution to this puzzle is exactly the negative effect of the technology level $\bar{A}$ on the speed of technological progress that is at the center of the model in Bloom et al. (2020). Less research output early in the century would have led to a lower level of $\bar{A}$, which would have in turn made research later in the century lead to faster technological progress that it did in the base case. This can be seen by examining the top panel of Figure 3 where we use the "naive" parameterization in which the effect just described is shut down. In this case, even if city size is restricted to one million, productivity in 2020 is 8% below its baseline level, while if city size is restricted to fifty thousand, the reduction in productivity in 2020 is roughly one quarter.

Table 3 shows the sensitivity of this result to value of σ_B, the parameter that measures the agglomeration effect in R&D. We focus on the case where city size in the alternative scenario is limited to one million, and consider the same combinations of λ and β that were examined in Figure 3. As the table shows, the cumulative effect of limiting city size on productivity is roughly linear in the value of σ_B. To the extent that this effect is relatively small under our baseline parameterization, it would take a very large adjustment in the parameter to produce large negative effects on productivity.

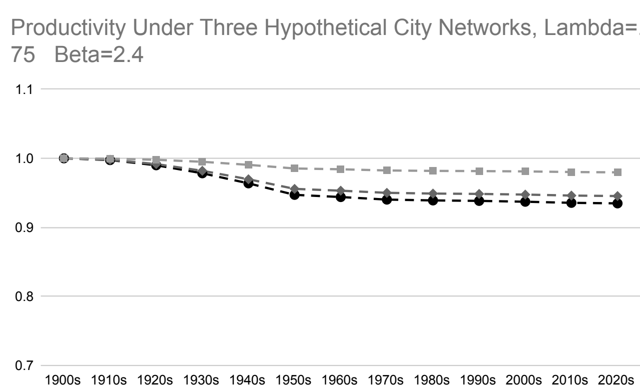
Figure 3: Share of total patents produced under three hypothetical city networks



(a)



(b)



(c)

Note: Counterfactual patents as a fraction of actual patents reported in CUSP when city sizes are capped at 1m (squares), 100k (diamonds), and 50k(circles). Calculations assume $\sigma_B = 0.06$.

5. Conclusion

In order to assess the effect of agglomeration on economic growth in the United States, we have considered the effect of counter factually limiting city sizes starting in the year 1900 on GDP per capita in the year 2020. We allow for both a static effect of city size on productivity and a dynamic effect of city size on research output, which then accumulates over time to determine the level of productive technology.

Our conclusion is the the effects of limited city size would have been surprisingly small – or put differently, that there was surprisingly little benefit from agglomeration. To give an example, consider the case in which city size was limited to one million people. Our estimate of the static productivity effect is that in this case (holding the level of technology constant), output would have been 92% of its baseline level, using our standard set of parameters. The dynamic effect of limiting city size in this fashion over the 120 year period that we consider would be that the level of technology would have been 98% of its baseline level. Multiplying these effects, output in the case with limited agglomeration would have been 10% lower that the baseline. While this is certainly not a trivial effect, it indicates that agglomeration was not the primary engine of economic growth.

References

Berkes, E. (2018). Comprehensive universe of us patents (cusp): data and facts. *Unpublished, Ohio State University*.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144.

Ciccone, A. and Hall, R. (1996). Productivity and the density of economic activity. *American Economic Review*, 86(1):54–70.

Combes, P.-P., Duranton, G., and Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of urban economics*, 63(2):723–742.

Desmet, K. and Rossi-Hansberg, E. (2013). Urban accounting and welfare. *American Economic Review*, 103(6):2296–2327.

Duranton, G. and Puga, D. (2019). Urban growth and its aggregate implications. Technical report, National Bureau of Economic Research.

Fogel, R. W. (1964). *Railroads and American economic growth*. Johns Hopkins Press Baltimore.

Forstall, R. and NBER (1995). U.s. decennial county population data, 1900-1990. Technical report. Accessed, January 2, 2024, https://www.nber.org/research/data/census-us-decennial-county-population-data-1900-1990.

Jones, C. I. (2002). Sources of us economic growth in a world of ideas. *American economic review*, 92(1):220–239.

Moretti, E. (2021). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, 111(10):3328–3375.

Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*, 39(3):312–320.

USDOC/BEA/RD (2023). Gross domestic product (gdp) by county and metropolitan area. Technical report. Accessed, January 2, 2024, https://www.bea.gov/sites/default/files/2023-12/lagdp1223.xlsx.