C.I.R.P.É.E

Centre Interuniversitaire sur le Risque, les Politiques Économiques et l'Emploi

Cahier de recherche/Working Paper **09-19**

Survival of the Fittest in Cities: Agglomeration Polarization, and Income Inequality

Kristian Behrens
Frédéric Robert-Nicoud

Juin/June 2009

Behrens: Département des sciences économiques, Université du Québec à Montréal (UQAM), Montréal, Canada ; CIRPÉE and CEPR
behrens.kristian@uqam.ca
Robert-Nicoud : Université de Genève, Département d'économie politique ; Spatial Economics Research Centre and Centre for Economic Performance, London School of Economics (LSE), London, UK ; and CEPR
Frederic.Robert-Nicoud@unige.ch

This paper extends some parts of the working papers CEPR #7018 and CEP #894, and omits some others. Sébastien Rodrigue-Privé provided excellent research assistance. Klaus Desmet, Gilles Duranton, Wen-Tai Hsu, Wilfried Koch, Muriel Meunier, Yasusada Murata, Volker Nocke, Esteban Rossi-Hansberg, Takaaki Takahashi, as well as participants at seminars and conferences at LSE, INRS Montréal, Munich, Nagoya, Passau, and Warwick provided valuable comments and suggestions. We gratefully acknowledge financial support from FQRSC Québec (Grant NP-127178). Any remaining errors are of course ours.

Abstract:
Using a large sample of US urban areas, we provide systematic evidence that mean household income rises with city size ('agglomeration'), that this effect is stronger for the top of the income distribution ('polarization'), and that household income inequality increases at a decreasing rate in city size ('inequality'). To account simultaneously for these facts, we develop a microfounded model of endogenous city formation in which urban centres select the most productive agents. Income inequality is driven by both the 'poverty' and the 'superstar' margins: whereas the least productives agents fail in a thougher urban environment, which increases 'poverty', the most productive agents become 'superstars' who reap the benefits from a larger urban market. At equilibrium, the returns to skills are increasing in city size, thereby dilating the income distribution. Our model is both rich and tractable enough to allow for a detailed investigation of when cities emerge, what determines their size, how they interact through the channels of trade, and how inter-city trade influences intra-city income inequality.

Keywords: City size, agglomeration, income inequality, heterogeneity, firm selection

JEL Classification: D31, F12, R11, R12

1 Introduction

The world's population is increasingly clustered into a few locations: since 2006, more than half of humanity is urbanized. Wealth creation and innovation are even more spatially concentrated than the distribution of population itself. Consider, for example, the three main Japanese metropolitan areas: in 1990, Tokyo, Osaka, and Nagoya together made up for a third of the Japanese population, or about 2.6% of East Asia's, but for as much as 40% of Japanese GDP and 29% of East Asia's manufacturing production (Fujita and Thisse, 2002). A back-of-the-envelope calculation therefore suggests that the inhabitants of these three metro areas were on average a third more productive than the rest of Japan, and an amazing eighteen times more productive than the rest of East Asia. As an additional example consider Figure 1a, which plots the mean household income against city population for the year 2006 in a sample of almost five hundred US cities: the raw log-log correlation of .52 is strong and statistically significant.

Insert Figures 1a–1b about here.

The cases of Japan and the US are illustrative, not exceptional: the link between city size and productivity is well documented (Rosenthal and Strange, 2004). By contrast, it is often overlooked that large cities are also more unequal and host many poor households (Glaeser, 1998). The first contribution of this paper is to uncover a series of stylized facts about city size, productivity, and the intra-city distribution of household income (Section 2). The second and main contribution is then to present a unified theoretical framework that provides a synthetic explanation showing how these facts might be the by-product of each other (Sections 3 and 4). We also extend the model to investigate how urban size, productivity and inequality interact in a system of cities that are linked through the channels of trade (Section 5).

Starting with the stylized facts, we first illustrate by way of Figure 1b that *income inequality is increasing in city size* in our US sample that includes the largest metro areas as well as smaller 'micro' areas.[1] The positive log-log correlation of .25 proves to be robust to the inclusion of many socio-economic controls, as we will establish in Section 2.

Insert Figures 2a–2b about here.

Second, income inequality also increases with city size for reasons that go beyond the socio-economic composition of cities: indeed, *the returns to city size increase along the income distribution*. To see this, we cut the income distribution of each city into quintiles. A close inspection of Figure 2a reveals that the relationship between the means of the 1st and of the 5th income quintiles and city size is statistically positive for both, and strongest for the top quintile (with log-log correlations of .24 and .56 for the 1st and the 5th quintiles, respectively). We show in

[1] On the link between MSA characteristics and income inequality, see Nord (1980), Madden (2000) and Glaeser, Resseger and Tobio (2008). This literature exclusively focuses on a subset of the largest MSAs and provides no microfounded theoretical explanations.

Section 2 that this monotonic 'dilatation' of the income distribution by city size is robust to the inclusion of various socio-economic controls. This phenomenon suggests that a 'superstar' effect *à la* Rosen (1981) might be at work, whereby the returns to skills are magnified by city size and/or the top income earners sort themselves into the largest cities. Figure 2b, which depicts the log-log correlation of .30 between the ratio of the average income of the top 5% incomes relative to the overall average income (henceforth, top-five-to-mean ratio) and city size is strongly suggestive of the existence of such an effect.

Our third stylized fact reveals that focusing on large cities, as the bulk of the literature has done to date, hides interesting patterns arising for smaller urban areas (Michaels, Rauch and Redding, 2008, provide a remarkable exception, albeit in a different context). To highlight this, we revisit the size-productivity relationship, which finds a persistent backing in the empirical literature: the elasticity of labor and firm productivity with respect to city size or density is positive and typically falls in the 3% – 8% range (e.g., Ciccone and Hall, 1996; Rosenthal and Strange, 2004). The reasons usually put forth are that city size makes workers more productive via various 'agglomeration economies' (e.g., Marshall, 1890; Duranton and Puga, 2004) and because the most productive agents sort themselves into the largest cities (e.g., Combes, Duranton and Gobillon, 2008; Mion and Naticchioni, 2009). Our contribution to this body of work is to establish a quantitative distinction between the small and the large cities of our sample ('micro' and 'metro' areas, respectively). We uncover that *the elasticity of mean income with respect to city size is larger for micro than for metro areas*. By way of quantile regressions, we also find that this elasticity decreases monotonically as we move from the bottom to the top quintile of the whole sample. To summarize our stylised facts, a large size contributes to both productivity and inequality and this contribution is strongest for the smallest cities.

The second and main contribution of the paper is to provide a theory that can account for the multiple links between city size and productivity (what we refer to as 'agglomeration' for short), the differential returns to skill ('polarization'), and the unequal distribution of income ('inequality'). While some of these features have been addressed individually in the literature there is, to the best of our knowledge, as yet no theory that addresses them simultaneously. For example, Rosen's (1981) pioneering work focusses on imperfectly competitive markets with quality-differentiated sellers and predicts that a larger market size leads to the entry of less productive agents so that the average productivity of sellers falls, tough the income distribution gets more skewed. This runs counter the stylized fact that larger cities (markets) are on average more productive (competitive). Furthermore, the size of the market is considered as exogenously given in his analysis. However, one of the key questions in an urban setting is to investigate what determines a city's equilibrium size. In our model, which blends heterogeneous managerial talent or skills in the spirit of Lucas (1978) and Rosen (1981) with functional forms taken from the heterogeneous firm models put forth by Asplund and Nocke (2006) and Melitz and Ottaviano (2008), larger cities are places that make workers and firms more productive yet where *failure*

is more likely than elsewhere because of tougher selection. The market size is endogenously determined by individual location decisions of entrepreneurs who are indifferent between entering the city or not at equilibrium: tougher selection in larger cities is ex ante compensated for by the more than proportional returns for the successful, thereby generating ex post larger income inequalities. In sum, the model suggests that urban productivity and polarization are a by-product of the survival of the fittest in a tough, competitive environment.[2]

Since city size has a strong impact on both productivity and inequality, and as size itself must be endogenously determined, we require a model in which we can tackle the questions of when cities *emerge* and what determines their equilibrium *size*. Our theoretical framework allows us to parsimoniously deal with these issues. Starting with a single city, we link the equilibrium city size, the average productivity, and the resulting income distribution to a few underlying key parameters such as the dispersion of skills, commuting costs, and the degree of product differentiation. We then extend the model to multiple cities to investigate how they *interact* with one another through the channels of trade. One key insight in this extended setting is that the positive relationship between city size and productivity becomes an equilibrium relationship in a system of cities, and that larger cities provide a larger array of goods and services (which is reminiscent of central place theory *à la* Lösch, 1940). Focussing on symmetric equilibria, in which all cities are identical, we show that the comparative static results derived in the single-city setup carry over to this new environment. We also show that lower inter-city trade costs are conducive to city formation and city growth: access to larger markets, brought about by transport innovations or by trade liberalization, increase the prospect of urbanization and the equilibrium city sizes. Income inequality also increases with trade openness at the symmetric equilibrium, but for reasons that are different from those unveiled in the international trade model of Helpman, Itskhoki and Redding (2008), who focus on labor market frictions.

The remainder of the paper is organized as follows. Section 2 analyzes the three stylized facts pertaining to US urban areas in 2006 in more detail and establishes their robustness. Section 3 then introduces the basic model and derives the equilibrium conditions. Section 4 deals with the single-city case, whereas Section 5 extends the model to multiple cities and trading networks. Section 6 concludes. We relegate most proofs, the guide to various calculations, as well as some extra material, to an extensive set of appendices.

[2]There are to our knowledge only three papers that investigate a subset of these issues in a spatial context. First, as pointed out to us by Esteban Rossi-Hansberg, the model in Lucas and Rossi-Hansberg (2002) should also imply a positive relationship between city size and income inequality because of the presence of externalities and a CBD with a spatial dimension. This relationship may be increasing at a decreasing rate as the emergence of new subcentres curbs the rent and wage gradients. Second, Combes, Duranton, Gobillon, Puga and Roux (2009) embed reduced-form agglomeration and dilatation forces in the Melitz and Ottaviano (2008) framework. Their focus is mostly empirical and aims at disentangling selection effects from agglomeration economies. A complementary approach is Okubo (2009), who casts a heterogeneous firm model into a 'new economic geography' framework to investigate the equilibrium patterns of agglomeration. Both Okubo and Combes *et al.* disregard urban structure.

2 Three stylized facts

Using data for 499 US Core Based Statistical Areas (henceforth CBSAs, or 'cities' for short), this section uncovers some new macro evidence on the links between city size and mean income ('agglomeration'), city size and the income structure ('polarization'), and city size and income distribution ('inequality'). A detailed description of the data, as well as summary statistics, are relegated to Appendix A and to Table 1. It is worth stressing from the outset that our aim is to highlight the various correlations in the data, but to keep the identification of any causal relationship among the variables for future work.

2.1 Agglomeration

Productivity rises with city size or density. The reasons put forward by the literature are essentially subsumed by 'Marshall's trinity': human density is conducive to better matching in thicker labor markets, the transmission of knowledge spillovers, and the sharing of intermediate inputs and infrastructure (for surveys, see Duranton and Puga, 2004, on theory; and Rosenthal and Strange, 2004, on empirics). Furthermore, productivity increases with city size also because the most highly skilled workers sort themselves into large metropolitan areas. As productivity and earnings rise with city size, so do household incomes, of which labor earnings constitute the largest component.

Insert Table 2 about here.

The visual findings of Figure 1a are confirmed by a simple OLS regression of average household income on city size, which yields an elasticity of about .1 (column 1 of Table 2). Two words of caution about the suggestive evidence illustrated by Figure 1a are in order. Firstly, the income of a household depends on its composition and on the individual earnings of its members. Therefore, average houshold income in a city depends on the presence of multiple-earner or single-person households in that city (Madden, 2000). Secondly, the ideal way to capture the size-productivity relationship directly is to estimate the response of individual earnings to city size by controlling for worker characteristics and unobserved heterogeneity (Wheeler, 2001; Combes *et al.*, 2008; Bacolod, Blum and Strange, 2009). This can be achieved using panel microdata but has the drawback of reducing the number of small cities that can be included in the sample as only few individual observations are available for them.[3] To keep as many CBSAs as possible in the sample, while controlling for cross-city heterogeneity that is likely to affect income and productivity, we

[3]There are only few observations for small cities in microdata samples such as the Current Population Survey (CPS) or the 5 percent Public Use Micro Sample (PUMS). Madden (2000) only works with the 182 largest MSAs, whereas Glaeser, Resseger and Tobio (2008) work with 242 of them. Our results reveal that the size-productivity-inequality relationships are strongest for small and medium-sized cities, which should thus be part of the analysis.

regress citywide average household income on measures of household composition, educational attainment, ethnic composition, poverty, geographic location, and industrial structure of the city. Table 2 (colums 2–8) reports the results. The positive size-income relationship survives the inclusion of all these controls and remains highly significant: doubling city size raises the average household income by about 3–4%, which is in line with previous findings in the literature. Comparing colums 6 and 7 of Table 2, we see that *the size-income relationship is twice as strong for the micropolitan statistical areas than for the metropolitan statistical areas.* This finding vindicates our choice of including the medium-sized cities ('micro areas') in establishing the stylized facts. Whereas doubling city size raises mean household income by 3.9% for the metro areas, it raises mean household income by 7.7% for the micro areas. This finding is further confirmed by Table 3, which summarizes the quantile regressions of mean household income on city size and the aforementioned controls. The size elasticity of mean income is significant for all quintiles and monotonically decreasing: doubling city size raises mean household income by 5.5% in the bottom quintile but only by 2.1% in the top quintile.

Insert Table 3 about here.

When taken together, our results suggest that there is a positive and highly significant relationship between city size and mean income (**Stylized Fact 1a**), and that this relationship is more important for smaller cities than for bigger cities (**Stylized Fact 1b**). In sum, the relationship between mean income and city size is increasing and concave (**Stylized Fact 1** for short).

2.2 Polarization

Does city size benefit disproportionately some individuals? Recall from Figure 2a that *the higher income quintiles benefit more from increases in city size than the lower income quintiles do.* Put differently, city size widens the gap between the means of the different income quintiles. Wheeler (2001) shows that the returns to city size are increasing in the level of educational achievement. Going further, Bacolod *et al.* (2009) show that this urban wage premium is larger for cognitive skills than for interactive skills (i.e., skills that ease face-to-face interactions, or 'people skills' in their terminology), which is in turn larger than the urban wage premium for motor skills and physical strength. Insofar as income, educational achievement and skills are highly correlated, we may view these results as the 'micro' counterparts to our 'macro' findings. As can be seen from Table 4, this is confirmed by regressing the mean household incomes by income quintile on city size: controlling for the same demographic and economic variables as in Tables 2 and 3, we find a monotonically increasing and highly significant relationship between mean income by quintile and city size. A simple F-test of equality reveals that the coefficients on city size of the first and the fifth quintiles are statistically different at the 1 percent level.

Insert Table 4 about here.

A larger city size thus does not benefit all inhabitants equally, but the benefits predominantly accrue to the agents higher up in the income distribution. The fact illustrated by Figure 2b that *city size dilates the upper tail of the income distribution* is confirmed by regressing the top-five-to-mean ratio on city size and our list of controls.

Insert Table 5 about here.

Table 5 presents the results. As can be seen, larger cities have higher top-five-to-mean ratios than smaller cities: doubling city size roughly increases the ratio of the top 5% mean income to the overall mean by 1.5% to 3.7%. Clearly, 'superstars' live in New York City or in the Bay area around San Francisco, not in micro areas, and city size is associated with a 'polarization' of the income distribution.

To summarize, city size is positively associated with a dilatation of the income distribution (**Stylized Fact 2a**). In particular, larger cities have disproportionately higher mean incomes for the top 5% of the income distribution (**Stylized Fact 2b**). These findings suggest that there may be a strong and direct link between city size and income inequality.

2.3 Income inequality

We measure income inequality using the household income Gini coefficient.[4] Recall from Figure 1b that the relationship between size and income inequality is positive, as suggested by the existence of polarization. See also Long, Rasmussen and Haworth (1977) and the contemporaneous paper by Glaeser, Resseger and Tobio (2008) on the positive correlation between city size or density and income inequality.[5] To get a sense of what is driving inequality, it is instructive to first regress the Gini coefficient on the poverty rate and on the top-five-to-mean ratio. Using OLS with standardized regression coefficients, we find that an increase of one standard deviation in the top-five-to-mean ratio leads to a .78 standard deviation increase in the income Gini, whereas a one standard deviation increase in poverty leads to a .33 standard deviation increase in the income Gini (both coefficients are significant at the 1 percent level). In words, more 'poverty' (which affects the left tail of the income distribution) and more 'superstars' (which affects the right tail of the income distribution) both increase income inequality in cities, yet superstars seem to contribute more to measured income inequality than poverty does. Using the interquantile distance as an alternative inequality measure yields similar results, with standardized regression coefficients of .81 for the top-five-to-mean ratio and .32 for the poverty rate (again significant

[4]As an alternative measure, we compute the gap between the fifth and the first income quintile means and take its ratio to the overall mean. As can be seen from Table 1, both measures are very strongly correlated, so that we can only focus on the Gini coefficient in what follows.

[5]These authors use the five percent Integrated Public Use Micro Samples from the census years 1980 and 2000 to construct income Gini coefficients. They find that the partial correlation between the log Gini coefficient of the household income distribution and log population density is positive.

at the 1 percent level). Table 6 reports regression results for measures of income inequality on city size and our list of controls. The coefficient on city size is almost always significant, with elasticities in the .01 to .03 range: doubling city size increases measured income inequality by about 1% to 3%. Comparing the results in columns 6 and 7 of Table 6, we also see that the impact of city size on inequality is roughly four times as large for the micro areas than for the metro areas, though the precision of the estimates decreases with the smaller sample sizes. This finding suggests that the relationship between city size and income inequality is increasing and concave (**Stylized Fact 3**).

Insert Table 6 about here.

The foregoing finding is further confirmed by Table 7, which presents results of quantile regressions of the household income Gini on city size and our controls. As can be seen, the coefficient on city size is almost monotonically decreasing as we move up the quintiles, thus suggesting that city size becomes less important for explaining income inequality in larger cities. A simple F-test of equality reveals that the coefficients on city size of the first and the fifth quintile are borderline statistically different at the 1 percent level.

Insert Table 7 about here.

To conclude this section, let us emphasize that our controls for industrial composition, namely an Isard index of industrial diversification and the share of employment in the higher level service sectors, do not appear to significantly influence income inequality whereas they influence mean income in cities (compare Tables 2 and 6). Thus, while specialization and industry structure do matter for productivity, income inequality seems to be more driven by within-sector than by between-sector heterogeneity (Lemieux, 2006). For this reason, we will build a model with a representative sector which, by construction, explains the relationship between inequality and city size by abstracting from compositional issues related to industry structure.[6]

3 The model

We start by sketching the model. There are Λ regions, labeled $l = 1, 2, \ldots, \Lambda$, and variables associated with each region will be subscripted accordingly. Each region has a large and fixed population L_l of ex ante undifferentiated workers. All workers are endowed with some amount of a numéraire good and one unit of labor that they can use either for producing the numéraire good as unskilled workers, or for becoming skilled entrepreneurs. Becoming an entrepreneur involves

[6]It is well known that the industrial mix of larger cities is more diverse than that of smaller ones (e.g., Henderson, 1988 and 1997). Duranton and Puga (2005) show that the same holds increasingly true for the 'functional mix' of cities. In both cases, these facts regard the horizontal diversification of cities, where productivity differences may be driven by compositional effects. In this paper, we show that large cities are vertically differentiated as well, i.e., productivity differences are driven by individual effects even in the absence of compositional effects.

a net entry (or education) cost and requires that the worker moves to a city which provides the adequate environment for acquiring human capital and starting a business. 'Learning in cities' which serve as 'incubators for innovation' is in accord with empirical evidence (Glaeser and Maré, 2001; Duranton and Puga, 2001) and the fact that most universities are located there. Becoming an entrepreneur entails however a risk of failure, in which case the agent is stuck in the city, does not produce, and consumes solely from her initial endowment. There are thus three types of agents at equilibrium: successful skilled agents in the cities (entrepreneurs); unsuccessful skilled agents in the cities; and unskilled agents who stay in the rural area.

The economy has two sectors. The first one produces a continuum of varieties of a horizontally differentiated good or service, whereas the second one produces a homogenous good. Production of the homogenous good requires no entrepreneurial skills, occurs under constant returns to scale, and takes place outside the city. Furthermore, the homogenous good is traded in a competitive market that is perfectly integrated. Hence, its price is equalized across regions, which makes this good a natural choice for the numéraire. Perfect competition ensures that marginal cost pricing prevails, which implies a unit wage everywhere as long as the homogenous good is produced in all regions, which we henceforth assume to be the case. The differentiated good is produced by the entrepreneurs using entrepreneurial skills and the numéraire good. The latter is obtained either from the entrepreneur's endowment or from the countryside. Trading the differentiated good across cities is costly.

Previewing our subsequent results, only those entrepreneurs who are productive enough survive and produce at equilibrium, whereas low-ability entrepreneurs leave the market immediately without setting up production at all. The minimum ability that entrepreneurs have to achieve to survive is an equilibrium feature of the model that we refer to as *selection*. Entry into the city occurs in reponse to economic opportunities on both the production and the consumption side, which depend largely on the ability threshold required for producing successfully. We view the determination of city size $H_l \leq L_l$ at equilibrium as the result of a tension between agglomeration and dispersion forces that will be made precise below.

3.1 Timing

There are two stages. In the first one, risk-neutral workers decide whether or not to become entrepreneurs, in which case they incur the sunk entry cost $f^E \geq 0$ (paid in terms of the numéraire and including the opportunity cost of foregoing the unskilled wage), or to stay as uneducated workers in the countryside. Henceforth, superscript 'E' is a mnemonic for 'entry' or 'education', whereas superscript 'U' is a mnemonic for 'unskilled' or 'uneducated'. Interpreting f^E as a cost to acquiring education, this means that agents decide first whether to acquire skills or not and, if so, become urban dwellers, and only then learn their ability. Living in a city gives rise to extra costs and benefits, which will be made precise below. Once the education-cum-location

decision is taken, *nature* attributes to each entrepreneur a horizontal characteristic ν and a vertical characteristic c: we think about the former as her product variety (or her skill *type*) and about the latter as her entrepreneurial ability (or her skill *level*). Specifically, entrepreneurs discover a variety ν and nature draws the marginal cost c at which they can produce this variety from some common and known distribution g. Upon observing their draw, entrepreneurs chose whether to produce or not and to which markets to sell.

Entrepreneurial skills are an indivisible fixed input and we assume that production occurs where entrepreneurs live, i.e., entrepreneurs who entered market l in the first stage produce and consume in city l. Our maintained assumption that all agents are immobile and do not relocate to another city after having observed their skill level implies that we disregard issues of spatial sorting along skills in the model. We do so mostly for analytical convenience; actually, selection and agglomeration generate higher productivity in larger cities regardless of sorting, so adding sorting to our model would reinforce our results. We also do so on theoretical and empirical grounds: our model with sorting would share the counterfactual property of a 'perfect sorting', like virtually any existing model of spatial sorting (e.g., Mori and Turini, 2005; Nocke, 2006). Also, recent empirical evidence suggests that while workers in larger cities are more educated and more skilled than are those in smaller cities, they are so to a modest degree only (Berry and Glaeser, 2005; Bacolod *et al.*, 2009). Last, we also assume that upon observing their skill level, unsuccessful agents do not migrate back to the countryside. Allowing for this would imply that the income distribution is independent of the city size, which runs counter the stylized facts highlighted in the foregoing section (see Appendix T.2 for additional details).

In the second stage, entrepreneurs set profit maximizing prices and all markets clear. We solve the game for subgame perfect equilibria by backward induction.

3.2 Preferences, demand, and urban structure

Following Asplund and Nocke (2006) and Melitz and Ottaviano (2008), all agents have identical quasi-linear preferences over the homogenous good and the varieties of the horizontally differentiated good. Furthermore, each agent is endowed with $\overline{d}^0$ units of the numéraire. Varieties of the differentiated good available in region l are indexed by $\nu \in \mathcal{V}_l$. In what follows, we denote by $\mathcal{V}_{hl}$ the set of varieties produced in h and consumed in l; by $\mathcal{V}_l^+ \subseteq \mathcal{V}_l \equiv \cup_h \mathcal{V}_{hl}$ the subset of varieties effectively consumed at equilibrium in region l; and by N_l the measure of $\mathcal{V}_l^+$ (i.e., the mass of varieties consumed in l). The subutility over the differentiated varieties is assumed to be quadratic, so that utility for a resident in region l is given by:

$$U_l^i = \kappa^i \left\{ \alpha \int_{\mathcal{V}_l} d_l(\nu) \mathrm{d}\nu - \frac{\gamma}{2} \int_{\mathcal{V}_l} [d_l(\nu)]^2 \, \mathrm{d}\nu - \frac{\eta}{2} \left[\int_{\mathcal{V}_l} d_l(\nu) \mathrm{d}\nu \right]^2 \right\} + d_l^0, \tag{1}$$

where $\alpha, \eta, \gamma > 0$ are preference parameters; where d_l^0 and $d_l(\nu)$ stand for the consumption of the numéraire and of variety ν, respectively; and where $\kappa^i = 1$ if $i = E$ (the agent lives in the

city) or $\kappa^i = 0$ if $i = U$ (the agent lives in the countryside). We assume that the differentiated good is sold and consumed exclusively in the cities, whereas the homogenous good is available and consumed everywhere. The parameter κ^i captures this assumption in a parsimonious way.[7]

Since marginal utility at zero consumption is bounded for each variety, urban dwellers will in general not have positive demand for all of them. In what follows, we assume that all agents have positive demand for the numéraire, i.e., $d_l^0 > 0$, which rules out income effects on the differentiated good. A sufficient condition for this to hold is that the initial numéraire endowment $\overline{d}^0$ is large enough, which we henceforth assume to be the case.

All entrepreneurs reside in a monocentric city and, therefore, pay commuting costs and land rents. Furthermore, there is at most one city per region.[8] The aggregate land rent is redistributed among urban dwellers, each of whom has a claim to an equal share of it. The urban costs (commuting plus housing) in region l, when its size is H_l, is given by θH_l, where $\theta > 0$ is a parameter positively related to commuting costs (see Appendix T.1). In sum, becoming an urban dweller involves two types of costs: the urban costs proper, namely θH_l, and the entry cost, namely f^E. Let $\Pi_l(c)$ denote the entrepreneurial profit of an agent with ability c in city l. The budget constraint is then given by

$$\kappa^i \left[\int_{\mathcal{V}_l} p_l(\nu) d_l(\nu) \mathrm{d}\nu + \theta H_l + f^E \right] + d_l^0 = w_l^i(c) + \overline{d}^0, \tag{2}$$

where $w_l^i(c) = w_l = 1$ if $i = U$, and $w_l^i(c) = \Pi_l(c)$ if $i = E$. In the latter case, the entrepreneur's income also depends on her inverse ability c, as will be made clear below.

Maximizing (1) subject to (2) allows us to express the indirect utility of a type-i agent in l as $V_l^i(c) = w_l^i(c) + \kappa^i \mathrm{CS}_l + \overline{d}^0$, where CS_l denotes the consumer surplus (see Appendix T.3 for computational details).

3.3 Production

We assume that markets are segmented and that entrepreneurs are free to price-discriminate. The *delivered cost* in city h of a unit produced with marginal cost c in city l is τc, with $\tau > 1$ if $h \neq l$ and c if $h = l$. Hence, $(\tau - 1)c$ may be interpreted as the frictional trade cost incurred in

[7] This assumption is a short-cut, the purpose of which is to make the market size for the differentiated good endogenous while retaining a simple model. A more elegant microfoundation would be to assume that shipping varieties from the city to the rural areas entails some cost. Assuming prohibitive trade costs may be a strong assumption, yet it is worth keeping in mind that a large share of urban output is made-up of non-tradable consumer services (restaurants, cinemas, theaters). Also, many tradables like branded products can be aquired only in downtown arcades or suburban shopping malls carrying large inventories.

[8] The first assumption is made for analytical convenience. The key element is that urban costs rise with city size, a property that is also encountered in non-monocentric city models like Fujita and Ogawa (1982) or Lucas and Rossi-Hansberg (2002). The second assumption is made without loss of generality since we can always redefine 'regions' so that there is, indeed, only at most one city in each.

transporting a unit of any variety of the differentiated good across the two cities. We interpret such a cost broadly as stemming from all distance-related barriers to the exchange of goods, and we assume that this cost is symmetric across all city-pairs. We impose symmetry for two reasons. Firstly, it eases the analysis and the notational burden without significantly modifying our main theoretical insights. As shown among others by Tabuchi, Thisse and Zeng (2005), the general case of an urban hierarchy leads to a complex taxonomy and only allows for clear-cut results in a few special cases involving at least some symmetry. Secondly, making these assumptions underscores that our model is rich enough to generate various urban configurations *despite all cities sharing a priori the same symmetric fundamentals*.

The variable production component requires using the ubiquitous numéraire good as an intermediate input so that the cost of an entrepreneur in l with a draw c is given by $c\big(q_{ll}+\tau\sum_{l\neq h} q_{lh}\big)$, where q_{lh} is output produced in l and sold in h.

3.4 Parametrization

To obtain clear analytical results, we henceforth assume that productivity draws $1/c$ in all regions follow a Pareto distribution with lower productivity bound $1/c_{\max}$ and shape parameter $k \geq 1$. This implies a distribution of cost draws given by:

$$G(c) = \left(\frac{c}{c_{\max}}\right)^k, \quad c \in [0, c_{\max}].$$

The shape parameter k is related to the dispersion of cost draws. When $k = 1$, the cost distribution is uniform on $[0, c_{\max}]$. As k increases, the relative number of low productivity firms increases, and the productivity distribution is more concentrated at these low productivity levels. Any truncation of the Pareto distribution from above at $c_l < c_{\max}$ is also a Pareto distribution with shape parameter k. To avoid a taxonomy of special cases that involve corner solutions and that do not add any additional insights, we impose $\alpha < c_{\max}$ in what follows. This assumption implies that an isolated entrepreneur who gets a really bad draw is not productive enough to remain active at equilibrium.

3.5 Market outcome

Let $p_{hl}(c)$ and $q_{hl}(c)$ denote the price and the quantity sold by an entrepreneur with inverse productivity c when she produces in region h and serves region l. Since markets are segmented and marginal costs are constant, operating profits earned from sales to different regions are independant from one another. Let $\pi_{hl}(c) = [p_{hl}(c) - \tau c]\, q_{hl}(c)$ denote these operating profits, expressed as a function of c.

Each firm sets profit-maximizing prices, taking the other firms' equilibrium pricing strategies as given. Profit maximization may thus be described by the following for each entrant: a pricing

strategy $p_{hl}(c)$, i.e., a mapping $c \in \mathbb{R}_+ \to \{p_{hl}(\cdot)\}_{l=1}^{\Lambda} \in \mathbb{R}_+^{\Lambda}$; and Λ 'entry-or-exit' decisions, i.e., a mapping $c \in \mathbb{R}_+ \to \{\mathrm{I}_{hl}(\cdot)\}_{l=1}^{\Lambda} \in \{0,1\}^{\Lambda}$. Obviously, both depend on her marginal cost c. We show in Appendix T.4 that only the most efficient firms make positive profits, whereas the least productive ones chose to exit (Lucas, 1978; Asplund and Nocke, 2006; Melitz and Ottaviano, 2008). More precisely, only entrepreneurs with inverse productivity c smaller than some cutoff c_l are productive enough to sell in city l. As shown in Appendix T.4, the Nash equilibrium prices can be expressed as follows:

$$p_{hl}(c) = \frac{c_l + \tau c}{2}, \quad \text{where} \quad c_l \equiv \frac{2\alpha\gamma + \eta N_l \overline{c}_l}{2\gamma + \eta N_l} \quad \text{and} \quad \overline{c}_l = \frac{k}{k+1} c_l \tag{3}$$

denote the *domestic cost cutoff in region* l and the average marginal cost of active entrepreneurs, respectively. The consumer price is decreasing in the degree of competition in the destination market, which is inversely related to c_l (see (5) below). For each pair of cities l and h, there exists an *export cost cutoff* c_{lh} such that only entrepreneurs with c lower than c_{lh} export from l to h. This cutoff must satisfy the zero-profit cutoff condition $c_{hl} = \sup\{c \mid \pi_{hl}(c) > 0\}$, which can be expressed as either $p_{hl}(c_{hl}) = \tau c_{hl}$ or $q_{hl}(c_{hl}) = 0$, which from (3) yields:

$$c_{hl} = \frac{c_l}{\tau}. \tag{4}$$

Expression (4) implies that $c_{hl} \leq c_l$ since $\tau \geq 1$. Put differently, trade barriers make it harder for exporters to break even relative to their local competitors because of higher market access costs. Using (3), the mass of entrepreneurs selling in region l is given as follows:

$$N_l \equiv \sum_h H_h G(c_{hl}) = \frac{2\gamma(k+1)(\alpha - c_l)}{\eta c_l}. \tag{5}$$

Note that (5) establishes a positive equilibrium relationship between the number of competitors selling in city l and the toughness of selection there: only the entrepreneurs with productivity larger than $1/c_l$ survive. *The larger the number of competitors, the smaller the share* $G(c_l)$ *of entrepreneurs that are fit enough to survive.* Accordingly, we refer to $1 - G(c_l)$ as the 'failure rate' in the urban market. Using (3) and (5), the consumer surplus can be expressed very compactly as follows:

$$\mathrm{CS}_l \equiv \mathrm{CS}(c_l) = \frac{\alpha - c_l}{2\eta} \left(\alpha - \frac{k+1}{k+2} c_l \right). \tag{6}$$

Thus, c_l is a sufficient statistic to analyze the impact of any policy or parameter change on consumer welfare. Clearly, $\partial \mathrm{CS}_l / \partial c_l < 0$ if $c_l \leq \alpha$ (which holds at equilibrium).[9]

[9] To see this, note first that α is the demand intercept (Appendix T.4) and assume $c_l > \alpha$ for some l. Then, there is a strictly positive mass of entrepreneurs who have a c larger than α and who make negative operating profits as a result. This establishes a contradiction with profit maximisation. We thus conclude that $\alpha > c_l$ holds at any equilibrium.

3.6 Equilibrium

It is readily verified that

$$\begin{aligned}\Pi_l(c) &= \sum_h \left[p_{lh}(c) - \tau c\right] q_{lh}(c) \\ &= \mathrm{I}_{ll}(c)\frac{H_l}{4\gamma}(c_l - c)^2 + \sum_{h\neq l} \mathrm{I}_{lh}(c)\frac{H_h}{4\gamma}(c_h - \tau c)^2, \end{aligned} \tag{7}$$

where $\mathrm{I}_{lh}(c) = 1$ if $c < c_{lh}$ and $\mathrm{I}_{lh}(c) = 0$ otherwise. We define a *short-run equilibrium* as a situation in which, contingent on entry decisions summarized by the Λ-dimensional vector $\{H_l\}_{l=1}^{\Lambda}$, the following holds: (i) entrepreneurs decide whether to produce or not and set prices so as to maximize profits; (ii) consumers maximize utility; and (iii) the masses of sellers obey the identity (5). The latter identity can be rewritten as:

$$\frac{\alpha - c_l}{A\eta c_l^{k+1}} \equiv H_l + \tau^{-k}\sum_{h\neq l} H_h, \tag{8}$$

where $A \equiv 1/[2c_{\max}^k(k+1)\gamma]$ is a recurrent bundle of parameters that captures the underlying productivity of the economy: A is decreasing in the upper bound $c_{\max}$ of the support of G and in the shape parameter k. As k rises, the mass of low-productivity entrepreneurs rises relative to the mass of highly productive ones. Observe that A also encapsulates the 'desirability' of the differentiated good in the sense that it contains γ, which inversely captures consumers' preference for variety: a larger γ implies that the good is less differentiated.

The indirect utility differential for a worker with entrepreneurial ability c between remaining unskilled in the countryside or becoming an urban entrepreneur in city l is given by:

$$\Delta V_l(c) \equiv \Pi_l(c) + \mathrm{CS}_l - \theta H_l - f^E. \tag{9}$$

A worker decides to become an urban entrepreneur if her expected indirect utility is larger than the (certain) equivalent that she could secure in the numéraire sector in the countryside. Formally, this is so when $\mathbb{E}(\Delta V_l) \geq 0$. Entry into the city takes place as long as it is profitable, i.e., $\mathbb{E}(\Delta V_l) \leq 0$ must hold at equilibrium, which we henceforth refer to as the *free-entry condition*. In words, expected profits, net of urban and entry costs, are non-positive at equilibrium.

Prices adjust more quickly than entry decisions. We thus define a *long-run equilibrium* (an *equilibrium* for short) as a 2Λ-tuple $(\{H_l, c_l\}_{l=1}^{\Lambda})$ such that the free-entry and the short-run equilibrium conditions hold simultaneously. In other words, at an equilibrium: (i) entrepreneurs maximize profits; (ii) consumers maximize utility; (iii) the masses of sellers obey (8); and (iv) agents decide whether to become urban entrepreneurs or whether to stay put as rural workers.

Using (6) and as shown in Appendix T.5, the expected value of (9) is given by:

$$\mathbb{E}(\Delta V_l) = A\frac{H_l c_l^{k+2} + \tau^{-k}\sum_{h\neq l} H_h c_h^{k+2}}{k+2} + \frac{\alpha - c_l}{2\eta}\left[\alpha - \frac{k+1}{k+2}c_l\right] - \theta H_l - f^E. \tag{10}$$

Expectations are rational and, at equilibrium, perfect. Agents are negligible and hence rationally disregard the impact of their actions on equilibrium market aggregates; they also take all other agents' decisions as given. The identities (8) and the inequalities $\mathbb{E}(\Delta V_l) \leq 0$ in (10) constitute a system of 2Λ conditions in the 2Λ unknowns $\{H_l\}_{l=1}^{\Lambda}$ (city sizes) and $\{c_l\}_{l=1}^{\Lambda}$ (cost cutoffs).

4 Equilibrium with one region: 'Urbanization'

To set the stage, we start by analyzing the equilibrium with a single region. In so doing, we can identify the three-way relationships among agglomeration, polarization and income inequality in a parsimonious way. As we shall see, two types of equilibria may arise in this simple case: an equilibrium in which no city forms, and an equilibrium in which a city forms.

4.1 Urban and rural equilibria

To ease notation, we suppress the h and l subscripts for the time being, except for the cutoff c_l (which may otherwise be mixed up with the firms' individual inverse productivity c). Using (10) the free entry condition then reduces to

$$\frac{A}{k+2} H c_l^{k+2} + \frac{\alpha - c_l}{2\eta} \left[\alpha - \frac{k+1}{k+2} c_l \right] - \theta H - f^E \leq 0, \tag{11}$$

with equality if $H > 0$ and strict inequality if $H = 0$. The first two terms in (11) collect the expected profits and the consumer surplus, respectively, whereas the last two terms collect the urban and the entry costs. Turning to condition (8), it can be solved for H as follows:

$$H = \frac{\alpha - c_l}{A\eta c_l^{k+1}}. \tag{12}$$

Two aspects of (12) are noteworthy. First, at any equilibrium with a strictly positive city size ($H^* > 0$), the equilibrium cutoff is strictly smaller than α. Second, $\partial H / \partial c_l < 0$ and $\partial^2 H / \partial c_l^2 > 0$ at equilibrium, thus revealing that there is a positive (and convex) equilibrium relationship between *agglomeration* and *selection*. In plain English, only the fittest entrepreneurs survive and produce, and this effect is particularly strong in large cities. Substituting (12) into (11), and rearranging, we obtain:

$$\frac{\alpha - c_l}{2\eta} \left[\alpha - \frac{k-1}{k+2} c_l - \frac{2\theta}{A c_l^{k+1}} \right] - f^E \equiv f(c_l; \mathbf{Z}) \leq 0, \tag{13}$$

where $\mathbf{Z} = \{\alpha, \eta, \gamma, f^E, c_{\max}, \theta\}$ is the vector of parameters in the model. The equilibrium condition (13) is central to the analysis that follows. It is expressed only as a function of c_l and of the model's parameters: conveniently, c_l is thus a summary statistic for expected profits, consumer surplus and congestion at once. Also, the nature and number of equilibria are fully characterized

by the properties of f. An interior equilibrium with a city ($H^* > 0$ and $0 < c_l^* < \alpha$), which we henceforth refer to as an *urban equilibrium*, is such that $f(c_l^*) = 0$; whereas an equilibrium without a city ($H^* = 0$ and $c_l^* = \alpha$), which we henceforth refer to as a *rural equilibrium*, necessarily implies $f(\alpha) \leq 0$. A rural equilibrium is always stable whenever it exists, whereas an urban equilibrium is *locally stable* if and only if $\partial f(c_l^*)/\partial c_l > 0$. This latter condition implies that, at a locally stable equilibrium, any small perturbation of city size is such that the free-entry condition will bring the economy back to its initial situation.[10]

It is readily verified that $\lim_{c_l \to 0} f(c_l) = -\infty$, which shows quite naturally that there is always an upper limit to city size. Furthermore, whenever a rural equilibrium does not exist there exists, by continuity, at least one stable urban equilibrium. A by-product of this latter property is that the smallest root of f (whenever one exists), which corresponds to the largest equilibrium city size, is a stable equilibrium as in Henderson (1974). We can summarize those findings as follows:

Proposition 1 (existence and number of equilibria) *The function f has either one or three positive roots, of which at most two are in $[0, \alpha)$. Consequently, there exist at most two stable equilibria: an urban equilibrium and the rural equilibrium. If no stable urban equilibrium exists, then the rural equilibrium is unique. Furthermore, the equilibrium associated with the smallest value of c_l (the largest H) is always stable.*

Proof. See Appendix B.1. ■

Given the equilibrium structure, how do the equilibria change with the values of the underlying parameters? We show that lower commuting costs (lower θ), a stronger preference for the differentiated good (larger α), a better productivity support (lower $c_{\max}$ and thus higher A) and stronger product differentiation (lower γ and thus higher A), all weakly increase city size and city productivity at any stable equilibrium. Formally:

Proposition 2 (urban equilibrium: monotonicity) *At any stable equilibrium, the equilibrium productivity cutoff $1/c_l^*$ and the equilibrium city size H^* are both non-increasing in θ and f^E and non-decreasing in α and A.*

Proof. See Appendix B.2. ■

We next investigate when which type of equilibrium arises. The following lemma, which pertains to the special case where $f^E = 0$, is useful before we proceed:

Lemma 3 *Assume that there are no net entry costs for becoming an entrepreneur ($f^E = 0$). Then: (i) f is increasing, negative and concave in the neighborhood of $c_l = 0$; (ii) $c_l = \alpha$ is always a root of f; and (iii) f admits at most one root on $(0, \alpha)$.*

Proof. See Appendix B.3. ■

[10] This is the case if, following standard 'new economic geography' practice, we specify the following law of motion for H: $\dot{H} = \mathbb{E}(\Delta V)H(L - H)$. This replicator dynamics can be microfounded (see Baldwin, 2001).

A consequence of the foregoing lemma is that the model admits a unique stable equilibrium if $f^E = 0$. If f^E and θ are large enough, then the rural equilibrium exists and is stable, as we establish formally in the following proposition:

Proposition 4 (rural equilibrium: stability) *(i) The rural equilibrium ($H^* = 0$ and $c_l^* = \alpha$) exists and is stable for any $f^E > 0$. (ii) If $\theta \geq \theta^R$ and $f^E \geq f^R$, where*

$$\theta^R \equiv \frac{3A\alpha^{k+2}}{2(k+2)} \quad \text{and} \quad f^R \equiv \frac{\alpha^2}{2\eta}\frac{k-1}{k+2}, \tag{14}$$

then the rural equilibrium is the unique equilibrium. (iii) If $f^E = 0$ then the rural equilibrium exists and is a stable equilibrium if and only if $\theta \geq \theta^R$.

Proof. See Appendix B.4. ■

Note that the sufficient conditions $f^E \geq f^R$ and $\theta \geq \theta^R$ for the rural equilibrium to be the unique stable equilibrium are less likely to hold if the technology used to produce the urban good is efficient and if consumers value it a lot (i.e., if A and α are high). The equilibrium structure of the model is depicted in Figure 3, where bold lines denote stable and where dashed lines denote unstable equilibria. As one can see, when $f^E = 0$ the rural equilibrium is stable for sufficiently high commuting costs and becomes unstable otherwise. When $f^E > 0$, the rural equilibrium is always stable; it is even the unique equilibrium when commuting and entry costs are together prohibitive, which is the case if $f^E \geq f^R$ and $\theta \geq \theta^R$. However, for sufficiently low values of commuting and entry costs, two urban equilibria appear, the one associated with the largest city size being stable and the other one being unstable.

Insert Figure 3 about here.

More formally, the equilibrium structure is characterized as follows:

Proposition 5 (equilibrium structure) *(i) Let $f^E \geq f^R$ and $\theta \geq \theta^R$; then the rural equilibrium $H^* = 0$ is the unique equilibrium. (ii) Let $f^E = 0$ and $\theta > \theta^R$; then $H^* = 0$ is the unique stable equilibrium. (iii) Let $f^E = 0$ and $\theta \in (0, \theta^R)$; then there exists a unique pair $\{H^*, c_l^*\}$ in $\mathbb{R}_{++} \times (0, \alpha)$ that consitutes a stable equilibrium (the urban equilibrium). (iv) Let $f^E > 0$; then there exists a θ, denoted as $\theta^U(f^E)$ with $\theta^U(f^E) < \theta^R$ and $\lim_{f^E \to 0} \theta^U(f^E) = \theta^R$, such that there is at most one pair $\{H^*, c_l^*\}$ in $\mathbb{R}_{++} \times (0, \alpha)$ that consitutes a stable equilibrium if $\theta \leq \theta^U(f^E)$.*

Proof. See Appendix B.5. ■

Parts (i) and (ii) in Proposition 5 establish conditions for the rural equilibrium to be the unique one, whereas part (iii) does the same for the urban equilibrium. Parts (iv) and (v) together establish the conditions for both the rural and urban equilibria to exist and to be stable.

We can now summarize the properties of the model, as collected in Propositions 2, 4 and 5 by focusing our attention on the *urbanization threshold* $\theta^U(f^E)$ (or θ^U henceforth for short): no city can emerge for values of θ larger than θ^U. Conversely, for all θ smaller than θ^U, urbanization *may* occur and both an urban and a rural equilibrium can be sustained. As stated in the foregoing, the largest city size is always stable when $\theta < \theta^U$, while a smaller yet unstable equilibrium city size coexists. Inspection of the urbanization threshold θ^U reveals several equilibrium characteristics worth stressing. First, any improvement in the benefits of living in cities, either as consumers or entrepreneurs, makes the emergence of cities more likely and maps into larger equilibrium city sizes and higher city productivity. By Proposition 4, the rural equilibrium exists and is stable provided that either: entrepreneurs draw their productivities from a bad support, i.e., $c_{\max}$ is large; acquiring entrepreneurial skills is expensive, i.e., f^E is large; products are sufficiently homogenous, i.e., γ is large; preferences for the differentiated good α are weak; and urban costs θ are large. Conversely, declining urban costs θ and rising benefits of living in cities A and α are both conducive to rural-urban migration, thereby ensuring that urbanization does arise (Proposition 5) or that cities grow (Proposition 2). These findings are consistent with the three 'classical' conditions stressed by, e.g., Bairoch (1988), for cities to emerge and to develop. First, there must be an agricultural surplus so that the rural population may feed the urban dwellers (in our model, this condition is satisfied via the initial endowment in the numéraire $\overline{d}^0$). Conversely, there must be some demand for urban goods and services: the extent of this demand is captured by the parameter α and urban production is more valuable if products are more differentiated (low γ). Second, the urban population must supply goods and services to sustain itself. It is able to produce more the lower is $c_{\max}$. Last, any reduction in urban costs that stems from an improvement in urban transportation is conducive to urban growth (Duranton and Turner, 2008). To sum up, a large α and a low γ, θ, f^E or $c_{\max}$ are all conducive to the emergence of (large) cities.

4.2 Polarization and income inequality

From expression (12) we know that, at equilibrium, larger cities are more productive. How does this map into average income? This question is warranted since only $G(c_l)H$ entrepreneurs produce at equilibrium, whereas the remaining ones exit the market and consume from their endowments. The *failure rate* $1 - G(c_l)$ in the urban market thus influences the distribution of income across successful and unsuccessful entrepreneurs.

Our model allows us to take a theoretical perspective on this question. To do so, we first compute the average (operating) profit of all entrants, including those who end up failing at equilibrium, which is given by

$$\overline{\Pi}(H, c_l) = A\frac{Hc_l^{k+2}}{k+2}. \tag{15}$$

Making use of the equilibrium relationship (12) between size and productivity, we then obtain $\partial\overline{\Pi}/\partial c_l = (\alpha - 2c_l)/[\eta(2+k)]$. Hence, $\overline{\Pi}$ is $\cap$-shaped and concave in c_l: average profits first increase in c_l for $0 \leq c_l \leq \alpha/2$ and then decrease for $\alpha/2 \leq c_l \leq \alpha$. This relationship is non-monotonic because operating in a large city has both pros and cons: a large market size increases profits (the 'H' component in the expression for $\overline{\Pi}$), but also induces tougher competition, thus reducing markups and profits (the 'c_l' component in the expression for $\overline{\Pi}$). Qualitatively, these findings are in line with our **Stylised Fact 1**: average income rises with city size (**Stylised Fact 1a**) but *at a decreasing rate* (**Stylised Fact 1b**).[11]

As an alternative measure of income, we then compute the average productivity conditional on producing, denoted by $\widetilde{\Pi} \equiv \overline{\Pi}\,\big|_{c\leq c_l}$, and we find that it is monotonically increasing in city size. To summarize our key findings thus far:

Proposition 6 (agglomeration) *(i) The average profit is non-decreasing and concave in city size if cities are smaller than* $1/\left[A\eta\left(\alpha/2\right)^k\right]$. *(ii) The average profit, conditional on survival, is monotonically increasing in city size.*

Proof. See Appendix B.6. ■

A central focus of the paper is on how city size affects heterogeneous individuals differently. To address this issue more formally, we characterize the average profit of the top quantile Q of the distribution and define $\widetilde{q}(Q)$ as $G(\widetilde{q}) = Q$ (for instance, the least productive entrepreneur of the top 20% has a c such that $G(c) = .2$). Letting $q \equiv \min\{\widetilde{q}, c_l\}$, we may then write the average profit of the top quantile Q as follows:

$$\overline{\Pi}_q(H, c_l) \equiv \frac{1}{G(q)}\int_0^q \Pi(c)\mathrm{d}G(c) = k\frac{H}{4\gamma}\left[\frac{c_l^2}{k} - \frac{2c_l q}{k+1} + \frac{q^2}{k+2}\right]. \tag{16}$$

Two features of (16) are noteworthy. First, evaluating this expression for $Q = 1$ naturally gives the average profit for the entire distribution; to see this, note that $\widetilde{q} = c_{\max}$ implies $\overline{\Pi}_{c_{\max}} = \overline{\Pi}$. Second, the ambiguity of city size on profits is again apparent in (16), for the term in the square bracket is decreasing in city size (increasing in c_l). Despite this ambiguity, the share of income accruing to the top of the skill distribution is unambiguously increasing in city size:

Proposition 7 (polarization) *Let* $\sigma_q \equiv \overline{\Pi}_q/\overline{\Pi}$ *denote the average income of the top* Q% *of the distribution relative to the overall average. Then: (i)* $\partial\sigma_q/\partial H > 0$*; and (ii)* $\partial^2\sigma_q/\partial H\partial q < 0$ *if and only if* $q < \overline{c}_l = c_l k/(k+1)$.

Proof. See Appendix B.7. ■

[11] When reinterpreted in light of our model, the evidence presented in Section 2 suggests that all US cities in our sample are not too large, i.e., $0 < H \leq 1/\left[A\eta\left(\alpha/2\right)^k\right]$ and $\alpha/2 \leq c_l < \alpha$ so that we are on the increasing part of the relationship. Au and Henderson (2006) show that there is a $\cap$-shaped relationship between city size and average real income (productivity) in Chinese cities. Our results suggest that, beyond some threshold, the relationship may also be $\cap$-shaped between city size and average nominal income.

Part (i) of Proposition 7 is the theoretical counterpart of **Stylised Fact 2a**, whereby the average income of the top 5% relative to the overall average increases with city size. This 'superstar' effect is also consistent with **Stylised Fact 2b**, whereby the elasticity of average income with respect to city size is increasing as we move up the income distribution. Part (ii) suggests that this expansion of the share of income accruing to the wealthiest comes at the expense of both the bottom half of the population of successful entrepreneurs, as well as those who simply fail. This polarization effect naturally leads us to study the relationship between city size and income inequality as highlighted by our **Stylised Fact 3**. To this aim, we compute the Gini coefficient of the income distribution as follows (see Appendix B.8 for details):

$$\text{Gini}(k, c_l) = 1 - \frac{k+2}{4k+2}\left(\frac{c_l}{c_{\max}}\right)^k. \tag{17}$$

Note that this coefficient does not depend directly on city size H because the Gini coefficient is 'scale free'. The Gini is also increasing in the shape parameter k, which governs the extent to which abilities are unevenly distributed. In particular, for a given c_l, the fraction of successful entrepreneurs falls as k rises. Straightforward inspection of (17) yields the following results:

Proposition 8 (inequality) *Let income inequality be measured by the Gini coefficient. Then income inequality is: (i) increasing at an increasing rate in the productivity cutoff $1/c_l$; (ii) increasing at a decreasing rate in city size H; (iii) decreasing in k; and (iv) increasing in $c_{\max}$.*

Proof. (i) It can be verified that $\partial(\text{Gini})/\partial c_l < 0$ and $\partial^2(\text{Gini})/\partial c_l^2 < 0$. (ii) $\partial(\text{Gini})/\partial H > 0$ readily follows from the monotonicity of (12) and (17). To obtain the concavity of Gini with respect to H, invert (17) to get an expression for c_l as a function of Gini, and substitute this for c_l into (12). Then, standard algebra reveals that $\partial^2 H/\partial(\text{Gini})^2 > 0$ and thus $\partial^2(\text{Gini})/\partial H^2 < 0$. (iii) Using (17) again, we obtain:

$$\frac{\partial(\text{Gini})}{\partial k}(k, c_l) = [1 - \text{Gini}(k, c_l)]\left[-\frac{3}{(k+2)(2k+1)} + \ln\left(\frac{c_l}{c_{\max}}\right)^k\right],$$

which is negative by inspection (recall that $c_l < c_{\max}$). (iv) The last part of the proposition immediately follows by inspection of (17). ■

To summarize the findings of this section, *large urban areas generate more wealth and are at the same time more unequal and more polarized than smaller cities.* As shown in Section 2, these theoretical predictions of our model are robust features of the US data.

4.3 Urban poverty and consumer cities

The model can also shed some light on a couple of facts that have attracted attention in the literature. First, it is worth stressing that, after entry, unsuccesful entrepreneurs in the city have

lower nominal and real incomes than those of workers in the countryside, yet that their consumer surplus exceeds that in the countryside. The reason is that they have access to urban diversity even if their choice to move to the city turns out to be unsuccessful (recall that unsuccessful agents can still pay for urban goods using their initial endowment). This aspect is taken into account in the entry decision in our model. It is somewhat reminiscent of standard arguments for explaining the growth of cities in the Third World, where the massive urbanization in the face of urban poverty constitutes a classical puzzle (Harris and Todaro, 1970).

Second, in a cross-section of (isolated) cities, those that have unfavorable fundamentals are small and not very productive ($\{H^*, c_l^*\} \simeq \{0, \alpha\}$ when $\theta \simeq \theta^U$). In such economies, urban migration is primarily motivated by urban wages (entrepreneurs' profits in the model) that are large relative to rural wages. Furthermore, the 'failure rate' $1 - G(c_l) \simeq 1 - G(\alpha)$ is relatively low, i.e., the mass of unsuccessful entrepreneurs is small. However, the consumer surplus is rather small too in this case: $\mathrm{CS}(c_l) \simeq \mathrm{CS}(\alpha) = 0$ by (6). Cities with good underlying economic conditions are large, competitive and productive; as a result, expected profits are then no longer the primary driver of urban life (the failure rate is large and expected profits are low), but the city's local and specific service and product mixes work like local amenities that attract consumers who display preference for diversity. At the limit, when $c_l \to 0$, expected profits go to zero and the consumer surplus $\mathrm{CS}(0)$ reaches its maximum and compensates for urban costs on its own. We may also view the recent history of urbanization as an ongoing reduction in commuting costs θ (e.g., the invention of the streetcar in the late nineteenth century and the spread of the automobile in the twentieth). In light of this historical perspective, cities in the early ages of the Industrial Revolution were *producer cities* and not very pleasant places to live in, whereas large modern cities are predominantly *consumer cities* that offer a wide array of consumer goods and services.[12]

5 Extensions: Urban systems and trading cities

We now extend the model to include multiple cities and selectively report a series of theoretical results with two aims in mind. Firstly, we establish that the equilibrium relationship between city size and average productivity continues to hold true in a multi-city framework. More precisely, what was in the previous section essentially a comparative statics result pertaining to isolated cities becomes now an equilibrium relationship in a framework where cities are linked through the channels of trade. Secondly, an open question in the literature asks whether selection reinforces or weakens agglomeration forces, how those two forces interact, and how they can be disentangled empirically (Combes, Duranton, Gobillon, Puga and Roux, 2009). The multi-city extension of

[12]The terminology 'producer and consumer cities' was introduced by Weber (1958), though his concept of 'consumer city' captures more closely the predatory behavior of primate cities rather than the more modern concept of consumer city in the wake of Glaeser, Kolko and Saiz (2001).

our model enables us to address this question, and we show that *selection is both an agglomeration and a dispersion force*. In other words, the net impact of selection on agglomeration is a priori unclear.

5.1 Urban systems

Let $\Lambda \geq 2$ and define as an *urban system* a stable equilibrium configuration in which cities of possibly different sizes co-exist or in which some regions develop cities whereas others do not. To begin with, let us go back to the short-run equilibrium condition (8), while temporarily disregarding the long-run condition (10). For any region l, we may rewrite condition (8) as:

$$\frac{\alpha - c_l}{A\eta c_l^{k+1}} = H_l + \phi \sum_{h \neq l} H_h, \tag{18}$$

where $\phi \equiv \tau^{-k}$ is a measure of trade openness with values in the unit interval.[13] The right-hand side of (18) can be interpreted as a measure of the 'market potential' of city l (Head and Mayer, 2004). Rewriting the system in matrix form yields:

$$\underbrace{\begin{bmatrix} 1 & \phi & ... & \phi \\ \phi & 1 & ... & \phi \\ ... & & & ... \\ \phi & ... & \phi & 1 \end{bmatrix}}_{\mathbf{F}} \underbrace{\begin{bmatrix} H_1 \\ H_2 \\ ... \\ H_\Lambda \end{bmatrix}}_{\mathbf{h}} = (A\eta)^{-1} \underbrace{\begin{bmatrix} (\alpha - c_1)c_1^{-(k+1)} \\ (\alpha - c_2)c_2^{-(k+1)} \\ ... \\ (\alpha - c_\Lambda)c_\Lambda^{-(k+1)} \end{bmatrix}}_{\mathbf{x}} \tag{19}$$

where $\mathbf{F}$ is a Λ-dimensional invertible square matrix whose determinant is positive by inspection (all its off-diagonal elements are identical and smaller than its diagonal elements) and $\mathbf{h}$ and $\mathbf{x}$ are both Λ-dimensional vectors. We use (19) to show that the qualitative relationship established in (12) as a comparative statics result carries over to an equilibrium relationship in an urban system. Formally:

Proposition 9 (size and selection in an urban system) *Assume that regions are ex ante (or fundamentally) symmetric, i.e., they face the same bilateral trade barriers and have identical ability supports. Then, at any equilibrium, selection is tougher in larger cities:*

$$c_l \leq c_h \quad \Longleftrightarrow \quad H_l \geq H_h.$$

Furthermore, $\partial H_l / \partial c_l < 0$ *and* $\partial H_l / \partial c_h > 0$.

Proof. See Appendix C.1. ■

[13] Note that a distribution which is more skewed towards lower ability draws (i.e., a higher value of k) implies a lower ϕ for any given τ, as fewer entrepreneurs are productive enough to export to other cities.

In words, Proposition 9 establishes two important results. First, selection is tougher and, as a result, average productivity is higher (c_l is lower) in larger cities. Insofar as mean income and average productivity are positively related at the city level, the finding of Proposition 9 is consistent with **Stylised Fact 1**. As a corollary, the positive relationship between the number of available varieties and the toughness of selection in (18) implies a hierarchy of cities akin to the *Central Place Theory* of Lösch (1940). Second, own city size decreases with own cutoff (selection) and increases with foreign cutoffs (competition). Insofar as a large H_l is the flip side of a low c_l, this finding suggests that urbanization in l may hinder urbanization in h and vice versa. We refer to this negative dependence as the *cannibalization effect* of proximate cities (see Dobkins and Ioannides, 2000; Partridge, Rickman, Ali and Olfert, 2009).

To push the analysis further, we must impose the long-run condition (10) to study the properties of asymmetric equilibria, including 'core-periphery equilibria', i.e., those in which only a subset of regions develops cities at equilibrium. Such an analysis is more involved because cities either inhibit the emergence or favor the existence of other cities in complex ways (Fujita, Krugman and Venables, 1999; Tabuchi, Thisse and Zeng, 2005).[14] By contrast, the analysis of symmetric equilibria is much simpler, yet enables us to derive a handful of interesting insights. We thus turn to this issue next.

5.2 Trading cities

We now look for the existence of *symmetric equilibria* with multiple regions and characterize their properties. The analysis is similar to the case where $\Lambda = 1$, except for the existence of trading links. The proofs of Propositions 1 to 5 in Appendix B are provided to include the current setting, which encompasses $\Lambda = 1$ as a special case. Put differently, Propositions 1, 2 and 4 carry over to the more general setting of this section. We can thus exclusively focus on the impact of changes in trade costs on the existence and the properties of symmetric equilibria.

Let $\Phi \equiv (\Lambda - 1)\phi$, which is increasing in Λ and in ϕ and which takes value $\Phi = 0$ when $\phi \to 0$ (trade is prohibitive), or when $\Lambda = 1$ (there is a single isolated region). Since the model is perfectly symmetric by assumption, an equilibrium where all regions have the same size $H_l \equiv H$ and the same cutoff c_l always exists. Imposing symmetry in the short-run equilibrium condition (8) yields

$$\frac{\alpha - c_l}{(1+\Phi)A\eta c_l^{k+1}} = H. \tag{20}$$

Plugging this expression into the free-entry condition (10) and imposing symmetry allows us to

[14] Due to space constraints, we do not report those developments here. See Behrens and Robert-Nicoud (2008), for details and examples of asymmetric equilibria. We also provide more general stability conditions there.

rewrite (13) as follows:

$$\frac{\alpha - c_l}{2\eta}\left[\alpha - \frac{k-1}{k+2}c_l - \frac{2\theta}{(1+\Phi)Ac_l^{k+1}}\right] - f^E \equiv f(c_l) \leq 0. \tag{21}$$

Rural and urban equilibria are defined as in the single-city case of Section 3. Note that A and Φ always enter the equilibrium expressions together as $A(1+\Phi)$. Therefore, the whole analysis pertaining to the role of A in relation to the types and stability of equilibria and the comparative statics of the previous section readily extend to the role of Φ. In words, *the implications of an increase in the freeness of trade are isomorphic to those of an increase in the underlying productivity of the whole economy.* Concretely, we show that lower trade costs are conducive to city formation and city growth:

Proposition 10 (cities and trade) *A larger Φ (lower trade costs τ and/or more trading partners Λ), a larger A or a lower θ all make the existence of cities more likely and weakly increase their equilibrium size.*

Proof. We first show that a smaller value of Φ makes the rural equilibrium more likely to occur. Note that when $f^E = 0$, local stability of the rural equilibrium requires that $\partial f/\partial c_l > 0$ when evaluated at $\{H^*, c_l^*\} = \{0, \alpha\}$. This is equivalent to $\theta > \theta_\Phi^R$, where θ_Φ^R is given by

$$\theta_\Phi^R \equiv (1+\Phi)3A\frac{\alpha^{k+2}}{2(k+2)} = (1+\Phi)\theta^R.$$

As in the single-city case, the rural equilibrium exists and is stable for all $\theta \geq \theta_\Phi^R$, whereas the urban equilibrium is the unique stable equilibrium when $\theta < \theta_\Phi^R$. Clearly, θ_Φ^R is increasing in Φ (i.e., with freer trade), which proves our claim. Turn next to the case where $f^E > 0$. We have already established in the proof of Proposition 5 that f is continuously decreasing in both f^E and θ, which implies that the equilibrium city size is decreasing in f^E and increasing in Φ at the stable urban equilibrium. ■

As established before in Propositions 4 and 5, the *intra-city transportation system* must be efficient enough for cities to emerge in equilibrium. The new result in Proposition 10 is that cities are also more likely to emerge if the *inter-city transportation system* is efficient enough so that cities can trade with one another at a low cost. Note that this result may not be as obvious as it sounds. Indeed, from the perspective of entrepreneurs in each city, lower inter-city trade costs and a larger number of trading partners mean both a better market access and tougher competition from entrepreneurs established in other cities. To see the latter effect, evaluate (8) at the symmetric equilibrium to get $(\alpha - c_l)/[A\eta c_l^{k+1}] = H(1+\Phi)$ and observe that $\partial H/\partial\Phi|_{c_l} < 0$. As it turns out, Proposition 10 shows that in equilibrium the agglomeration effect dominates, i.e., $\partial H/\partial\Phi > 0$; the selection effect is also stronger, the lower the trade costs τ are and/or the more numerous the trading cities Λ are, i.e., $\partial c_l/\partial\Phi < 0$. Trade thus increases aggregate productivity and city sizes simultaneously.

5.3 Trade, profits and polarization

Does trade contribute to rising profits and larger income inequalities in cities? To answer this question, we first compute the average (operating) profit of the entrants in the symmetric case as follows:

$$\overline{\Pi}_{\Phi} = \frac{(1+\Phi)A}{k+2} H c_l^{k+2} = \frac{(\alpha - c_l)c_l}{\eta(k+2)} = (1+\Phi)\overline{\Pi},$$

where we have used the short-run equilibrium condition $(\alpha - c_l)/[A\eta c_l^{k+1}] = H(1+\Phi)$. Since $\partial c_l/\partial\tau > 0$, it is readily verified that the average profit in city l is $\cap$-shaped in τ. Hence, as in the single-city case with respect to size, average profits are first increasing as trade costs fall from high initial values, and then eventually decreasing as trade becomes sufficiently free. In the early stages of integration, access to a larger market raises entrepreneurs' profits, whereas in later stages of integration increased competition reduces them again as more agents fail due to tougher selection. The effect of Φ on the expected profit conditional on being successful, $\widetilde{\Pi}_{\Phi} = (1+\Phi)\widetilde{\Pi}$, is also qualitatively identical to the effect of A in the single-city case.

The effect of Φ on polarization is quite intuitive. Insofar as freer trade makes all markets more competitive, this hurts profits of *every* producer. However, higher trade openness also opens foreign markets to *some* entrepreneurs, the exporters. As a result, a higher Φ unambiguously raises the exporters' share of profits in the industry. Since only the most productive entrepreneurs export, by shifting profits from non-exporters to exporters, it follows logically that more trade openness increases income inequality. To establish this formally, we compute the Gini coefficient as follows:

$$\text{Gini}_l(\Lambda, \tau, k; c_l) = 1 - \lambda(\Lambda, \tau, k)\left(\frac{c_l}{c_{\max}}\right)^k, \tag{22}$$

where $\lambda(\cdot)$ is a bundle of parameters too unwieldy to be revealing (its expression is relegated to Appendix C.2). One can verify that (22) is equivalent to (17) in two special cases: first and naturally, when there is only one city in the economy ($\Lambda = 1$); second, when inter-city trade is perfectly free ($\tau = 1$).

Insert Figure 4 about here.

Figure 4 plots the equilibrium Gini coefficient as a function of τ for $\Lambda = 4$ and $\Lambda = 8$, respectively, by using the equilibrium value of (21) in (22). All the simulations we have conducted show that inequality rises as the number of trading partners increases, or $\partial(\text{Gini}_l)/\partial\Lambda > 0$. Turning to trade costs τ, we analytically show that income inequality increases at the city-level as trade becomes less costly:

Proposition 11 (trade and income inequality) *Let income inequality be measured by the Gini coefficient. Then income inequality is increasing at the symmetric equilibrium as trade gets freer, namely* $\partial(\text{Gini}_l)/\partial\tau < 0$.

Proof. By inspection of (22), Gini_l is decreasing in both c_l and $\lambda(\cdot)$. Thus, to establish the result, it is sufficient to show that $\partial\lambda/\partial\tau < 0$, which we do in Appendix C.2. ■

Proposition 11 establishes that trade integration or lower costs of shipping goods do increase income inequality, at least in the symmetric case. The reason is actually two-fold: freer trade makes selection tougher by expanding the market, which raises the failure rate (i.e. $(c_l/c_{\max})^k$ falls). As shown in Appendix C.2, a lower τ also redistributes income from the least productive entrepreneurs to the most successful ones, thus lowering λ. Consequently, the income distribution gets more skewed towards the most productive agents, who secure a larger share of total income. The recent trends of economy-wide integration are hence likely to spur more income inequality at the city level, and the redistribution of income may not predominantly be a cross-factor cross-sector issue, but may well take place across skills within the same sectors.

5.4 Linkages, competition, and selection

We conclude our theoretical analysis by studying the five key ingredients that shape the spatial outcome of our model: *selection, competition, agglomeration ('backward' and 'forward' linkages) and urban costs*. One concise way to illustrate how these work is to study the stability properties of the symmetric equilibrium. We already know that the symmetric equilibrium is (locally) stable if $\theta < \theta^R$ in the sense that an arbitrarily small shock (to the endogenous variables) common to *all* cities is self-correcting in this case, i.e., $\mathrm{d}f/\mathrm{d}c_l > 0$ at the symmetric equilibrium.

When there is more than one city, however, the (symmetric) equilibrium might be unstable in other ways. In what follows, we retain a simple definition of stability: we focus solely on the case without aggregate shocks, i.e. $\sum_l \mathrm{d}H_l = 0$, and where the shock affects *just two cities* at the symmetric equilibrium.[15] When is such a shock to city size self-correcting? In other words, starting from a symmetric equilibrium configuration, imagine that H_l increases by $\mathrm{d}H > 0$ and H_h decreases by the same $\mathrm{d}H > 0$ (and $\mathrm{d}H_i = 0$ for all $i \neq l, h$). Loosely speaking, this corresponds to an entrepreneur entering the 'wrong' city. If $\mathbb{E}[\Delta V_l(\mathrm{d}H)] = -\mathbb{E}[\Delta V_h(\mathrm{d}H)]$ is negative, then the shock is self-correcting and the symmetric equilibrium is indeed locally stable (the symmetry of the model implies $\mathbb{E}[\Delta V_i(\mathrm{d}H)] = 0$ for all $i \neq l, h$). Otherwise, symmetry is 'broken' and the symmetric equilibrium is locally unstable (Krugman, 1991; Baldwin, 2001).

To address this issue formally, let $H_l = H_h = H$, $\mathrm{d}H = \mathrm{d}H_l = -\mathrm{d}H_h > 0$ and $\mathrm{d}c_h = -\mathrm{d}c_l$.

[15] The absence of aggregate shocks is standard in economic geography models where the mass of firms and of mobile agents is usually fixed (Krugman, 1991; Ottaviano, Tabuchi and Thisse, 2002). In general, local stability requires that at the interior equilibrium, the Jacobian of f with respect to $\mathbf{c} = (c_1\ c_2\ \ldots\ c_\Lambda)$ be positive definite. As characterizing the eigenvalues of a non-numerical system is a daunting task, which leads to a complex taxonomy of different cases (see Tabuchi, Thisse and Zeng, 2005, for further details), we retain a less stringent condition.

Differentiating (8) and (10) around the symmetric equilibrium, and using (20), yields

$$\begin{aligned} \mathrm{d}\mathbb{E}(\Delta V_l) &= \underbrace{-\theta \mathrm{d}H}_{\text{congestion}} + \underbrace{\frac{1-\phi}{1+\Phi}\frac{\alpha-c_l}{\eta}\frac{c_l}{k+2}\frac{\mathrm{d}H}{H}}_{\text{backward linkage}} \\ &+ \underbrace{\frac{1-\phi}{1+\Phi}\frac{\alpha-c_l}{\eta}\mathrm{d}c_l}_{\text{selection and competition}} \underbrace{-\frac{1}{\eta(k+2)}\left[\frac{\alpha}{2}+(\alpha-c_l)(k+1)\right]\mathrm{d}c_l}_{\text{forward linkage}} \end{aligned} \tag{23}$$

and

$$(1+\Phi)\frac{\alpha+k(\alpha-c_l)}{\alpha-c_l}\frac{\mathrm{d}c_l}{c_l}+(1-\phi)\frac{\mathrm{d}H}{H}\equiv 0. \tag{24}$$

A few aspects of (23) are noteworthy. First, the expected value of becoming an entrepreneur in l is affected by both the mass of entrepreneurs H and by their average equilibrium inverse ability (which is proportional to c_l under the Pareto parametrization). Second, there is a 'pull' and a 'push' factor in both. To see this, consider the first line of the right-hand side of (23). An additional entrepreneur in l means one more urban dweller, which increases urban costs by θ (*congestion*); it also means one more consumer in l (and one fewer in h), which means a relatively larger market in l and thus higher profits. This demand linkage is also known as a *backward linkage* (Fujita, Krugman and Venables, 1999). The latter effect relies on market segmentation, hence it is increasing in τ (i.e., decreasing in ϕ and Φ); at the limit, when goods markets are fully integrated ($\phi = 1$), this effect vanishes as local market size becomes irrelevant. Consider next the second line of the right-hand side of (23). Less productive entrepreneurs (i.e., a larger value of c_l) is good news for expected profits: the failure rate is lower (less *selection*) and the pro-competitive effect is weaker (*competition* is softer). This effect is also directly affected by market segmentation: when $\phi = 1$, competition becomes *global* and thus shifting entrepreneurs around has no impact on *local* expected profits. By contrast, less productive local entrepreneurs is bad news for consumers because they pay higher prices for their consumption bundle. The reason is that, since all varieties are substitutes, less productive local entrepreneurs raise the prices of all varieties sold in l. Also, consumers substitute towards non-local varieties as c_l rises and pay trade costs on these imports. This cost linkage, also known as *forward linkage* (Fujita, Krugman and Venables, 1999), is indirectly affected by the degree of market segmentation. Indeed, as can be seen from the equilibrium constraint (24), the (negative) link between c_l and H weakens as ϕ decreases: swapping entrepreneurs between l and h has no effect on consumer surplus nor on competition when the market is global.

To summarize our key findings, as can be seen from the second line of (23), *selection is both an agglomeration and a dispersion force*. Whereas tougher selection decreases operating profits (the first term) and thus works against agglomeration on the production side, it also increases purchasing power through lower consumer prices (the second term) and thus favors agglomeration of the consumption side. Disentangling the relative importance of these different production and

consumption channels for agglomeration then becomes an empirical question that we keep for future research.

6 Conclusions

All empirical studies reveal that the elasticity of worker and firm productivity with respect to city size is positive and typically falls in the 3% – 8% range. Less well known is the stylized fact that larger cities are also more polarized: the average incomes of the highest income quantiles are magnified by city size, so that income inequality is increasing with urban size, at least in the US. The elasticity of the income Gini coefficient with respect to city size is also quite substantial, in the order of 1% for the metro areas to 3% for the micro areas. This paper has developped an integrated model to account for such facts simultaneously. In particular, it can replicate the stylized facts that *larger urban areas are more productive and have larger income inequalities* than smaller ones, and that the latter is a consequence of polarization due to both selection effects ('poverty') and the dilatation of the upper end of the income distribution ('superstars'). Our model is also flexible enough to shed light on phenomena such as urbanization, to investigate the impacts of trade integration on city size and inequality, and to highlight how selection works both as an agglomeration and as a dispersion force.

Our findings open the research agenda in two directions at least. On the empirical front, as we have said, we view the macro evidence provided in the paper as suggestive only. Indeed, we do not *identify* any specific mechanism, though our evidence on polarization does suggest two (the 'superstar' and the selection channels). This identification requires one to use microdata. Such data would also help us to understand whether larger cities are more unequal because they increase the dispersion of incomes and wages for a given level of observed skills or types (the so-called residual wage inquality; see, e.g., Lemieux, 2006; Helpman, Itskhoki and Redding, 2008), or whether they are more unequal as the result of the observable divergence of human capital levels across cities (a composition effect).

On the theory side, we are currently extending the model to include sorting according to skills in a non-trivial way. The aim is to build a comprehensive model combining agglomeration, selection and sorting. To our knowledge, such a model is missing to date. The theoretical analysis presented in this paper has also largely left untouched issues that can only be addressed in a rigorous manner by studying the asymmetric equilibria of the model. In the working paper version of this paper (Behrens and Robert-Nicoud, 2008), we extend the model so as to make the notion of 'space' more meaningful by relaxing the symmetry assumption. Specifically, we assume that cities are located on a circle and that trade takes place around this circle, with (the log of) trade costs being proportional to distance. Using this setup, we then construct 'core-periphery equilibria' and 'systems of cities' in which large cities inhibit the existence of small, nearby

cities, thereby 'cannibalizing' them and casting an 'agglomeration shadow'. Preliminary analysis reveals that the equilibrium conditions of the model can be estimated using spatial econometric techniques, and that the results provide some support for 'cannibalization'. To sum up, there are plenty of theoretical and empirical avenues to be explored further, and we leave them open for future work.

References

[1] Asplund, M. and V. Nocke (2006) Firm turnover in imperfectly competitive markets, *Review of Economic Studies* 73, 295-327.

[2] Au, C.-C. and J.V. Henderson (2006) Are Chinese cities too small?, *Review of Economic Studies* 73, 549-576.

[3] Bacolod, M., B.S. Blum and W.C. Strange (2009) Skills in the city, *Journal of Urban Economics* 65, 136-153.

[4] Bairoch, P. (1988) *Cities and Economic Development: From the Dawn of History to the Present*. Chicago, IL: University of Chicago Press.

[5] Baldwin, R.E. (2001) The Core-Periphery model with forward looking expections, *Regional Science and Economics* 31, 21-49.

[6] Behrens, K. and F.L. Robert-Nicoud (2008) Survival of the fittest in cities: Agglomeration, selection and polarisation. CERP Discussion Paper #7018, London; and CEP Discussion Paper #894, London School of Economics, London.

[7] Berry, C. and E.L. Glaeser (2005) The divergence of human capital levels across cities, *Papers in Regional Science* 84, 407-444.

[8] Ciccone, A. and R. Hall (1996) Productivity and the density of economic activity, *American Economic Review* 86, 54-70.

[9] Combes, P.-Ph., G. Duranton and L. Gobillon (2008) Spatial wage disparities: Sorting matters!, *Journal of Urban Economics* 63, 723-742.

[10] Combes, P.-Ph., G. Duranton, L. Gobillon, D. Puga and S. Roux (2009) The productivity advantages of large markets: Distinguishing agglomeration from firm selection. CEPR Discussion Paper #7191, London.

[11] Dobkins, L.H. and Y. Ioannides (2000) Dynamic evolution of the size distribution of US cities. In: Huriot, J.-M. and J.-F. Thisse (eds.) *Economics of Cities: Theoretical Perspectives*. Cambridge, MA: Cambridge University Press, pp. 217-260.

[12] Duranton, G. and D. Puga (2005) From sectoral to functional urban specialisation, *Journal of Urban Economics* 57, 343-370.

[13] Duranton, G. and D. Puga (2004) Micro-foundations of urban agglomeration economies. In: Henderson, J.V. and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics, Vol. 4*, North-Holland: Elsevier, pp. 2063-2117.

[14] Duranton, G. and D. Puga (2001) Nursery cities: Urban diversity, process innovation and the life cycle of products, *American Economic Review* 91, 1454-1477.

[15] Duranton, G. and M. Turner (2008) Urban growth and transportation. University of Toronto, mimeographed.

[16] Fujita, M. (1989) *Urban Economic Theory. Land Use and City Size*. Cambridge, MA: Cambridge University Press.

[17] Fujita, M., P.R. Krugman and A.J. Venables (1999b) *The Spatial Economy. Cities, Regions and International Trade.* Cambridge, MA: MIT Press.

[18] Fujita, M. and H. Ogawa (1982) Multiple equilibria and structural transition of non-monocentric urban configurations, *Regional Science and Urban Economics* 12, 161-196.

[19] Fujita, M. and J.-F. Thisse (2002) *Economics of Agglomeration. Cities, Industrial Location and Regional Growth*. Cambridge MA: Cambridge University Press.

[20] Glaeser, E.L. (1998) Are cities dying?, *Journal of Economic Perspectives* 12, 139-160.

[21] Glaeser, E.L., J. Kolko and A. Saiz (2001) Consumer city, *Journal of Economic Geography* 1, 27-50.

[22] Glaeser, E.L. and D.C. Maré (2001) Cities and skills, *Journal of Labor Economics* 19, 316-342.

[23] Glaeser, E.L., M. Resseger and K. Tobio (2008) Urban inequality. Harvard University, mimeographed.

[24] Harris, J. and M. Todaro (1970) Migration, unemployment and development: A two-sector analysis, *American Economic Review* 60, 126-142.

[25] Head, K. and T. Mayer (2004) The empirics of agglomeration and trade. In: Henderson, J.V. and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics, Vol. 4*, North-Holland: Elsevier, pp. 2609-2670.

[26] Helpman, E., O. Itskhoki and S.J. Redding (2008) Inequality and unemployment in a global economy. Harvard University *and* London School of Economics, mimeographed.

[27] Henderson, J.V. (1974) The size and types of cities, *American Economic Review* 64, 640-656.

[28] Henderson, J.V. (1988) *Urban Development: Theory, Fact and Illusion*. Oxford: Oxford University Press.

[29] Henderson, J.V. (1997) Medium size cities, *Regional Science and Urban Economics* 27, 583-612.

[30] Krugman, P.R. (1991) Increasing returns and economic geography, *Journal of Political Economy* 99, 483-499.

[31] Laguerre, E.N. (1883) Mémoire sur la théorie des équations numériques, *Journal de Mathématiques Pures et Appliquées* 3, 99-146. Translated into English by S.A. Levin (2002), On the theory of numeric equations, Stanford University.

[32] Lemieux, T. (2006) Increasing residual wage inequality: Composition effects, noisy data or skill returns?, *American Economic Review* 96, 461-98.

[33] Long, J.E., D.W. Rasmussen and C.T. Haworth (1977) Income inequality and city size, *Review of Economics and Statistics* 59, 244-246.

[34] Lösch, A. (1940) *Die räumliche Ordnung der Wirtschaft*. Jena: G. Fischer. English translation (1954): *The Economics of Location*. New Haven, CT: Yale University Press.

[35] Lucas, R.E. (1978) On the size distribution of business firms, *Bell Journal of Economics* 9, 508-523.

[36] Lucas, R.E. and E. Rossi-Hansberg (2002) On the internal structure of cities, *Econometrica* 70, 1445-1476.

[37] Madden, J.F. (2000) *Changes in Income Inequality within U.S. Metropolitan Areas*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

[38] Marshall, A. (1890) *Principles of Economics*. London: Macmillan (8th edition, published in 1920).

[39] Melitz, M.J. and G.I.P. Ottaviano (2008) Market size, trade, and productivity, *Review of Economic Studies* 75, 295-316.

[40] Michaels, G., F. Rauch and S. Redding (2008) Urbanization and structural transformation. CEPR Discussion Paper #7016, London.

[41] Mion, G. and P. Naticchioni (2009) The spatial sorting and matching of skills and firms, *Canadian Journal of Economics* 42, 28-55.

[42] Mori, T. and A. Turrini (2005) Skills, agglomeration, and segmentation, *European Economic Review* 49, 201-225.

[43] Nocke, V. (2006) A gap for me: Entrepreneurs and entry, *Journal of the European Economic Association* 4, 929-956.

[44] Nord, S. (1980) Income inequality and city size: An examination of alternative hypotheses for large and small cities, *Review of Economics and Statistics* 62, 502-508.

[45] Okubo, T. (2009) Trade liberalisation and agglomeration with firm heterogeneity: Forward and backward linkages, *Regional Science and Urban Economics*, forthcoming.

[46] Ottaviano, G.I.P., T. Tabuchi and J.-F. Thisse (2002) Agglomeration and trade revisited, *International Economic Review* 43, 409-436.

[47] Partridge, M.D., D.S. Rickman, K. Ali and M.R. Olfert (2009) Do New Economic Geography agglomeration shadows underlie current population dynamics across the urban hierarchy?, *Papers in Regional Science*, forthcoming.

[48] Rosen, S. (1981) The economics of superstars, *American Economic Review* 71, 845-858.

[49] Rosenthal, S. and W. Strange (2004) Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V. and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics, Vol. 4*, North-Holland: Elsevier, pp. 2713-2739.

[50] Tabuchi, T., J.-F. Thisse and D.-Z. Zeng (2005) On the number and size of cities, *Journal of Economic Geography* 5, 423-448.

[51] Weber, M. (1958) *The City*. New York, NY: The Free Press (translated into English by D. Martindale and G. Neuwirth).

[52] Wheeler, C.H. (2001) Search, sorting, and urban agglomeration, *Journal of Labor Economics* 19, 879-899.

Appendix A: Data

We mainly use data from the 2006 American Community Survey (henceforth ACS) released by the US Census Bureau. This dataset reports a large number of socio-economic variables from a sample of three million housing units covering 507 Core Based Statistical Areas (henceforth

CBSAs) in the 50 US states and in Puerto Rico.[16] While the ACS is conducted on an annual basis since 2000, household income Gini coefficients are only reported either since 2006 or in the 2005-2007 3-year estimates sample. The latter dataset offers the advantage of providing more observations (921 CBSAs instead of 507), but limits the number of useable observations de facto as some data for several of the smaller CBSAs is missing. Furthermore, since we supplement our dataset with other data sources (see below for more details), for which many data for the smaller CBSAs is also missing, we rely in what follows on the 2006 ACS to maximize the number of complete observations we can use.

We obtain the following variables for each CBSA from the 2006 ACS (see Table 1 for further details and summary statistics): the population total (`size`); the household median income in 2006 US$ (`medi`); the mean of the i-th quintile of the income distribution in 2006 US$ (`Q`i, for $i = 1, 2, \ldots 5$); the household mean income in 2006 US$, which we compute as the mean of the quintile means of the household income distribution (`meani`); the household income Gini coefficient in 2006 US$ (`gini`); the interquintile range between the means of the 5th and the 1st quintiles, divided by the overall mean income (`interq`); the ratio of the top 5% mean income to the overall mean income (`top5ratio`); the share of population with at least some college education or more, where we impute half of the category of those with some college education to this measure (`edu`); the share of the population below the poverty line (`pov`); a dummy variable for the southern states, as defined by the US Census Bureau classification (`south`); the share of households with two or more income earners (`hh_2plus`); and the share of single-person households in the CBSA (`sphh`).

We further augment the dataset with four variables that capture the ethnic composition of the CBSAs: the share of African-Americans (`black`); the share of Hispanic (`hisp`); the share of Asians and Islanders (`asian`); the share of American-Indians and other natives (`indian`). The data on racial composition is obtained from the 2000 Housing Patters of the US Census Bureau, Housing and Household Economic Statistics Division. The CBSAs in that dataset are matched to fit as closely as possible those of the 2006 ACS data, which leaves us with 495 observations. We compute for each CBSA the share of the corresponding ethnic group in 2000. Note that although the ACS itself provides data on ethnic groups, it only reports that data for the subset of metropolitan statistical areas (367 in the 2006 ACS). We hence prefer to rely on the more exhaustive 2000 data to maximize the number of observations of smaller areas in our sample.

[16] We exclude the Puerto Rican CBSAs from the analysis, which leaves 499 observations. CBSAs collectively refer to both metropolitan and micropolitan statistical areas. Metro areas contain a core urban area with population in excess of 50,000, whereas micro areas contain an urban core with population ranging from 10,000 to 50,000. As explained by the US Census Bureau, metro and micro areas are made up of "one or more counties and include the counties containing the core urban area, as well as any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core". CBSAs constitute hence a natural unit of analysis when the object of study is, as in this paper, centered essentially on the concept of local labor markets.

We checked that the 2000 Housing Patterns series and the 2006 ACS series are highly correlated, thus pointing to the persistence of ethnic composition over short time periods.

We finally construct two variables to control for industrial composition and one variable related to market potential. Firstly, we construct a measure of industrial specialization (`gini_3d`). This is computed as follows:

$$\text{gini\_3d}_i = \sum_k \left| \frac{\text{empl}_{i,k}}{\text{empl}_i} - \frac{\text{empl}_k}{\text{empl}} \right|,$$

where k denotes the sectors and i the CBSA. The employment data comes from the 2002 US Manufacturing Census, and we use the 3-digit level classification (79 sectors). We match again as closely as possible the 2002 Manufacturing Census CBSAs with the 2006 ACS CBSAs. Due to disclosure rules at a small geographical scale, there are a substantial number of withheld data for smaller CBSAs in different sectors: in that case, only an employment range is provided. We chose to assign the average of the range interval to each withheld observation. In the case of the highest top-coded category, we assing the average of the observed highest values in the other CBSAs. Using the employment data, we also construct a variable that reflects the employment share in higher level service industries (`hs_share`). This is defined as the CBSA employment share in the 3-digit sectors 511–562 (which include, among others: Securities and intermediation; Insurance carriers; Funds, trusts, and other financial vehicles; Real estate; Rental and leasing services; Professional, scientific, and technical services, ...). Finally, the net market potential (`net_mp`) is defined as

$$\text{net\_mp}_i = \sum_{k \neq i} \frac{\text{population}_k}{\text{distance}_{ik}},$$

where distance is computed as the great circle distance in kilometers between the largest core cities of the CBSAs. Note that the city's own population is excluded from net_mp_i.

Appendix B: Proofs for Section 4

We prove all propositions of section 4.1. for an arbitrary number Λ of *symmetric* cities, one per region. Since the model is perfectly symmetric by assumption, an equilibrium where all regions have the same size $H_l \equiv H$ and the same cutoff c_l always exists. Let $\Phi \equiv (\Lambda - 1)\tau^{-k}$ denote the 'freeness' of trade. The single-city case corresponds to the situation where $\Phi = 0$ (since $\Lambda = 1$), which also applies when $\tau \to \infty$ (trade is prohibitive). More generally, Φ is increasing in Λ and decreasing in τ and takes value $\Phi = \Lambda - 1$ when $\tau = 1$ (trade is costless). The reader can readily verify that all the proofs in this appendix apply to the special case where $\Phi = 0$, as in Section 4; and to the more general case where $\Phi > 0$, as in Section 5. It turns out that Φ and A enter all expressions together as $(1 + \Phi)A$ so that all comparative static exercises pertaining to the effect of a change in A readily extend to the effects of changes in the freeness of trade.

In the symmetric case with trade and with Λ regions, the free-entry condition (10) in each region reduces to:

$$\frac{(1+\Phi)A}{k+2}Hc_l^{k+2} + \frac{\alpha - c_l}{2\eta}\left[\alpha - \frac{k+1}{k+2}c_l\right] - f^E - \theta H \leq 0. \tag{B.1}$$

Likewise, the identity (8) becomes

$$H = \frac{1}{(1+\Phi)A\eta}\frac{\alpha - c_l}{c_l^{k+1}}. \tag{B.2}$$

Substituting (B.2) into (B.1), and rearranging, we then obtain (21).

B.1. Proof of Proposition 1

Proof. Rewriting f in decreasing order of its powers in c_l, we obtain:

$$f(c_l) = K_1 c_l^2 - K_2 c_l \pm K_3 + K_4 c_l^{-k} - K_5 c_l^{-k-1},$$

where all coefficients K_i are strictly positive. Note that the constant K_3 (which is associated with c_l to the power 0) may a priori be positive or negative, hence the $\pm$ sign in front of it. As one can see, in all cases *there are at most three sign changes from positive to negative or vice versa* between the coefficients of the consecutive powers. Let the number of positive roots be n and the number of sign changes be s. By Laguerre's (1883) generalization of *Descartes' rule of signs*, we know that $n \leq s$ (i.e., there are at most as many positive roots as sign changes) and $(s-n)$ is an even number if $n < s$. Hence, there are either 3 or 1 positive roots in our case. Applying Laguerre's generalization of Descartes' rule to the first and second derivatives of f reveals that f' changes sign at most twice and that f'' changes signs at most once. The final part of the proposition results from the fact that f increases from $-\infty$ at $c_l = 0$. Hence, $\partial f/\partial c_l$ must be strictly positive at the smallest root (whenever one exists). By continuity, and the changes in the signs of the derivatives when there are multiple roots, it follows that there at most two stable equilibria. To see that the third root of f is outside the relevant range $[0, \alpha]$, since $c_l > \alpha$ implies a negative city size which does not make any economic sense, it is sufficient to know that $f(\alpha) = -f^E$ and $\lim_{c_l \to +\infty} f(c_l) = \lim_{c_l \to +\infty} f'(c_l) = +\infty$. Thus, the largest root of f is (strictly) larger than α if (and only if) the parameter f^E is (strictly) positive. ■

B.2. Proof of Proposition 2

Proof. Using (13), it is readily verified that, for any given value of c_l, f is strictly increasing in α or A and decreasing in θ. Assume that $c_l^* \in (0, \alpha]$ is a stable equilibrium. Two cases may arise: either $f(c_l^*) \leq 0$ (with $c_l^* = \alpha$ and $H^* = 0$), which corresponds to the rural equilibrium; or $f(c_l^*) = 0$ with $0 < c_l^* < \alpha$ and $H^* > 0$ at an urban equilibrium. Consider first an increase

in θ. Then f shifts down everywhere, so it must be that $f(c_l^*) < 0$ in the first case: the rural equilibrium remains stable, and $H^* = 0$ is trivially non-increasing from its initial value. In the second case, $f(c_l^*) < 0$ after the shift. Since stability implies $\partial f(c_l^*)/\partial c_l > 0$ and by continuity of f, the new equilibrium must lie to the right of the previous one, hence c_l^* increases and H^* falls by (12). Consider next an increase in α or A. A symmetric argument to the foregoing ensures that c_l^* falls. The overall effect of a rise of A or α on H^* now involves a direct effect, seen in (12), that reinforces the indirect effect in the case of α but that works in the opposite direction in the case of A. This is because a more productive differentiated goods sector requires fewer entrepreneurs to produce the same quantity of output, ceteris paribus. In turn, fewer urban dwellers make it less costly to live in cities, thus triggering urban entry. The net effect turns out to be unambiguous, which can be established by contradiction. Assume that $\mathrm{d}A > 0$ but that $\mathrm{d}H^* < 0$. From (13), $\mathrm{d}H^* < 0$ implies that $(\alpha - c_l)\left[\alpha - (k-1)c_l/(k+2)\right]$ must fall. It turns out that this term is decreasing in c_l over $(0, \alpha]$, thus $\mathrm{d}H^* < 0$ implies $\mathrm{d}c_l^* > 0$ by (13). However, we have previously established that $\partial c_l^*/\partial A < 0$, a contradiction. Therefore, $\partial H^*/\partial A > 0$. ■

In addition, all stable symmetric equilibrium city sizes H^* are non-decreasing in trade freeness Φ by the same token (remember that in the symmetric trading cities model, $(1+\Phi)A$ replaces the term A in the one-city model). The rest of the proof is identical.

B.3. Proof of Lemma 3

Proof. (i) Taking limits, we readily obtain $\lim_{c_l \to 0} f(c_l) = -\infty$, $\lim_{c_l \to 0} \partial f(c_l)/\partial c_l = +\infty$ and $\lim_{c_l \to 0} \partial^2 f(c_l)/\partial c_l^2 = -\infty$. Part (ii) immediately follows by inspection. As to (iii), we have established in Proposition 1 that f admits at most three roots; this part of the lemma is therefore a direct implication of that result and of part (ii). ■

B.4. Proof of Proposition 4

Proof. (i) Condition (12) implies that $H^* = 0$ if and only if $c_l^* = \alpha$. Plugging this result into (13) shows that this inequality holds for any $f^E > 0$. Local stability of the rural equilibrium then immediately follows from the strict inequality. It is useful to show (iii) next. If $f^E = 0$, local stability of the rural equilibrium requires that $\partial f/\partial c_l > 0$ when evaluated at $\{H^*, c_l^*\} = \{0, \alpha\}$. Using (13), some straightforward computations show that this is equivalent to $\theta > \theta^R$, where θ^R is defined in (14). This establishes the stability of the rural equilibrium. To show its existence and to derive a sufficient condition for it to be the only equilibrium, add and subtract (14) in

(12) to obtain:

$$\begin{aligned} f(c_l;\mathbf{Z}) &= \frac{\alpha-c_l}{2\eta}\left[\alpha-\frac{k-1}{k+2}c_l-\frac{3\alpha}{k+2}\left(\frac{\alpha}{c_l}\right)^{k+1}+2\frac{\theta^R-\theta}{Ac_l^{k+1}}\right]-f^E \\ &< \frac{\alpha-c_l}{2\eta}\left[(\alpha-c_l)\frac{k-1}{k+2}+2\frac{\theta^R-\theta}{Ac_l^{k+1}}\right]-f^E, \end{aligned} \tag{B.3}$$

where the inequality stems from $c_l < \alpha$. Imposing $\theta \geq \theta^R$, we further have

$$\begin{aligned} \frac{\alpha-c_l}{2\eta}\left[(\alpha-c_l)\frac{k-1}{k+2}+2\frac{\theta^R-\theta}{Ac_l^{k+1}}\right]-f^E &\leq \frac{\alpha-c_l}{2\eta}\left[(\alpha-c_l)\frac{k-1}{k+2}-2\frac{\theta-\theta^R}{A\alpha_l^{k+1}}\right] \\ &< \frac{\alpha}{2\eta}\left[\alpha\frac{k-1}{k+2}-2\frac{\theta-\theta^R}{A\alpha^{k+1}}\right]-f^E \end{aligned} \tag{B.4}$$

where the first inequality in (B.4) is due to $c_l < \alpha$ and $\theta \geq \theta^R$ and where the second inequality comes from the fact that the second expression in (B.4) is decreasing in c_l. Consequently, when the right-hand side of (B.4) is (weakly) negative, then $f(c_l;\cdot) < 0$ for all values of c_l. In that case, the rural equilibrium is the unique equilibrium. A sufficient condition for this to be so is $f^E \geq f^R$, where

$$f^R \equiv \frac{\alpha^2}{2\eta}\frac{k-1}{k+2}.$$

This establishes the result. ■

To extend the proof to the multi-city case of Section 5, it suffices to replace θ^R by θ^R_Φ and A by $(1+\Phi)A$ in the proof above.

B.5. Proof of Proposition 5

Proof. Parts (i) and (ii) are a re-statement of Proposition 4. (iii) We are looking for a candidate equilibrium with $\alpha > c_l$. In this case, (13) is equivalent to

$$\frac{2\theta}{A} \geq c_l^{k+1}\left(\alpha-\frac{k-1}{k+2}c_l\right), \tag{B.5}$$

the right-hand side of which is strictly concave in c_l, increasing at the limit $c_l \to 0$, and its maximum value on $(0,\alpha]$ is given by $3\alpha^{k+2}/(k+2)$. Thefore, the condition $\theta < \theta^R$ is also sufficient to ensure that there exists a pair $\{H^*, c_l^*\}$ with $c_l^* \in (0,\alpha)$ and $H^* = H(c_l^*)$ from (12) that is compatible with an equilibrium. We finally invoke the continuity of f to establish (iv): at the limit $f^E \to 0$, there exists a finite $\theta^U(f^E)$ by (ii) such that a stable urban equilibrium exists, with $\lim_{f^E\to 0}\theta^U(f^E) = \theta^R$. Since f is continuously differentiable in both f^E and θ, it must be the case that $\theta^U(f^E)$ is positive in the neighborhood of $f^E = 0$ and, by $\partial f/\partial f^E < 0$ and $\partial f/\partial\theta < 0$ (Proposition 2), that $\theta^U(f^E)$ is smaller than θ^R for any f^E. ■

To extend the proof to the multi-city case of Section 5, it suffices to replace θ^R by θ^R_Φ and A by $(1+\Phi)A$ in the proof above.

B.6. Proof of Proposition 6

Proof. (i) The upward-sloping part of the $\cap$-shape of $\overline{\Pi}$ is easily established using (6) and the monotonicity of (12), whereby $\partial c_l/\partial H < 0$:

$$\frac{\partial \overline{\Pi}(H)}{\partial H} = \frac{\alpha - 2c_l}{\eta(k+2)} \frac{\partial c_l}{\partial H},$$

which is non-negative if and only if $\alpha/2 \leq c_l \leq \alpha$. We next compute the second derivative of $\overline{\Pi}$ with respect to city size, which is given by:

$$\frac{\partial^2 \overline{\Pi}(H)}{\partial H^2} = \frac{1}{\eta(k+2)} \left[(\alpha - 2c_l) \frac{\partial^2 c_l}{\partial H^2} - 2 \left(\frac{\partial c_l}{\partial H} \right)^2 \right].$$

A sufficient condition for this to be negative is $\alpha/2 \leq c_l$, since $\partial c_l/\partial H < 0$ and $\partial c_l^2/\partial H^2 > 0$. (ii) Let $\widetilde{\Pi} \equiv \overline{\Pi} \,\big|_{c \leq c_l}$ define the average profit conditional on survival. One can check that

$$\widetilde{\Pi}(c_l) = \frac{H c_l^2}{2\gamma(k+1)(k+2)} = \frac{(\alpha - c_l)c_l}{\eta(k+2)} \left(\frac{c_l}{c_{\max}} \right)^{-k},$$

where the second equality has been obtained from (12). Then

$$\frac{\partial \widetilde{\Pi}(c_l)}{\partial c_l} = -\left(\frac{c_l}{c_{\max}} \right)^{-k} \frac{c_l + (\alpha - c_l)(k-1)}{\eta(2+k)} < 0,$$

where the inequality is due to $0 < c_l \leq \alpha$ and $k \geq 1$. ■

B.7. Proof of Proposition 7

Proof. (i) From (16), we have

$$\sigma_q(c_l) \equiv \frac{k(k+1)(k+2)}{2} \left(\frac{c_l}{c_{\max}} \right)^{-k} \left[\frac{1}{k} - \frac{2}{k+1} \frac{q}{c_l} + \frac{1}{k+2} \left(\frac{q}{c_l} \right)^2 \right].$$

Therefore, for $q \in (0, c_l)$, we have

$$\frac{\partial \sigma_q(c_l)}{\partial c_l} = -\frac{1}{2c_l} \left(\frac{c_l}{c_{\max}} \right)^{-k} \left(1 - \frac{q}{c_l} \right)^2 < 0,$$

and thus $\partial \sigma_q/\partial H > 0$ by $\partial H/\partial c_l < 0$. (ii) Using the foregoing expression we get:

$$\frac{\partial^2 \sigma_q(c_l)}{\partial c_l \partial q} = -\frac{2}{c_l} \frac{\partial \sigma_q(c_l)}{\partial c_l} \left[1 - \frac{q}{c_l} \right]^{-1}$$

which is positive, as was to be shown. ■

B.8. Gini coefficient

In this appendix, we derive the Gini coefficient of income inequality as given by (17). First, since all agents with $c \geq c_l$ have zero income, aggregate income in city l across all draws c is given by

$$W_l(c_l) \equiv H_l G(c_l)\overline{\Pi}(H_l, c_l) = A\frac{H_l^2 c_l^{k+2}}{k+2},$$

where $\overline{\Pi}(H_l, c_l)$ is from (15). The total income accruing to agents with draw $q \leq c_l$ is thus given by

$$W_l(q) \equiv H_l G(c_l)\overline{\Pi}_q(H_l, c_l) = \frac{kH_l^2}{4\gamma}\left(\frac{q}{c_{\max}}\right)^k\left(\frac{c_l^2}{k} - \frac{2q}{k+1} + \frac{q^2}{k+2}\right),$$

where $\overline{\Pi}_q(H_l, c_l)$ is from (16), and their income share is $W_l(q)/W_l(c_l)$. To compute the Gini coefficient, we have to link the income share with the population share. To do so, we need to switch to the distribution in terms of population shares (and not in terms of cost levels c). Let $y \equiv (q/c_{\max})^k$, i.e., $q = y^{1/k}c_{\max}$. Using this change in variables, the new upper bound for integration is given by $y = (c_l/c_{\max})^k$, and we obtain the integral of the Lorenz curve for the surviving agents as follows:

$$\int_0^{\left(\frac{c_l}{c_{\max}}\right)^k} \frac{W_l(y)}{W_l(c_l)}\mathrm{d}y - \int_0^{\left(\frac{c_l}{c_{\max}}\right)^k} x\mathrm{d}x = \frac{2+7k}{4+8k}\left(\frac{c_l}{c_{\max}}\right)^k - \frac{1}{2}\left(\frac{c_l}{c_{\max}}\right)^{2k} \tag{B.6}$$

To finally obtain the Gini coefficient, we need to add the integral of the Lorenz curve for the agents who do not produce. This is given by

$$\int_{\left(\frac{c_l}{c_{\max}}\right)^k}^1 (1-x)\mathrm{d}x = \frac{1}{2}\left[\left(\frac{c_l}{c_{\max}}\right)^k - 1\right]^2 \tag{B.7}$$

Summing (B.7) and (B.6) then yields the Gini coefficient of the income distribution as follows:

$$\mathrm{Gini}(k, c_l) = 1 - \frac{k+2}{4k+2}\left(\frac{c_l}{c_{\max}}\right)^k. \tag{B.8}$$

Appendix C: Proofs for Section 5

C.1. Proof of Proposition 9

Proof. Straighforward rearrangement of (18) yields

$$\frac{\alpha - c_l}{\alpha - c_h}\left(\frac{c_h}{c_l}\right)^{1+k} = \frac{(1-\phi)H_l + \phi\sum_{i=1}^{\Lambda} H_i}{(1-\phi)H_h + \phi\sum_{i=1}^{\Lambda} H_i},$$

which directly implies that

$$c_l < c_h \quad \Longleftrightarrow \quad H_l > H_h.$$

To get the second result, recall that the solution to the linear system $\mathbf{Fh} = \mathbf{x}$ is given by $\mathbf{h} = \det(\mathbf{F})^{-1}\text{cof}(\mathbf{F})^T\mathbf{x}$, where $\text{cof}(\mathbf{F})$ stands for the matrix of cofactors associated with $\mathbf{F}$ and where T denotes the transpose operator. As a result,

$$\frac{\partial H_l}{\partial c_l} = \frac{\det(\mathbf{F}_{l,l})}{\det(\mathbf{F})}$$

where $\det(\mathbf{F}_{l,l})$ is the minor of the $(\Lambda-1)\times(\Lambda-1)$ square matrix cut down from $\mathbf{F}$ by removing its l^{th} column and its l^{th} row. Both matrices $\mathbf{F}_{l,l}$ and $\mathbf{F}$ have only 1's on their main diagonals and ϕ off their main diagonals. Thus, their determinants are also positive, i.e. $\det(\mathbf{F}_{l,l}) > 0$ and $\det(\mathbf{F}) > 0$. By the same token,

$$\frac{\partial H_l}{\partial c_h} = \frac{\det(\mathbf{F}_{h,l})}{\det(\mathbf{F})}.$$

From the Gaussian elimination algorithm, we know that $\det(\mathbf{F}_{h,l}) = -\det(\mathbf{F}_{l,l})$ for $l \neq h$ since $\mathbf{F}_{h,l}$ and $\mathbf{F}_{l,l}$ differ by a column permutation only. Hence $\det(\mathbf{F}_{h,l}) < 0$, which completes the proof. ■

C.2. Gini coefficient and trading cities

Let $z(\Lambda,\tau,k) \equiv -\lambda(\Lambda,\tau,k)/2$ so that (22) may be rewriten as

$$\text{Gini}_l(\Lambda,\tau,k;c_l) = 1 + 2z(\Lambda,\tau,k)\left(\frac{c_l}{c_{\max}}\right)^k,$$

with $z(\cdot) < 0$ for all Λ, τ and k. Fastidious calculations similar to those leading to (B.6) in appendix C.6 yield

$$\begin{aligned} z(\Lambda,\tau,k) &= -1 + \frac{\phi}{2(1+2k)}\frac{(\Lambda-1)\left[(\tau-1)^2(1+2k)(2+k)(1+k) + 2(\tau-1)(2+k)(1+3k) + 2 + 7k\right]}{2\tau^2 + (\Lambda-1)\left[(\tau-1)^2(2+k)(1+k) + 2(\tau-1)(2+k) + 2\right]} \\ &\quad + \frac{1}{2(1+2k)}\frac{(2+7k)\tau^2}{2\tau^2 + (\Lambda-1)\left[(\tau-1)^2(2+k)(1+k) + 2(\tau-1)(2+k) + 2\right]} \end{aligned}$$

from which it follows that $-2z(1,\tau,k) = (2+k)/(2+4k)$ and that $-2z(\Lambda,1,k) = (2+k)/(2+4k)$. We are now equipped to prove the result.

Proof. Differentiating $z(\Lambda,\tau,k)$ with respect to τ yields:

$$\begin{aligned} \frac{\partial z(\Lambda,\tau,k)}{\partial \tau} &= -\frac{k(\Lambda-1)}{1+2k}\left\{\frac{\phi\left[\tau^2\kappa_2 + \tau\kappa_1 + \kappa_0\right]}{\left\{(\Lambda-1)\left[\tau^2(1+k)(2+k) - \tau(2+k)2k + (1+k)k\right] + 2\tau^2\right\}^2}\right. \\ &\quad + \frac{\tau(2+7k)\left[(\tau-1)(2+k)+1\right]}{\left\{(\Lambda-1)\left[\tau^2(1+k)(2+k) - \tau(2+k)2k + (1+k)k\right] + 2\tau^2\right\}^2} \\ &\quad \left. - \frac{\partial\phi}{\partial\tau}\frac{1}{k}\frac{(\tau-1)^2(1+k)(2+k)(1+2k) + 2(\tau-1)(2+k)(3k+1) + (2+7k)}{(\Lambda-1)\left[\tau^2(1+k)(2+k) - \tau(2+k)2k + (1+k)k\right] + 2\tau^2}\right\} \end{aligned}$$

where $\kappa_2 \equiv (\Lambda-1)(1+k)(2+k)^2 - 4k(2+k)$, $\kappa_1 \equiv 3(\Lambda-1)(1+k)(2+k) - 2k(7+2k)$ and $\kappa_0 \equiv (\Lambda-1)(2+k) - 6k$ all have ambiguous signs; therefore, the term in the first line of the

right-hand side above cannot be signed a priori. By contrast, the terms on the second and third lines are positive by inspection. However, if $\phi\left[\tau^2\kappa_2+\tau\kappa_1+\kappa_0\right]$ is negative, then it is larger than $\tau^2\kappa_2+\tau\kappa_1+\kappa_0$ and, adding the terms of the first and second lines, implies that

$$\begin{aligned}
&\phi\left[\tau^2\kappa_2+\tau\kappa_1+\kappa_0\right]+\tau(2+7k)\left[(\tau-1)(2+k)+1\right]\\
&\quad>\tau^2\kappa_2+\tau\kappa_1+\kappa_0+\tau(2+7k)\left[(\tau-1)(2+k)+1\right]\\
&\quad=(2+k)(\Lambda-1)\left[(1+k)(2+k)(\tau-1)^2+3(1+k)(\tau-1)+1\right]\\
&\qquad+(2+k)\left[(2+3k)(\tau-1)^2+3(1+k)(\tau-1)+1\right]>0
\end{aligned}$$

which in turn implies that $\partial z(\Lambda,\tau,k)/\partial\tau<0$ for all Λ, τ and k. We have already established in Proposition 10 that selection gets tougher as trade gets freer ($\partial c_l/\partial\tau>0$), therefore $\partial(\text{Gini}_l)/\partial\tau\equiv 2\left[c_l(\cdot)/c_{\max}\right]^k\left\{\partial z(\cdot)/\partial\tau+z(\cdot)c_l^{-1}\partial c_l(\cdot)/\partial\tau\right\}<0$. ■

For the sake of completeness, note that

$$\begin{aligned}
\frac{\partial z(\Lambda,\tau,k)}{\partial\Lambda} &= -\frac{1}{1+2k}\left\{\frac{-\tau^2\phi\left[(1+k)(2+k)(1+2k)(\tau-1)^2+2(2+k)(1+3k)(\tau-1)+2+7k\right]}{\left\{(\Lambda-1)\left[\tau^2(1+k)(2+k)-\tau(2+k)2k+(1+k)k\right]+2\tau^2\right\}^2}\right.\\
&\quad\left.+\frac{(2+7k)\tau^2\left[(1+k)(2+k)(\tau-1)^2+2(2+k)(\tau-1)+2\right]}{2\left\{(\Lambda-1)\left[\tau^2(1+k)(2+k)-\tau(2+k)2k+(1+k)k\right]+2\tau^2\right\}^2}\right\}\\
&< -\frac{k}{1+2k}\frac{\tau^2(\tau-1)(2+k)\left[3(1+k)(\tau-1)+2\right]}{2\left\{(\Lambda-1)\left[\tau^2(1+k)(2+k)-\tau(2+k)2k+(1+k)k\right]+2\tau^2\right\}^2}<0.
\end{aligned}$$

Therefore, *given* c_l, granting access to more urban markets increases wages of the less productive exporters relative to the wages of the most productive ones; this positive effect is strong enough to overcome the negative one on income inequality that arises as a result of the wages of all successful entrepreneurs going up. However, since selection gets tougher as trade gets freer ($\partial c_l/\partial\tau>0$), the two effects work in opposite directions. Our numerical simulations suggest that the latter indirect effect always dominates the former, direct effect. More precisely, the fact that a larger Λ increases the Gini coefficient *is entirely due to the increase in selection*. By contrast, the fact that a lower τ increases the Gini is due to the increase in selection *and* to the increase of the profits of the most productive entrepreneurs relative to those of the least productive entrepreneurs.

Table 1: Descriptive statistics and correlations

Variable	Description	Obs	Min	Max	Mean	Median	Std deviation
size	CBSA population size	507	65583	18818536	525278	157193	1353798
medi	household median income	507	11717	80638	43749.47	42914	8876.98
meani	household mean income	507	18269	124665	56598.92	55551	11334.47
Q1	household mean income 1st quintile	507	1128	19933	10745.72	10614	2668.657
Q2	household mean income 2nd quintile	507	6192	50095	26811.79	26281	5941.083
Q3	household mean income 3rd quintile	507	11971	80110	43892.1	43304	8852.877
Q4	household mean income 4th quintile	507	21666	122093	66678.73	65904	12416.4
Q5	household mean income 5th quintile	507	50360	362103	134866.3	130528	29667.12
gini	household income Gini coefficient	507	.353	.568	.438	.439	.031
interq	household income interquintile gap to mean	507	1.7608	2.8622	2.1926	2.1804	.1619
top5ratio	household top 5% to mean income	507	2.753	6.310	3.993	3.994	.516
edu	share of college educated	507	.163	.519	.299	.293	.064
pov	share with poverty ratio < 1	507	.046	.570	.149	.140	.064
south	southern states dummy	507	0	1	.410	0	.492
hh_2plus	share of households with 2+ earners	507	.0502	.152	.503	.342	.050
sphh	share of single person households	507	.138	.339	.267	.272	.034
black	share of African-Americans	495	.003	.613	.101	.061	.111
asian	share of Asians-Islanders	495	.003	.735	.025	.013	.056
hisp	share of Hispanic	495	.005	.943	.079	.033	.128
indian	share of Natives-Indians	495	.002	.764	.018	.008	.047
gini_3d	3-digit industrial Isard index	496	.100	.652	.273	.270	.080
hs_share	share of higher level services	496	.036	.389	.142	.132	.056
net_mp	net market potential	507	44.157	805.507	289.405	284.203	92.349

Notes: See the Data Appendix A for further details on data sources and definitions.

Selected correlations

	log(gini)	log(interq)	log(top5ratio)	pov	log(medi)	log(meani)	log(size)
log(gini)	1.000						
ln(interq)	.9901	1.000					
log(top5ratio)	.8501	.8714	1.000				
pov	.5226	.5086	.1855	1.000			
log(medi)	−.3359	−.3117	−.0538	−.8334	1.000		
log(meani)	−.0865	−.0712	.1817	−.7613	.9549	1.000	
log(size)	.2495	.2551	.3066	−.1636	.4050	.5002	1.000

Table 2: City size and mean income

	Base(1)	Robust(2)	Robust(3)	Robust(4)	Robust(5)	Robust(6)	Robust(7)	Robust(8)
Obs.	499	499	499	495	493	359	136	493
Sample	All	All	All	All	All	Metro	Micro	All
Dependent variable	log(meani)	log(meani)	log(meani)	log(meani)	log(meani)	log(meani)	log(meani)	log(meani)
Coefficients:								
log(size)	.0963	.0637	.0610	.0389	.0405	.0392	.0768	.0342
	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.018)	(.000)
log(net_mp)								.0586
								(.000)
edu		.9737	1.148	1.117	1.025	.9834	1.143	1.095
		(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
pov		−1.438	−1.338	−1.912	−1.885	−2.140	−1.184	−1.881
		(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
south		−.0111	−.0007	−.0065	−.0076	−.0181	.0299	−.0041
		(.249)	(.934)	(.464)	(.384)	(.075)	(.071)	(.630)
hh_2plus			.2275	.2118	.2019	.0368	.6073	.1750
			(.055)	(.050)	(.063)	(.801)	(.000)	(.086)
sphh			−.9977	−.5839	−.5760	−.5798	−.4700	−.6247
			(.000)	(.000)	(.000)	(.007)	(.053)	(.000)
black				.2587	.2477	.2873	.0394	.2370
				(.000)	(.000)	(.000)	(.591)	(.000)
asian				.3385	.3232	.3370	.2425	.4648
				(.000)	(.000)	(.000)	(.000)	(.000)
hisp				.2778	.2582	.2823	.1971	.3094
				(.000)	(.000)	(.000)	(.277)	(.000)
indian				.3489	.3258	.1783	.2140	.4252
				(.001)	(.001)	(.040)	(.059)	(.000)
gini_3d					.1418	.2778	.0186	.1055
					(.156)	(.004)	(.928)	(.266)
hs_share					.3004	.3343	.2281	.3401
					(.003)	(.009)	(.180)	(.000)
R^2	.3513	.7162	.7605	.8044	.8089	.7939	.8025	.8182

Notes: p-values in parentheses, robust standard errors. Clustering standard errors by state does not significantly affect our key results. All regressions exclude the CBSAs located in Puerto Rico.

Table 3: City size and mean income (quintile regressions)

Quintile	1	2	3	4	5
log(size)	.0554	.0505	.0484	.0347	.0212
	(.000)	(.000)	(.000)	(.000)	(.000)
Psd R^2	.5833	.5837	.5782	.5854	.6153

Notes: Dependent variable in all regressions is the CBSA log mean income (meani); 493 observations. p-values in parentheses, robust standard errors. Controls included as in specifications Robust(5,6,7) in Table 2 (asian, black, hisp, indian, edu, south, pov, hh_2plus, sphh, gini_3d, hs_share).

Table 4: City size and income (regression on income quintile means)

Income quintile	1st	2nd	3rd	4th	5th
log(size)	.0209	.0340	.0344	.0368	.0477
	(.003)	(.000)	(.000)	(.000)	(.000)
R^2	.8589	.8758	.8452	.8096	.7127

Notes: Dependent variables are the log means of the CBSA income quintiles; 493 observations. p-values in parentheses, robust standard errors. Controls included as in specifications Robust(5,6,7) in Table 2 (asian, black, hisp, indian, edu, south, pov, hh_2plus, sphh, gini_3d, hs_share).

Table 5: City size dilates the income distribution

	Base(1)	Robust(2)	Robust(3)	Robust(4)
Obs.	499	499	499	499
Sample	All	All	All	All
Dependent variable	log(top5ratio)	log(top5ratio)	log(top5ratio)	log(top5ratio)
Coefficients:				
log(size)	.0367	.0310	.0207	.0146
	(.000)	(.000)	(.000)	(.076)
edu		.6221	.5883	.5654
		(.000)	(.000)	(.000)
pov		.6064	.2208	.1920
		(.000)	(.176)	(.009)
south		.0727	.0757	.0777
		(.000)	(.000)	(.000)
hh_2plus			−.3908	−.3991
			(.005)	(.005)
sphh			.6285	.6128
			(.004)	(.005)
black			.0087	.0063
			(.891)	(.920)
asian			.1805	.1982
			(.172)	(.150)
hisp			.2237	.2305
			(.000)	(.000)
indian			.0345	.0478
			(.657)	(.547)
gini_3d				−.1132
				(.288)
hs_share				−.0081
				(.950)
R^2	.0975	.2601	.3180	.3195

Notes: p-values in parentheses, robust standard errors. Clustering standard errors by state does not significantly affect our key results. All regressions exclude the CBSAs located in Puerto Rico.

Table 6: City size and income inequality

	Base(1)	Robust(2)	Robust(3)	Robust(4)	Robust(5)	Robust(6)	Robust(7)	Robust(8)
Obs.	499	499	499	495	493	359	136	493
Sample	All	All	All	All	All	Metro	Micro	All
Dependent variable	log(gini)	log(gini)	log(gini)	log(gini)	log(gini)	log(gini)	log(gini)	log(interq)
Coefficients:								
log(size)	.0309	.0198	.0174	.0128	.0088	.0083	.0324	.0211
	(.000)	(.000)	(.000)	(.000)	(.032)	(.065)	(.110)	(.007)
log(medi)	−.1702	−.0458	.0241	−.0081	−.0018	.0359	−.0891	.0060
	(.000)	(.115)	(.501)	(.826)	(.987)	(.382)	(.136)	(.864)
log(net_mp)								−.0175
								(.054)
edu		.4159	.3426	.3629	.3473	.2921	.5226	.3214
		(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
pov		.7525	.5847	.6381	.6321	.6585	.7075	.6522
		(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
south		.0400	.0381	.0371	.0386	.0388	.0443	.0381
		(.000)	(.000)	(.000)	(.000)	(.000)	(.000)	(.000)
hh_2plus			−.2213	−.2192	−.2268	−.2545	−.1350	−.2238
			(.002)	(.001)	(.001)	(.001)	(.300)	(.001)
sphh			.2465	.3539	.3504	.3564	.3666	.3735
			(.004)	(.000)	(.000)	(.003)	(.037)	(.000)
black				.0434	.0409	.0275	.0146	.0397
				(.133)	(.152)	(.424)	(.778)	(.163)
asian				.1073	.1172	.0071	.2038	.1011
				(.050)	(.040)	(.843)	(.000)	(.062)
hisp				.0877	.0918	.0863	.0444	.0807
				(.000)	(.000)	(.002)	(.670)	(.001)
indian				−.0117	−.0046	.0684	−.0743	−.0255
				(.810)	(.923)	(.379)	(.139)	(.605)
gini_3d					−.0769	−.0551	−.0641	−.0517
					(.139)	(.339)	(.600)	(.351)
hs_share					−.0224	−.0021	−.0218	−.0255
					(.741)	(.987)	(.853)	(.705)
R^2	.2236	.5117	.5179	.5658	.5675	.5317	.6253	.5703

Notes: p-values in parentheses, robust standard errors. Clustering standard errors by state does not significantly affect our key results. All regressions exclude the CBSAs located in Puerto Rico. When using the 2007 ACS one-year estimates of the US Census in Base(1) we obtain similar results (510 observations, log(size) = .0289 (.000) and adj R^2 = .3011). The same holds true when using the 2005-2007 ACS three-year estimates of the US Census in Base(1) (921 observations, log(size) = .0298 (.000) and adj R^2 = .3902), thus suggesting that our results are robust.

Table 7: City size and income inequality (quintile regressions)

Quintile	1	2	3	4	5
log(size)	.0196	.0166	.0104	.0111	.0037
	(.000)	(.000)	(.013)	(.001)	(.652)
Psd R^2	.4200	.3839	.3842	.3827	.3596

Notes: Dependent variable in all regressions is the log income Gini coefficient; 493 observations. p-values in parentheses, robust standard errors. Controls included as in Robust(5,6,7) in Table 6 (medi, asian, black, hisp, indian, edu, south, pov, hh_2plus, sphh, gini_3d, HS_share).

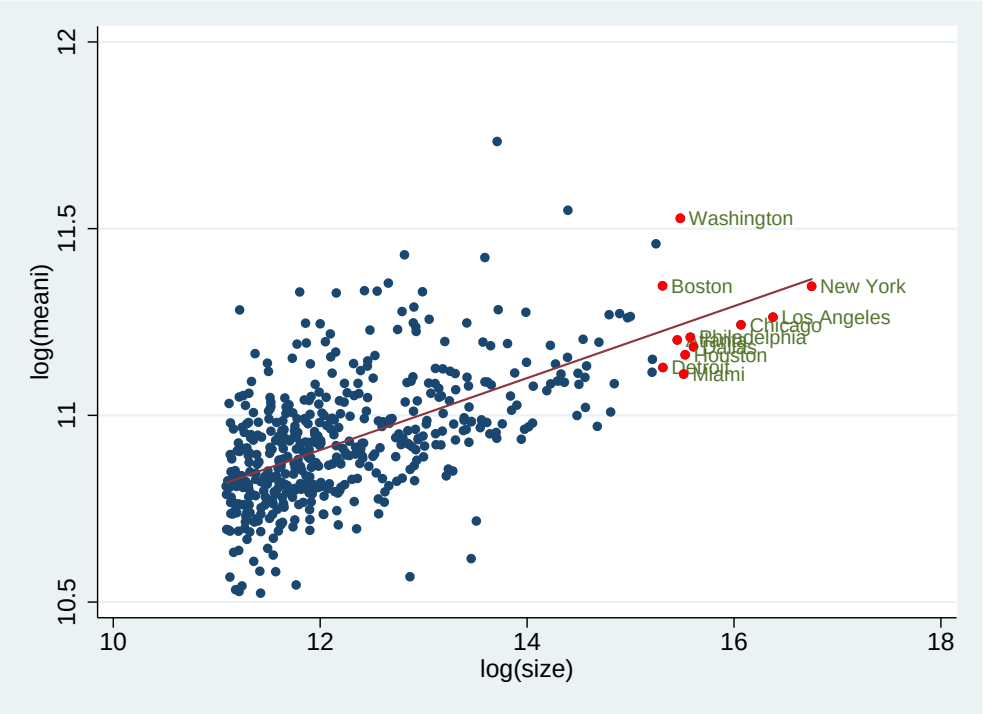

Figure 1a. Mean household income and CBSA size

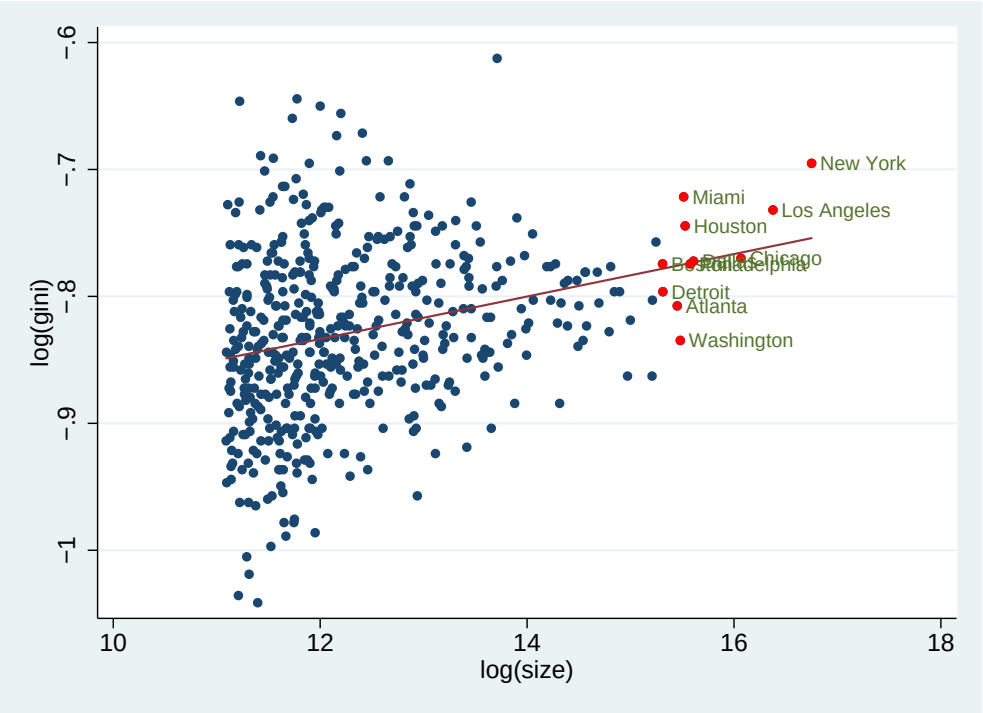

Figure 1b. Household income Gini coefficient and city size

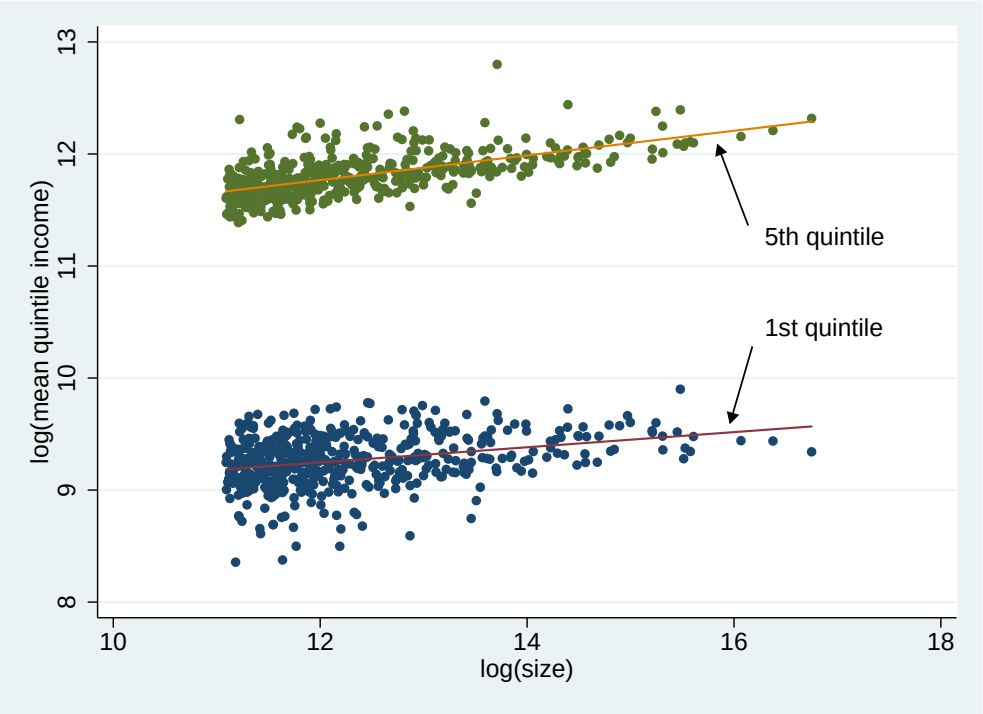

Figure 2a. 1st and 5th income quintile means and city size

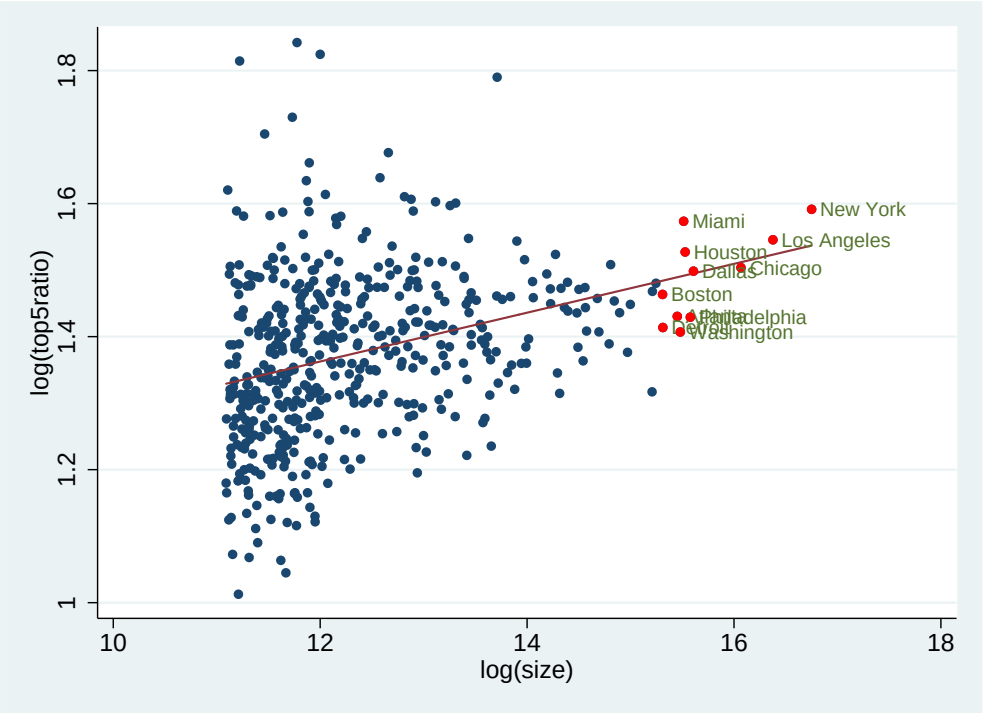

Figure 2b. Top 5% mean income to overall mean income and city size

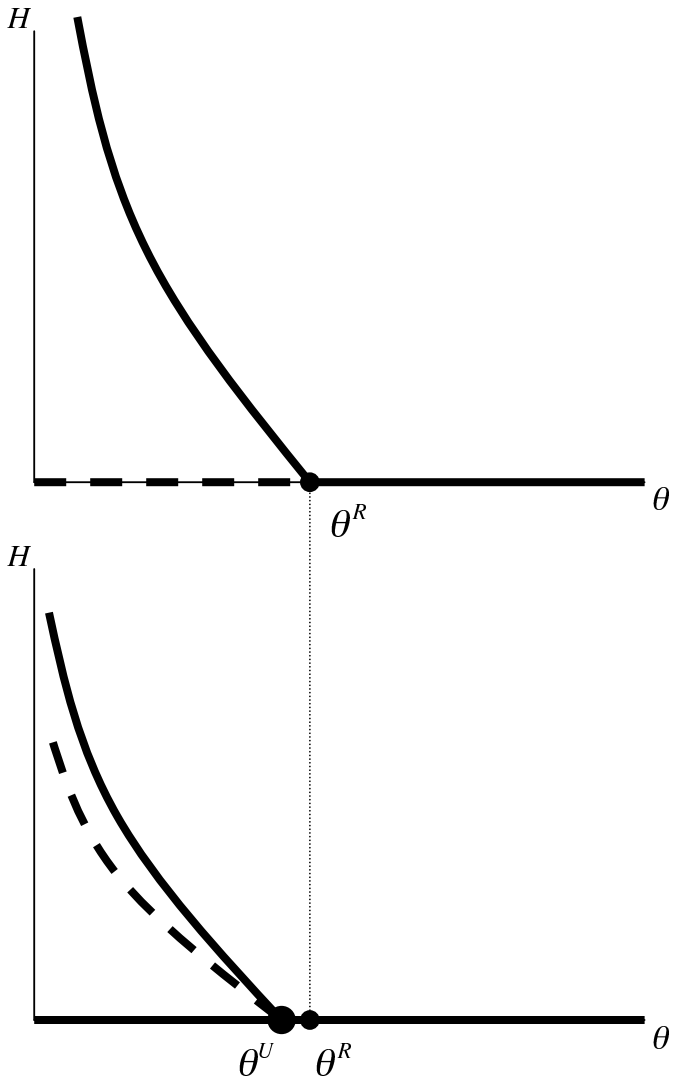

Figure 3. Structure of equilibria with $f^E = 0$ (top panel) and $f^E > 0$ (bottom panel)

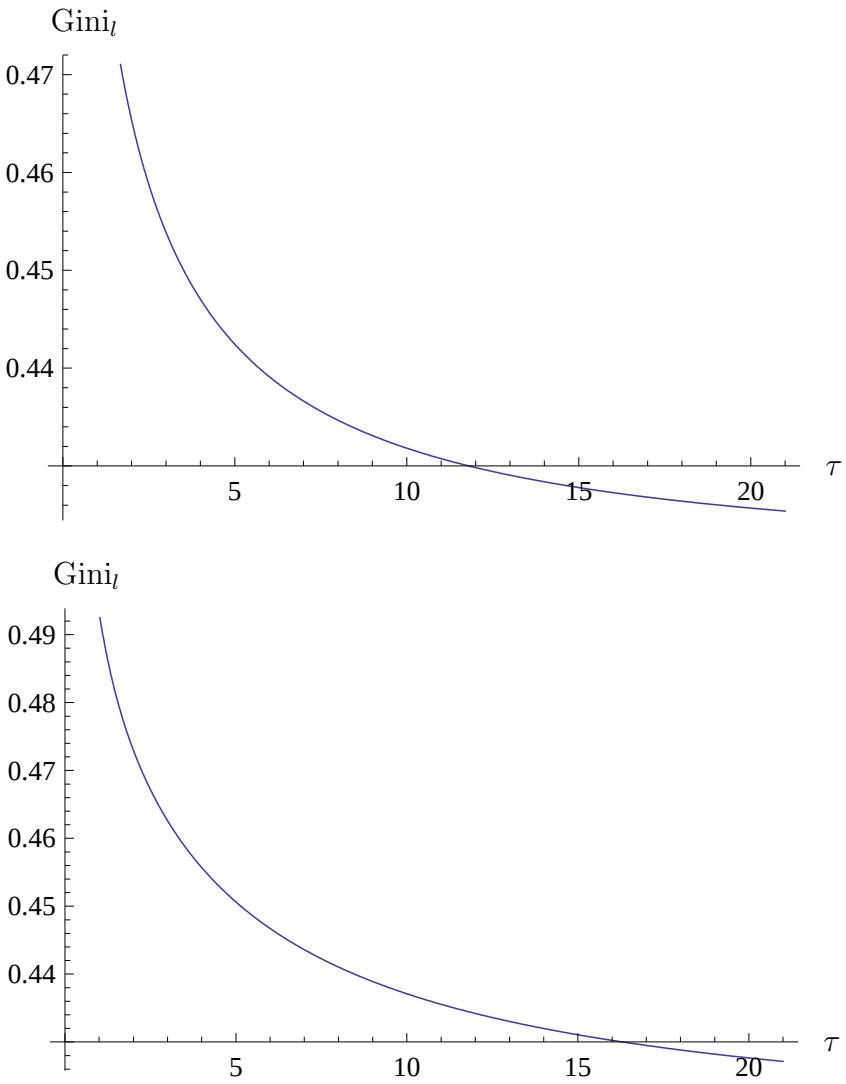

Figure 4. Gini$_l$ as a function of τ for stable c_l^* ($\Lambda = 4$, top; $\Lambda = 8$, bottom)

Technical Appendix: Guide to calculations and extensions, not to be published

T.1. Urban costs

Assume that all city dwellers consume one unit of land, as in standard fixed lot-size models (see Fujita, 1989). Assume further that the central business district (CBD) is located at $x = 0$, so that a city of size H stretches out from $-H/2$ to $H/2$. Without loss of generality, we normalize the opportunity cost of land at the urban fringe to zero: $R(H/2) = R(-H/2) = 0$. Each city dweller commutes to the CBD at constant unit-distance cost $\xi > 0$. Hence, an agent located at x incurrs a commuting cost of $\xi|x|$. Because expected profits and consumer surplus do not depend on city location (see Section 2), the sum of commuting costs and land rent must be identical across locations at a residential equilibrium. This implies that

$$\underbrace{R\left(\frac{H}{2}\right)}_{=0} + \xi\frac{H}{2} = R(x) + \xi|x|,$$

for all x, which yields the equilibrium land rent schedule $R(x) = \xi\left(H/2 - |x|\right)$. The aggregate land rent is thus given by

$$\text{ALR} = \int_{-\frac{H}{2}}^{\frac{H}{2}} R(x)\mathrm{d}x = \frac{\xi}{4}H^2.$$

When ALR is equally redistributed to all agents, equilibrium total urban costs are given by

$$-\frac{\text{ALR}}{H} + R(x) + \xi|x| = \frac{\xi}{4}H.$$

Letting $\theta \equiv \xi/4 > 0$ then yields the expression θH_l for urban costs.

T.2. Return migration

Assume that, upon learning their inverse ability c, entrepreneurs who fail to be successful may return to the countryside at no cost. Assume further that all agents know this piece of information and include it in their entry decision. Starting from the equilibrium conditions of the benchmark model, the mass of people staying in the city is now $G(c_l)H$ with $G(c_l) = \left(c_l/c_{\max}\right)^k$. Let $\widetilde{A} \equiv 1/[2(k+1)\gamma]$. The short run equilibrium condition, which characterizes the mass of varieties actually supplied to consumers, may be rewritten as:

$$\frac{\alpha - c_l}{\widetilde{A}\eta c_l} = G(c_l)H.$$

Next, the profit is given by $\Pi(c) = G(c_l)\frac{H}{4\gamma}(c_l - c)^2$, so that the average profits of the stayers may be written as

$$\widetilde{\Pi} = \int_0^{c_l} \frac{\alpha - c_l}{4\gamma c_l \widetilde{A}\eta}(c_l - c)^2 k\frac{c^{k-1}}{{c_l}^k}\mathrm{d}c = \frac{1}{\eta(2+k)}(\alpha - c_l)c_l. \tag{T.1}$$

Note that (T.1) is positive and concave, as well as decreasing in c_l over $(\alpha/2, \alpha)$. Furthermore, it is readily verified that

$$\begin{aligned}\widetilde{\Pi}(q) &\equiv \int_0^q \frac{\alpha - c_l}{4\gamma c_l \widetilde{A}\eta}(c_l - c)^2 k \frac{c^{k-1}}{c_l^k}\mathrm{d}c \\ &= \frac{(\alpha - c_l)k(k+1)}{2\eta}\left[\frac{{c_l}^{-k+1}q^k}{k} - \frac{2{c_l}^{-k}q^{k+1}}{k+1} + \frac{c_l^{-k-1}q^{k+2}}{k+2}\right],\end{aligned}$$

so that the share of profits accruing to entrepreneurs with a draw smaller than q is given by

$$\sigma(q) \equiv \frac{\widetilde{\Pi}(q)}{\widetilde{\Pi}} = \frac{k(k+1)(k+2)}{2}\left[\frac{{c_l}^{-k}q^k}{k} - \frac{2{c_l}^{-k-1}q^{k+1}}{k+1} + \frac{c_l^{-k-2}q^{k+2}}{k+2}\right],$$

which depends on the inverse average productivity c_l. For any given q, the income share is larger in larger cities (smaller c_l). The Gini coefficient can then finally be computed as follows:

$$\text{Gini} = 1 - 2\left[1 - \int_0^{c_l}\sigma(q)k\frac{q^{k-1}}{{c_l}^k}\mathrm{d}q\right] = \frac{3k}{4k+2}, \tag{T.2}$$

which is independent of city size and solely depends on the distributional parameter $k \geq 1$, despite the fact that $\sigma(q)$ is a function of c_l. Thus, the model with return migration delivers the counterfactual prediction that city size does not matter for income inequality (see Section 2).

T.3. Consumer surplus

Denote by $D_l \equiv \int_{\mathcal{V}_l} d_l(\nu)\mathrm{d}\nu$ the demand for all varieties of the differentiated good. The inverse demand of an agent of type $i = E$ for each variety ν of that good is obtained by maximizing (1) subject to (2) and can be expressed as follows:

$$p_l(\nu) = \alpha - \gamma d_l(\nu) - \eta D_l \tag{T.3}$$

whenever $d_l(\nu) \geq 0$. Denote by $\mathcal{V}_l^+ \subseteq \mathcal{V}_l$ the subset of varieties *effectively consumed* in region l. Expression (T.3) can be inverted to yield a linear demand system as follows:

$$q_l(\nu) \equiv H_l d_l(\nu) = H_l\left[\frac{\alpha}{\eta N_l + \gamma} - \frac{p_l(\nu)}{\gamma} + \frac{\eta N_l}{\eta N_l + \gamma}\frac{\overline{p}_l}{\gamma}\right], \quad \forall \nu \in \mathcal{V}_l^+, \tag{T.4}$$

where $\overline{p}_l \equiv (1/N_l)\int_{\mathcal{V}_l^+} p_l(\nu)\mathrm{d}\nu$ stands for the average price. By definition, $\mathcal{V}_l^+$ is the largest subset of $\mathcal{V}_l$ satisfying

$$p_l(\nu) \leq \frac{\gamma\alpha + \eta N_l \overline{p}_l}{\eta N_l + \gamma} \equiv p_l^d. \tag{T.5}$$

For any given level of product differentiation γ, lower average prices $\overline{p}_l$ or a larger number of competing varieties N_l increase the price elasticity of demand and decrease the price bound p_l^d defined in (T.5). Stated differently, a lower $\overline{p}_l$ or a larger N_l generate a 'tougher' competitive

environment, thereby reducing the maximum price at which entrepreneurs still face positive demand. Letting $p_{hl}(\nu)$ stand for the price of variety ν produced in h and sold in l, the consumer surplus is given by:

$$\begin{aligned} \mathrm{CS}_l &= \frac{\alpha^2 N_l}{2(\eta N_l+\gamma)} - \frac{\alpha}{\eta N_l+\gamma}\sum_h \int_{\mathcal{V}_{hl}^+} p_{hl}(\nu)\mathrm{d}\nu \\ &\quad + \frac{1}{2\gamma}\sum_h \int_{\mathcal{V}_{hl}^+} p_{hl}^2(\nu)\mathrm{d}\nu - \frac{\eta}{2\gamma(\eta N_l+\gamma)}\left[\sum_h \int_{\mathcal{V}_{hl}^+} p_{hl}(\nu)\mathrm{d}\nu\right]^2. \end{aligned} \tag{T.6}$$

T.4. Nash price equilibrium

Let $\pi_{hl}(c) = [p_{hl}(c) - \tau c]\, q_{hl}(c)$ denote operating profits, expressed as a function of the entrepreneur's inverse productivity c. The firms sets prices in order to maximize these profits for each market separately. Then, the profit maximizing prices and output levels must satisfy (for $h \neq l$, with $\tau = 1$ substituted for when $h = l$):

$$p_{hl}(c) = \frac{\gamma\alpha + \eta N_l \overline{p}_l}{2(\eta N_l+\gamma)} + \frac{\tau c}{2} \quad \text{and} \quad q_{hl}(c) = \frac{H_l}{\gamma}\left[p_{hl}(c) - \tau c\right]. \tag{T.7}$$

Integrating the prices in (T.7) over all available varieties, summing across regions and rearranging yields the average delivered price in market l as follows:

$$\overline{p}_l = \frac{\gamma\alpha + \eta N_l \overline{p}_l}{2(\eta N_l+\gamma)} + \frac{\overline{c}_l}{2} \quad \Rightarrow \quad \overline{p}_l = \frac{\gamma\alpha + (\gamma+\eta N_l)\overline{c}_l}{2\gamma+\eta N_l}, \tag{T.8}$$

where

$$\overline{c}_l \equiv \frac{\tau}{N_l}\sum_h \int_{\mathcal{V}_{lh}^+} c\, \mathrm{d}G(c)$$

stands for the average delivered cost of surviving firms selling to l. Plugging (T.8) into (T.7), some straightforward rearrangements show that the Nash equilibrium prices can then be expressed as follows:

$$p_{hl}(c) = \frac{c_l + \tau c}{2}, \quad \text{where} \quad c_l \equiv \frac{2\alpha\gamma + \eta N_l \overline{c}_l}{2\gamma+\eta N_l}$$

denotes the *domestic cost cutoff in region l*. Only entrepreneurs with c 'sufficiently smaller' than c_l are productive enough to sell in city l. This can be seen by expressing q_{hl} in (T.7) more compactly as follows:

$$q_{hl}(c) = H_l \frac{c_l - \tau c}{2\gamma}. \tag{T.9}$$

Clearly, selling in a 'foreign' market l when producing in h requires that $c \leq c_l/\tau$, whereas the analogous condition for selling in the 'domestic' market is given by $c \leq c_l$. In what follows, we denote by c_{hl} the *export cost cutoff for firms producing in region h and selling to region l*. This cutoff must satisfy the zero-profit cutoff condition $c_{hl} = \sup\{c \mid \pi_{hl}(c) > 0\}$. From expressions

(3) and (T.9), this condition can be expressed as either $p_{hl}(c_{hl}) = \tau c_{hl}$ or $q_{hl}(c_{hl}) = 0$, which then yields: $c_{hl} = c_l/\tau$. Clearly, $c_{hl} \leq c_l$ since $\tau \geq 1$. Put differently, trade barriers make it harder for exporters to break even relative to their local competitors because of higher market access costs. Since $p_l^d = p_{ll}(c_l) = c_l$, the zero-profit cutoff condition (T.5) can be expressed as follows:

$$\frac{\gamma\alpha + \eta N_l \overline{p}_l}{\eta N_l + \gamma} = c_l, \quad \text{with} \quad \overline{p}_l = \frac{\alpha\gamma + (\gamma + \eta N_l)\overline{c}_l}{2\gamma + \eta N_l}.$$

We can thus solve for the mass of entrepreneurs selling in region l as follows:

$$N_l = \frac{2\gamma}{\eta}\frac{\alpha - c_l}{c_l - \overline{c}_l}. \tag{T.10}$$

Using the Pareto parametrization of Section 3.4, the average price and the average marginal cost in region l are computed as follows:

$$\overline{p}_l = \frac{2k+1}{2k+2}c_l \quad \text{and} \quad \overline{c}_l = \frac{k}{k+1}c_l,$$

i.e., they are given by a scaling of the domestic cutoff. Using this expression, as well as (T.10), we can then express the mass of sellers in l as follows:

$$N_l \equiv \sum_h H_h G(c_{hl}) = \frac{2\gamma(k+1)(\alpha - c_l)}{\eta c_l},$$

where the first equality comes from the definition of N_l.

The consumer surplus is finally derived by substituting the equilibrium prices into (T.6).

T.5. Expected profits

The expected profit in region l in the symmetric case under the Pareto parametrization is given as follows:

$$\begin{aligned}
\mathbb{E}(\Pi_l) &= \frac{1}{H_l}\left[\frac{H_l}{4\gamma}\int_0^{c_l}(c_l - c)^2\, H_l \mathrm{d}G_l(c) + \sum_{h\neq l}\frac{H_h}{4\gamma}\int_0^{\frac{c_h}{\tau}}(c_h - \tau c)^2\, H_l \mathrm{d}G_l(c)\right] \\
&= \frac{c_{\max}^{-k}[H_l c_l^{k+2} + \tau^{-k}\sum_{h\neq l} H_h c_h^{k+2}]}{2\gamma(k+1)(k+2)} = \frac{A[H_l c_l^{k+2} + \tau^{-k}\sum_{h\neq l} H_h c_h^{k+2}]}{k+2}.
\end{aligned}$$

Using this expression, and noting that neither the consumer surplus nor the urban costs depend on the entrepreneur's ability, we readily obtain expression (10).