Firm selection and agglomeration

Kristian Behrens* Giordano Mion† Yasusada Murata‡ Jens Südekum§

October 15, 2008

Abstract

In this paper we develop a two-region economic geography model with heterogeneous firms. The static model consists of two stages: First, all individuals can freely choose their location. Second, firm entry and selection occurs for given population distribution. This yields an endogenous cost cutoffs for firms to survive in each of the two regions. We show that the larger region has the lower cutoff, and this tougher selection translates into welfare gains in that region. Firm selection can thus be though of as an agglomeration force, which is more pronounced the higher inter-regional transport costs are. This agglomeration force is countervailed by intra-regional commuting congestion. If agglomeration arises the larger region pays the higher wage, has larger consumption variety, and is more "competitive" as it has a lower average consumer price and a lower average markup.

Keywords: heterogeneous firms; new economic geography; monopolistic competition; variable demand elasticities

JEL Classification: F12; F15; F17

Incomplete and preliminary draft. Not for circulation.

*Department of Economics, Université du Québec à Montréal (UQAM), Canada; CORE, Université catholique de Louvain, Belgium; and CIRPÉE, Canada. E-mail: behrens.kristian@uqam.ca.

†CORE, Université catholique de Louvain, Belgium; and FNRS. E-mail: giordano.mion@uclouvain.be

‡Advanced Research Institute for the Sciences and Humanities (ARISH), Nihon University, Tokyo, Japan. E-mail: murata@arish.nihon-u.ac.jp

§Mercator School of Management, Universität Duisburg-Essen, Germany; Ruhr Graduate School of Economics, Germany; and IZA, Germany. E-mail: jens.suedekum@uni-due.de

1 Introduction

The seminal papers by Krugman (1979, 1980) on intra-industry trade under monopolistic competition have revolutionized the field of international trade and inspired several further significant contributions. One crucial development, pioneered by Krugman himself, has been the so-called "new economic geography" (NEG) which incorporates factor mobility into the models of the new trade theory and provides a theory of spatial agglomeration based on endogenous market size effects (see Krugman 1991). A second, more recent extension is the introduction of firm heterogeneity in the Krugman-style trade theory (see Melitz 2003). This important step was motivated by abundant empirical evidence showing that there is substantial firm-level heterogeneity even in narrowly defined industries, and that only a small share of firms is engaged in foreign trade.[1] In Melitz (2003) only the most productive firms self-select into export markets. Trade liberalization forces the least efficient firms to exit, thus leading to an reallocation of market shares and to aggregate productivity gains.[2]

Somewhat surprisingly, these two famous offsprings of the new trade theory have hardly been brought together so far. In this paper we present a framework in the tradition of the NEG with heterogeneous firms. The element of firm heterogeneity is introduced in a similar way as in Melitz (2003) and Melitz and Ottaviano (2008): Firms pay a sunk entry cost and randomly draw their productivity level from a known probability distribution. Only firms which draw a sufficiently high productivity level (inter alia, a sufficiently low marginal cost level) will be able to remain active in the market. That cutoff level is endogenous and depends, in particular, on market size. We show below that the cost cutoff in some region r is decreasing in the population size of that region, which means that larger markets have tougher selection and thus higher average productivity. This in turn implies a welfare gain. Put differently, tougher firm selection in larger markets can be thought of as an agglomeration force.

We consider a two-region economy where all workers are perfectly mobile. It is costly to ship goods across regions, and to commute to the Central Business District (CBD) within each region. Our model is static and has two stages of interaction. In the first stage individuals choose their location. Then, in the second stage, firm entry and selection occurs for given population distribution and yields the endogenous cutoffs in the two regions.[3] The two crucial

[1] These firms which are active on international markets differ along various dimensions from purely domestic ones. Exporters tend, in particular, to be larger and more productive than non-exporters. These firm-level productivity differences act as channels through which trade liberalization brings about aggregate productivity gains, by forcing the least efficient firms to leave the market and by reallocating market shares from low to high productivity firms (e.g., Bernard and Jensen, 1999; Aw *et al.*, 2000; Pavcnik, 2002; Bernard *et at.*, 2007).

[2] Other important contributions to the heterogeneous firms literature with a different focus include Helpman *et al.* (2004), Antràs and Helpman (2004), Bernard et al. (2007), and Melitz and Ottaviano (2008). For an overview of recent advances in the literature, see Helpman (2006).

[3] This simple setting is useful, because it avoids modelling the relocation process of heterogeneous firms and

parameters that affect individual location choices and the spatial equilibrium structure are the level of trade costs and the level of urban costs. If trade costs are high, individuals have a stronger incentive to concentrate in one region, similar as in the NEG-model with homogeneous firms by Helpman (1998). Yet, the agglomeration area not only has more aggregate economic activity, but it also has better (i.e., on average more productive) firms. As the economy gradually integrates, it is important which type of spatial transaction cost (trade costs or urban costs) falls more rapidly. Whereas decreasing transport costs for goods have a dispersive impact, a fall in urban costs is agglomerative.[4]

The model that we develop in this paper is based on our recent multi-region framework with heterogeneous firms (see Behrens et al., 2008), which is extended here to allow for labor mobility across and commuting costs within regions. By building upon that framework, we use a full-fledged general equilibrium model as the underlying basis in which both wages and price-cost margins are endogenous. This is important, because it allows for a more complete account of the process of endogenous spatial agglomeration than in previous analyses. As is well know, a lion's share of the literature is based on the Dixit-Stiglitz-model of monopolistic competition with constant elasticity of substitution across varieties (see Dixit and Stiglitz 1977). This framework is behind both the seminal models of the NEG (Krugman 1991) and the heterogeneous firms literature (Melitz 2003). It generates, however, the vastly unrealistic outcome that all firms (even if they differ in productivity) charge the same constant price markup which is also independent of the spatial distribution of demand, the level of trade integration, etc. This artefact of the Dixit-Stiglitz model has been addressed in the NEG-literature with homogeneous firms by Ottaviano et al. (2002), and in the heterogeneous firms literature with immobile labor by Melitz and Ottaviano (2008). These papers are based on a monopolistic competition model where mark-ups are endogenous, so that they differ across heterogeneous firms and are responsive to trade integration. However, these models assume quasi-linear preferences defined over the set of manufacturing varieties and a homogeneous outside good. This setup rules out all income effects of demand for varieties, and generates equal wages across all regions. Hence, there is no nominal "agglomeration wage premium" in these models, even though empirical evidence strongly suggests that such a premium exists (see Glaeser and Maré 2001; Head and Mayer 2006). In our model the centre and the periphery differ along all these important economic dimensions, and the endogenous regional differences are related to and, in turn, affect the process of agglomeration.

---

the change in the regional distributions of firm productivities that is generated by that relocation. Among the few papers that introduce firm heterogeneity in a NEG-style model are Baldwin and Okubo (2006) and Okubo (2006). Both papers focus on the relocation process among heterogeneous firms, showing that more productive firms are more likely to move first.

[4]See Glaeser and Ponzetto (2007) for a broader discussion how the "death of distance" (decreasing transaction costs for transport, communication and commuting) has affected the spatial structure of the economy.

We show that the larger region pays the higher wage, has tougher selection, features more consumption variety, and is more "competitive" as average consumer prices and markups are lower. Still, despite the various economic mechanisms and channels that are at work in our model, it remains analytically tractable and yields clear-cut results how changes in the exogenous spatial transaction costs affect the spatial equilibrium structure.

The rest of this paper is organized as follows. In section 2 we develop a closed economy version of our model in order to highlight the basic structure of our framework. In section 3 we extend this to a multi-region framework with an arbitrary number of regions, and in section 4 we analyze the case of two (symmetrical) regions in greater depth. Section 5 concludes.

2 Closed economy

Consider a monocentric city endowed with a mass $L > 0$ of identical consumers/workers, as well as with a large amount of homogeneous land. The land stretches out along a one-dimensional space X, and the amount of land available at each location $x \in X$ is set to one. All firms in the city are set up at an exogenously given Central Business District (henceforth, CBD). In what follows, we assume that labor is the only factor of production and that land is used for housing only, i.e., firms do not consume land and the CBD is dimensionless. Without loss of generality, we label locations such that this CBD is located at $x = 0$. Each agent consumes inelastically one unit of land, supplies inelastically one unit of labor, and commutes to the CBD for work. This implies that workers are symmetrically distributed around the CBD and that the city covers the interval $[-L/2, L/2]$.

Following Murata and Thisse (2005), we assume that commuting costs are of the 'iceberg' type: the effective labor supply of a worker living at a distance $|x| \leq L/2$ from the CBD is given by

$$s(x) = 1 - 2\theta|x|. \tag{1}$$

In expression (1), the parameter $\theta > 0$ captures the efficiency loss due to commuting. For the labor supply in efficiency units to be positive regardless of the worker's location x in the city, we assume throughout the paper that $\theta < 1/L$. Consequently, the *total effective labor supply* at the CBD is given by

$$S = \int_{-\frac{L}{2}}^{\frac{L}{2}} s(x)\mathrm{d}x = L\left(1 - \frac{\theta L}{2}\right). \tag{2}$$

Let w stand for the wage rate paid to the workers by the firms at the CBD. Then, the wage income net of commuting costs earned by a worker residing at either city edge is such that $s(-L/2)\,w = s(L/2)\,w = (1-\theta L)w$. Without loss of generality, we normalize the opportunity cost of land to zero. Because workers are identical, the wages net of commuting costs and land rents are equalized across all locations: $s(x)w - R(x) = s(-L/2)w = s(L/2)w$, where $R(x)$ is the land rent prevailing at x, and $R(L/2) = R(-L/2) = 0$. For a given spatial distribution of workers, the equilibrium land rent schedule in the city is therefore given by $R^*(x) = \theta(L - 2|x|)w$, which yields the following aggregate land rents:

$$\text{ALR} = \int_{-\frac{L}{2}}^{\frac{L}{2}} R^*(x)\mathrm{d}x = \frac{\theta L^2 w}{2}.$$

In what follows, we assume that each worker owns an equal share of land in the city and has equal claims to firms' profits. Accordingly, in addition to her wage, each worker receives an equal share ALR$/L$ of aggregate land rents from her land ownership, and an equal share of aggregate profits Π.

2.1 Preferences and demands

In this closed economy there is a final consumption good which is provided as a continuum of horizontally differentiated varieties. We denote by Ω the endogenously determined set of available varieties with measure N.

Following Behrens and Murata (2007) and Behrens et al. (2008), we assume that all consumers have identical preferences which display 'love of variety' and give rise to demand functions featuring variable price elasticities. The utility maximization problem of a representative consumer is given by:

$$\max_{q(j),\, j\in\Omega} \; U \equiv \int_{\Omega} \left[1 - \mathrm{e}^{-\alpha q(j)}\right]\mathrm{d}j \qquad \text{s.t.} \quad \int_{\Omega} p(j)q(j)\mathrm{d}j = E,$$

where $E \equiv (1 - \theta L)w + \text{ALR}/L$ stands for expenditure; $p(j) > 0$ and $q(j) \geq 0$ stand for the price and the per capita consumption of variety j; and $\alpha > 0$ is a parameter. As shown by Behrens et al. (2008), this preference structure yields the following demand function for a single variety:

$$q(i) = \frac{E}{N\overline{p}} - \frac{1}{\alpha}\left\{\ln\left[\frac{p(i)}{N\overline{p}}\right] + h\right\},$$

where

$$\overline{p} \equiv \frac{1}{N}\int_{\Omega} p(j)\mathrm{d}j \quad \text{and} \quad h \equiv -\int_{\Omega} \ln\left[\frac{p(j)}{N\overline{p}}\right]\frac{p(j)}{N\overline{p}}\mathrm{d}j$$

denote the average price and the differential entropy of the price distribution, respectively. Since marginal utility evaluated at zero consumption is bounded, the demand for a variety need not be positive. Indeed, as can be seen, the demand for variety i is strictly positive if and only if its price is lower than the reservation price p^d. Formally,

$$q(i) > 0 \quad \Longleftrightarrow \quad p(i) < p^d \equiv N\overline{p}\,\mathrm{e}^{\frac{\alpha E}{N\overline{p}} - h}.$$

Note that the reservation price p^d is a function of the price aggregates $\overline{p}$ and h. Combining these expressions allows us to express the demands compactly as follows:

$$q(i) = \frac{1}{\alpha}\ln\left[\frac{p^d}{p(i)}\right]. \tag{3}$$

2.2 Technology and market structure

The labor market is assumed to be perfectly competitive so that all firms take the wage rate w as given. Firms are heterogeneous in our model. As in Behrens et al. (2008) we assume that prior to production, each firm engages in research and development, which requires a fixed amount F of labor paid at the market wage. Each entrant discovers its marginal labor requirement $m(i) \geq 0$ only after making this irreversible investment. We assume that $m(i)$ is drawn from a common and known, continuously differentiable distribution G. Since research and development costs are sunk, a firm will remain active in the market provided it can charge a price $p(i)$ above marginal cost $m(i)w$.

Each surviving firm sets its price to maximize operating profit

$$\pi(i) = L\big[p(i) - m(i)w\big]q(i), \tag{4}$$

Since there is a continuum of firms, no individual firm has any impact on p^d so that the first-order conditions for (operating) profit maximization are given by:

$$\ln\left[\frac{p^d}{p(i)}\right] = \frac{p(i) - m(i)w}{p(i)}, \quad \forall i \in \Omega. \tag{5}$$

A price distribution satisfying (5) is called a *price equilibrium*. Multiplying both sides of (5) by $p(i)$, integrating over Ω, and using (3) yield the average price as follows:

$$\overline{p} = \overline{m}w + \frac{\alpha E}{N}, \tag{6}$$

where $\overline{m} \equiv (1/N)\int_\Omega m(j)\mathrm{d}j$ denotes the average marginal labor requirement of the surviving firms. Observe that expression (6) displays pro-competitive effects, i.e., the average price is decreasing in the mass of surviving firms N.

Equations (3) and (5) imply that $q(i) = (1/\alpha)[1 - m(i)w/p(i)]$, which allows us to derive the upper and lower bounds for the marginal labor requirement. The maximum output is given by $q(i) = 1/\alpha$ at $m(i) = 0$. The minimum output is given by $q(i) = 0$ at $p(i) = m(i)w$, which by (5) implies that $p(i) = p^d$. Therefore, the cutoff marginal labor requirement is defined as $m^d \equiv p^d/w$. A firm that draws m^d is indifferent between producing and not producing, whereas all firms with a draw below (resp., above) m^d remain in (resp., exit from) the market.

Since firms differ only by their marginal labor requirement, we can express all firm-level variables in terms of m. Solving (5) by using the Lambert W function, the profit-maximizing prices and quantities as well as operating profits can be expressed as follows:

$$p(m) = \frac{mw}{W}, \quad q(m) = \frac{1}{\alpha}(1 - W), \quad \pi(m) = \frac{Lmw}{\alpha}\left(W^{-1} + W - 2\right), \tag{7}$$

where we suppress the argument $\mathrm{e}m/m^d$ of W to alleviate notation.

It is readily verified that $W' > 0$ for all non-negative arguments and that $W(0) = 0$ and $W(\mathrm{e}) = 1$. Hence, $0 \leq W \leq 1$ if $0 \leq m \leq m^d$. The expressions in (7) then show that a firm with draw m^d charges a price equal to marginal cost, faces zero demand, and earns zero profit. Since $W' > 0$, we readily obtain $\partial p(m)/\partial m > 0$, $\partial q(m)/\partial m < 0$ and $\partial \pi(m)/\partial m < 0$. In words, firms with better draws charge lower prices, sell larger quantities, and earn higher operating profits than firms with worse draws. [5]

2.3 Equilibrium

We derive the equilibrium conditions for the closed economy, which consist of zero expected profits and labor market clearing. First, given the mass of entrants N^E, the mass of surviving firms can be written as $N = N^E G(m^d)$. Using (4), the zero expected profit condition for each firm is given by

$$L\int_0^{m^d} [p(m) - mw]\, q(m)\mathrm{d}G(m) = Fw. \tag{8}$$

Furthermore, the labor market clearing condition can be expressed as follows:[6]

$$N^E\left[L\int_0^{m^d} mq(m)\mathrm{d}G(m) + F\right] = S. \tag{9}$$

Making use of the profit-maximizing prices and quantities in terms of the Lambert W function (given above in (7)), we can rewrite the zero expected profit condition (8) as follows

$$\frac{L}{\alpha}\int_0^{m^d} m\left(W_m^{-1} + W_m - 2\right)\mathrm{d}G(m) = F. \tag{10}$$

This equation yields a unique equilibrium cutoff level as the left-hand side of (10) is strictly increasing in m^d from 0 to ∞ (see Appendix A.1). Using expression (7), the labor market clearing condition (9) can be rewritten as follows:

$$N^E\left[\frac{L}{\alpha}\int_0^{m^d} m\left(1 - W_m\right)\mathrm{d}G(m) + F\right] = S, \tag{11}$$

[5] Further details about the Lambert W function, which is the key technical property that renders our model analytically tractable, can be found in Appendix A.1 of Behrens et al. (2008)

[6] Note that using (8) and the budget constraint $N^E\int_0^{m^d} p(m)q(m)\mathrm{d}G(m) = E$, we obtain $EL/(wN^E) = L\int_0^{m^d} mq(m)\mathrm{d}G(m) + F$, which together with (9), yields $E = (S/L)w$ in equilibrium.

Given the equilibrium cutoff m^d, equation (11) can be uniquely solved for N^E. In other words, in the closed economy there is always a unique equilibrium.

Turning to a comparative-static analysis we can address how market size L affects the firms' entry and survival probability. Using the equilibrium conditions (10) and (11), we can show that larger market size leads to a smaller cutoff m^d and a greater mass of entrants N^E, respectively (see Appendix A.2). This is true irrespective of the the distribution function G from which firms draw their marginal labor requirements, but it can be seen more explicitly when making the common assumption that firms draw from a Pareto distribution

$$G(m) = \left(\frac{m}{m^{\max}}\right)^k,$$

where $m^{\max} > 0$ denotes the upper bound of the support and $k \geq 1$ is a shape parameter.[7] Under the Pareto parametrization, we readily obtain the closed-form solutions for the equilibrium cutoff and the equilibrium mass of entrants:[8]

$$m^d = \left[\frac{\alpha F \left(m^{\max}\right)^k}{\kappa_2 L}\right]^{\frac{1}{1+k}} \quad \text{and} \quad N^E = \frac{S}{F}\left(1 + \frac{\kappa_1}{\kappa_2}\right)^{-1},$$

where κ_1 and κ_2 are positive constants that solely depend on the shape parameter k (see Appendices B.1 and B.2). As can be seen a larger population size L decreases the cutoff level m^d and thus also the average marginal labor requirement, $\overline{m} = [k/(k+1)]m^d$, which under the Pareto parametrization is proportional to the cutoff. In other words, average firm productivity is higher in larger markets as there is tougher firm selection. A larger population size L also increases effective labor supply ($dS/dL > 0$), and thereby the equilibrium number of entrants N^E. The effect of a larger population size on the number of surviving firms N is ambiguous, however.

2.4 Welfare

Finally, welfare can be written as follows. Since $\mathrm{e}^{-\alpha q(m)} = p(m)/p^d$ by (3), the indirect utility is given by

$$U = N^E \int_0^{m^d} \left[1 - \mathrm{e}^{-\alpha q(m)}\right] \mathrm{d}G(m) = N\left(1 - \frac{\overline{p}}{p^d}\right).$$

Using expression (6), one can verify that $\overline{p} = [k/(k+1)]p^d + \alpha E/N$, which allows to express the indirect utility as follows:

$$U = \frac{N}{k+1} - \frac{\alpha E}{m^d w} = \frac{\alpha \kappa S}{m^d L} \quad \text{where} \quad \kappa \equiv \frac{1}{(k+1)\kappa_3} - 1 > 0.$$

[7]The Pareto distribution has been extensively used in the previous literature on heterogeneous firms (Bernard *et al.*, 2007; Helpman *et al.*, 2008; Melitz and Ottaviano, 2008). Such a distribution is also consistent with the U.S. firm size distribution (see Axtell, 2001).

[8]For this solution to be consistent, we have to make sure that $m^d \leq m^{\max}$, i.e., that $m^{\max} \geq (\alpha F)^{1/k} L^{-1/k}$. In what follows, we assume that this condition always holds by restricting the parameter values accordingly.

The equilibrium indirect utility depends negatively on the cutoff level m^d, i.e., tougher selection implies welfare gains in this model. Secondly indirect utility depends positively on the effective labor supply rate S/L. Using (2) indirect utility can be rewritten as

$$U = \alpha\kappa(\mu^{\max})^{-\frac{1}{1+k}} \left(1 - \frac{\theta L}{2}\right) L^{\frac{1}{1+k}} \quad \text{where} \quad \mu^{\max} \equiv \frac{\alpha F \left(m^{\max}\right)^k}{\kappa_2 L}, \tag{12}$$

and this expression gives rise to the following result:

Proposition 1 *U is single peaked with respect to L.*

Proof. Differentiating (12) with respect to L, it is readily verified that

$$\frac{\partial U}{\partial L} = \alpha\kappa(\mu^{\max})^{-\frac{1}{1+k}} \left[\frac{2-(2+k)\theta L}{2(1+k)L^{\frac{k}{1+k}}}\right],$$

thus showing that U is increasing when L is small and is decreasing when L is large. ■

An increase of the population size L leads to a decline in the cutoff m^d, i.e., a larger population size raises average productivity by inducing tougher selection. This mechanism can be thought of as an agglomeration force in our model. The increase in population size also raises labor supply S, but at a smaller rate so that the labor supply rate S/L decreases. This effect represents the congestion force in our model, the efficiency loss due to commuting. The latter is more important in populous economies since the overall supply of land is fixed. Hence the net effect of an increase in population size on welfare is only positive if the level of L is sufficiently small.

3 Open economy

We now turn to the open economy case. In this section we first derive the equilibrium conditions for the general case with K asymmetric regions, assuming that population sizes (and therefore regional labor supplies) are given. Using this setup we impose the following timing of events in our model: First, workers/consumers choose their location, and second, the firms' entry, productivty draws and the subsequent production decisions take place. In other words we solve the second stage of this game in the current section and focus on a short-run equilibrium constellation where utility may still differ for a given population distribution. In the next section we then analyze the special case of two symmetric regions in greather depth. In that section we also move to the long-run equilibrium perspective (the first stage) where individuals endogenously choose their location.[9]

[9]Notice that this timing greatly simplies our model with heterogeneous firms, because we need not specify which type of firms relocates as workers move across regions. Furthermore, notice that we assume that *all* workers are mobile across regions.

3.1 Preferences and demands

Preferences are analogous to the ones described in the previous section. Let $p_{sr}(i)$ and $q_{sr}(i)$ denote the price and the per capita consumption of variety i when it is produced in region s and consumed in region r. It is readily verified that the demand functions in the open economy case are given as follows:

$$q_{sr}(i) = \frac{E_r}{N_r^c \overline{p}_r} - \frac{1}{\alpha}\left\{\ln\left[\frac{p_{sr}(i)}{N_r^c \overline{p}_r}\right] + h_r\right\}, \quad \forall i \in \Omega_{sr},$$

where N_r^c is the mass of varieties consumed in region r; Ω_{sr} denotes the set of varieties produced in region s and consumed in region r; and

$$\overline{p}_r \equiv \frac{1}{N_r^c}\sum_s \int_{\Omega_{sr}} p_{sr}(j)\mathrm{d}j \quad \text{and} \quad h_r \equiv -\sum_s \int_{\Omega_{sr}} \ln\left[\frac{p_{sr}(j)}{N_r^c \overline{p}_r}\right]\frac{p_{sr}(j)}{N_r^c \overline{p}_r}\mathrm{d}j$$

denote the average price and the differential entropy of all varieties consumed in region r. As in the closed economy case, there exist reservation prices for the various demands to be positive. The reservation prices for domestic and foreign varieties in market r are given by

$$\begin{aligned} q_{rr}(i) > 0 \quad &\Longleftrightarrow \quad p_{rr}(i) < p_r^d \equiv N_r^c \overline{p}_r \mathrm{e}^{\frac{\alpha E_r}{N_r^c \overline{p}_r} - h_r} \\ q_{sr}(j) > 0 \quad &\Longleftrightarrow \quad p_{sr}(j) < p_r^d \equiv N_r^c \overline{p}_r \mathrm{e}^{\frac{\alpha E_r}{N_r^c \overline{p}_r} - h_r}. \end{aligned}$$

Note that p_r^d is a function of the price aggregates $\overline{p}_r$ and h_r.

Using the reservation price, the demands for domestic and foreign varieties can be concisely expressed as follows:

$$q_{rr}(i) = \frac{1}{\alpha}\ln\left[\frac{p_r^d}{p_{rr}(i)}\right] \quad \text{and} \quad q_{sr}(j) = \frac{1}{\alpha}\ln\left[\frac{p_r^d}{p_{sr}(j)}\right]. \tag{13}$$

3.2 Technology and market structure

Technology and the entry process are analogous to the ones described in Section 2. We assume that shipments from r to s are subject to transport costs $\tau_{rs} > 1$, incurred by firms. We do not exclude the case where $\tau_{rr} > 1$ in order to capture the fact that firms also incur shipping costs in their domestic markets. We further assume, in accord with abundant empirical evidence, that markets are segmented and that firms are free to price discriminate.

Firms in region r independently draw their productivities from a region-specific distribution G_r. Assuming that firms incur transport costs in terms of labor, the operating profit of firm i in region r is given by:

$$\pi_r(i) = \sum_s \pi_{rs}(i) = \sum_s L_s q_{rs}(i)\left[p_{rs}(i) - m_r(i)\tau_{rs} w_r\right]. \tag{14}$$

Each firm i in region r maximizes (14) with respect to its prices $p_{rs}(i)$ separately. Since it has no impact on the price aggregates and the wages, the first-order conditions are given by:

$$\ln\left[\frac{p_s^d}{p_{rs}(i)}\right] = \frac{p_{rs}(i) - m_r(i)\tau_{rs}w_r}{p_{rs}(i)}, \quad \forall i \in \Omega_{rs}. \tag{15}$$

We first solve for the average price in market r. To do so, multiply (15) by $p_{rs}(i)$, use (13), integrate over Ω_{rs}, and finally sum the resulting expressions to obtain

$$\overline{p}_r \equiv \frac{1}{N_r^c}\sum_s \int_{\Omega_{sr}} p_{sr}(j)\mathrm{d}j = \frac{1}{N_r^c}\sum_s \tau_{sr}w_s \int_{\Omega_{sr}} m_s(j)\mathrm{d}j + \frac{\alpha E_r}{N_r^c}, \tag{16}$$

where the first term is the average of marginal delivered costs for all varieties sold in market r. Clearly, the average price in region r is increasing in this term. It is also increasing in expenditure E_r and decreasing in the mass N_r^c of firms competing in market r. All these results mimic the ones established in the closed economy case.

Equations (13) and (15) imply that $q_{rs}(i) = (1/\alpha)[1 - \tau_{rs}m_r(i)w_r/p_{rs}(i)]$, which shows that $q_{rs}(i) = 0$ at $p_{rs}(i) = \tau_{rs}m_r(i)w_r$. It then follows from (15) that $p_{rs}(i) = p_s^d$. Hence, a firm located in r with draw $m_{rs}^x \equiv p_s^d/(\tau_{rs}w_r)$ is just indifferent between selling and not selling in region s. All firms with draws below m_{rs}^x are productive enough to sell to region s. In what follows, we refer to $m_{ss}^x \equiv m_s^d$ as the *domestic cutoff* in region s, whereas m_{rs}^x with $r \neq s$ is the *export cutoff*.

Export and domestic cutoffs are linked as follows:

$$m_{rs}^x = \frac{\tau_{ss}}{\tau_{rs}}\frac{w_s}{w_r}m_s^d. \tag{17}$$

Expression (17) reveals how trade costs and wage differentials affect firms' ability to break into foreign markets. When wages are equalized ($w_r = w_s$) and internal trade is costless ($\tau_{ss} = 1$), all export cutoffs must fall short of the domestic cutoffs since $\tau_{rs} > 1$. In that case, breaking into any foreign market is always harder than selling domestically. However, in the presence of wage differentials and internal trade costs, the domestic and the foreign cutoffs can no longer be clearly ranked. The usual ranking, namely that exporting to s is more difficult than selling domestically in s, prevails only when $\tau_{ss}w_s < \tau_{rs}w_r$.

Switching to notation in terms of m, we can solve the first-order conditions (15) for the profit-maximizing prices and quantities as a function of market aggregates as follows:

$$p_{rs}(m) = \frac{m\tau_{rs}w_r}{W\left(\mathrm{e}\,\frac{m\tau_{rs}w_r}{p_s^d}\right)} \quad \text{and} \quad q_{rs}(m) = \frac{1}{\alpha}\left[1 - W\left(\mathrm{e}\,\frac{m\tau_{rs}w_r}{p_s^d}\right)\right], \tag{18}$$

where W denotes again the Lambert W function. It is readily verified that prices and quantities exhibit the same behavior with respect to the marginal labor requirement m as in the closed economy case. Therefore, more productive firms charge lower prices and sell larger quantities.

Contrary to the closed economy case, the masses of varieties consumed and produced in each region need not be the same. Given a mass of entrants N_r^E, only a subset of the mass of firms actually produces varieties. More precisely, only

$$N_r^p = N_r^E \, G_r \left(\max_s \{ m_{rs}^x \} \right)$$

firms survive (those with a small enough marginal labor requirement that allows them to sell at least in one market). The mass of varieties consumed in region r is given by

$$N_r^c = \sum_s N_s^E G_s(m_{sr}^x),$$

which depends quite naturally on the region's own distribution G_r as well as on those of all its trading partners.

3.3 Equilibrium

The zero expected profit condition for each firm in region r is given by

$$\sum_s L_s \int_0^{m_{rs}^x} \left[p_{rs}(m) - w_r \tau_{rs} m \right] q_{rs}(m) \mathrm{d}G_r(m) = F_r w_r, \tag{19}$$

where F_r is the region-specific fixed labor requirement. Furthermore, each labor market clears in equilibrium, which yields

$$N_r^E \left[\sum_s L_s \tau_{rs} \int_0^{m_{rs}^x} m q_{rs}(m) \mathrm{d}G_r(m) + F_r \right] = S_r. \tag{20}$$

In addition, we assume that trade is balanced for each region:

$$N_r^E \sum_{s \neq r} L_s \int_0^{m_{rs}^x} p_{rs}(m) q_{rs}(m) \mathrm{d}G_r(m) = L_r \sum_{s \neq r} N_s^E \int_0^{m_{sr}^x} p_{sr}(m) q_{sr}(m) \mathrm{d}G_s(m).$$

In what follows, we assume that productivity draws $1/m$ follow a Pareto distribution with identical shape parameters $k \geq 1$. However, to capture local technological possibilities, we allow the upper bounds to differ across regions, i.e., $G_r(m) = (m/m_r^{\max})^k$. A lower $m_r^{\max}$ implies that firms in region r have a higher probability of drawing a better productivity.

Under the Pareto parametrization, these equilibrium conditions can be expressed quite compactly. First, using the expression in Appendix B.1, the labor market clearing conditions can be written as:

$$N_r^E \left[\frac{\kappa_1}{\alpha \left(m_r^{\max} \right)^k} \sum_s L_s \tau_{rs} \left(m_s^d \frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} \right)^{k+1} + F_r \right] = S_r. \tag{21}$$

Second, using the expression in Appendix B.2, zero expected profits imply that

$$\mu_r^{\max} \equiv \frac{\alpha F_r \left(m_r^{\max}\right)^k}{\kappa_2} = \sum_s L_s \tau_{rs} \left(m_s^d \frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} \right)^{k+1}, \tag{22}$$

where μ_r is a simple monotonic transformation of the upper bounds.

Last, using the expression in Appendix B.3, the trade balance conditions can be rewritten as:

$$\frac{N_r^E w_r}{\left(m_r^{\max}\right)^k} \sum_{s \neq r} L_s \tau_{rs} \left(m_s^d \frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r} \right)^{k+1} = L_r \sum_{s \neq r} \frac{N_s^E w_s}{\left(m_s^{\max}\right)^k} \tau_{sr} \left(m_r^d \frac{\tau_{rr}}{\tau_{sr}} \frac{w_r}{w_s} \right)^{k+1}. \tag{23}$$

The $3K$ conditions (21)–(23) depend on $3K$ unknowns: the wages w_r, the masses of entrants N_r^E, and the domestic cutoffs m_r^d. The different export cutoffs m_{rs}^x can then easily be obtained using (17). Combining (21) and (22) immediately shows that

$$N_r^E = \frac{S_r}{F_r} \left(1 + \frac{\kappa_1}{\kappa_2} \right)^{-1}, \tag{24}$$

The mass of entrants in region r therefore only depends positively on that region's size L_r and negatively on the fixed labor requirement F_r. Adding the term in r that is missing on both sides of (23), and using (22) and (24), we obtain the following equilibrium relationship:

$$\frac{S_r}{L_r} \frac{1}{\left(m_r^d\right)^{k+1}} = \sum_s S_s \tau_{rr} \left(\frac{\tau_{sr}}{\tau_{rr}} \frac{w_s}{w_r} \right)^{-k} \frac{1}{\mu_s^{\max}}. \tag{25}$$

Expressions (22) and (25)summarize how wages, upper bounds, cutoffs, trade costs and population sizes are related in general equilibrium.

3.4 Welfare

Finally, we can derive welfare in region r. This can be done as follows. Since $\mathrm{e}^{-\alpha q_{sr}(m)} = p_{sr}(m)/p_r^d$ by (13), the indirect utility in region r is given by

$$U_r = \sum_s N_s^E \int_0^{m_{sr}^x} \left[1 - \mathrm{e}^{-\alpha q_{sr}(m)}\right] \mathrm{d}G_s(m) = N_r^c \left(1 - \frac{\overline{p}_r}{p_r^d} \right).$$

Using expression (16), one can verify that $\overline{p}_r = [k/(k+1)]p_r^d + \alpha E_r / N_r^c$, which allows to express the indirect utility as follows:

$$U_r = \frac{N_r^c}{k+1} - \frac{\alpha}{\tau_{rr} m_r^d} = \frac{\alpha}{\tau_{rr}} \left[\frac{(m_r^d)^k}{(k+1)\kappa_3} \sum_s S_s \tau_{rr} \left(\frac{w_s \tau_{sr}}{w_r \tau_{rr}} \right)^{-k} \frac{1}{\mu_s^{\max}} - \frac{1}{m_r^d} \right]. \tag{26}$$

In equilibrium, expression (25) holds, so that the equilibrium indirect utility is given by

$$U_r = \frac{\alpha}{\tau_{rr} m_r^d} \left[\frac{1}{(k+1)\kappa_3} \frac{S_r}{L_r} - 1 \right]. \tag{27}$$

Thus, welfare in region r rises as the cutoff productivity level m^d falls, representing the welfare-enhancing effect of tougher firm selection. Welfare is also increasing in the (given) regional labor supply rate $\frac{S_r}{L_r}$.

4 Two regions

In this section we analyze the simplest case with two regions. In what follows, we first characterize the short-run equilibrium for given populations, which is simply an application of the general case analyzed in the previous section. Afterwards we turn to the long-run equilibrium where individuals choose their location in the first stage of the game. Following the traditions of the NEG literature we assume symmetrical regions in that part, showing under which conditions agglomeration endogenously arises starting from an ex-ante symmetrical configuration.

4.1 Short run equilibrium

Using (22)–(24), an equilibrium for given populations L_1 and L_2 can be characterized by a system of three equations with three unknowns $\omega \equiv w_1/w_2$, m_1^d and m_2^d as follows:

$$\left(\frac{w_1}{w_2}\right)^{2k+1} = \left(\frac{m_2^d}{m_1^d}\right)^{k+1} \left(\frac{\mu_2^{\max}}{\mu_1^{\max}}\right) \left(\frac{\tau_{12}}{\tau_{21}}\right)^{-k} \left(\frac{\tau_{22}}{\tau_{11}}\right)^{k+1} \left(\frac{S_1/L_1}{S_2/L_2}\right) \tag{28}$$

$$\mu_r^{\max} = L_r \tau_{rr} \left(m_r^d\right)^{k+1} + L_s \tau_{rs} \left(m_s^d \frac{\tau_{ss}}{\tau_{rs}} \frac{w_s}{w_r}\right)^{k+1}, \tag{29}$$

for $r = 1, 2$ and $s \neq r$.

Equation (29) for regions 1 and 2 can readily be solved for the cutoffs as a function of ω:

$$(m_1^d)^{k+1} = \frac{\mu_1^{\max}}{L_1 \tau_{11}} \frac{1 - \rho \omega^{-(k+1)} \left(\frac{\tau_{22}}{\tau_{12}}\right)^k}{1 - \left(\frac{\tau_{11}\tau_{22}}{\tau_{12}\tau_{21}}\right)^k} \quad \text{and} \quad (m_2^d)^{k+1} = \frac{\mu_2^{\max}}{L_2 \tau_{22}} \frac{1 - \rho^{-1} \omega^{k+1} \left(\frac{\tau_{11}}{\tau_{21}}\right)^k}{1 - \left(\frac{\tau_{22}\tau_{11}}{\tau_{21}\tau_{12}}\right)^k}, \tag{30}$$

where $\rho \equiv \mu_2^{\max}/\mu_1^{\max}$ captures relative technological possibility. Everything else being equal, a larger value of ρ implies that firms in region 2 face, upon entry, a higher probability of getting a worse draw than those in region 1. Substituting the cutoffs into the trade balance condition, and simplifying, then yields the following relationship that pins down the relative wage of the two regions:

$$\text{LHS} \equiv \omega^k = \rho \frac{S_1}{S_2} \left(\frac{\tau_{12}}{\tau_{21}} \right)^{-k} \frac{\rho \omega^{-(k+1)} \tau_{11}^{-k} - \tau_{21}^{-k}}{\tau_{22}^{-k} - \rho \omega^{-(k+1)} \tau_{12}^{-k}} \equiv \text{RHS}. \tag{31}$$

The LHS of (31) is increasing in ω, whereas the RHS is decreasing in ω on its relevant domain when trade within countries is less costly than trade between countries, i.e., $\tau_{11} < \tau_{21}$ and $\tau_{22} < \tau_{12}$. Suppose that these inequalities hold. Then, there exists a unique equilibrium such that the equilibrium relative wage ω^* is bounded by relative trade costs τ_{22}/τ_{12} and τ_{21}/τ_{11}, relative technological possibility ρ, and the shape parameter of the Pareto distribution k (see Appendix A.3).

Since the RHS of (31) is decreasing when $\tau_{11} < \tau_{21}$ and $\tau_{22} < \tau_{12}$ hold, also the comparative static results are straightforward to derive. Everything else equal, it can be shown that: (i) the larger region in terms of effective labor supply has the higher wage; (ii) the region with the better productivity support has the higher wage; (iii) higher internal transport costs in one region reduce the relative wage of that region; (iv) better access to the foreign market raises the domestic relative wage, whereas better access to the domestic market reduces the domestic relative wage; and (v) wages converge as bilateral trade barriers get smaller (see Appendix A.4).

4.2 Long-run equilibrium with symmetrical regions

Let us now consider the simplest case with two symmetrical regions. Let L denote the total population in the two-region economy, and λ is the share of individuals residing in region 1. The population size of region 2 is thus given by $(1-\lambda)L$. Furthermore assume that trade costs are symmetric ($\tau_{sr} = \tau_{rs} = \tau$), that there are no internal trade costs within a region ($\tau_{ss} = \tau_{rr} = 1$) and that regions have equal technological possibilities ($\mu_1^{\max} = \mu_2^{\max} = \mu^{\max}; \rho = 1$).

The equilibrium cutoff levels as a function of ω are then given by

$$(m_1^d)^{k+1} = \frac{\mu^{\max}}{\lambda L} \frac{1 - \tau^{-k} \omega^{-(k+1)}}{(1 - \tau^{-2k})} \quad \text{and} \quad (m_2^d)^{k+1} = \frac{\mu^{\max}}{(1-\lambda)L} \frac{1 - \tau^{-k} \omega^{(k+1)}}{(1 - \tau^{-2k})} \tag{32}$$

Furthermore, the relative wage ω is implicitly given by the following relationship

$$\omega^k = \frac{\lambda(2 - \lambda \theta L)}{(1-\lambda)(2 - (1-\lambda)\theta L)} \frac{\omega^{-(k+1)} - \tau^{-k}}{1 - \omega^{-(k+1)} \tau^{-k}} \tag{33}$$

Since we cannot solve explicitly for the equilibrium relative wage ω^* we can also not provide analytical expressions for the regional indirect utility levels U_1 and U_2 as given in (27) only in terms of the model parameters. However, we can conveniently explore the properties of the model with a numerical approach.

TABLE 1 HERE

To start, we illustrate the endogenous regional disparities for a given population distribution. In Table 1 we depict a constellation where region 1 is larger than region 2 ($\lambda = 0.6$) and has higher indirect utility, so that this situation with partial agglomeration would in fact be sustainable. As can be seen, the larger region 1 has both the higher (relative) wage and the lower cost cutoff (m_1^d).[10] More firms enter in the larger region, but as a result of tougher selection there is a lower survival probability. Still, there are in total more surviving firms in region 1. Second, there is larger consumption variety in that region, since only a subset of firms in both regions is productive enough to sell their variety also in the respective other market. Lastly, the tougher selection in region 1 leads to a lower average price for the consumers, and to a lower average markup. In other words, the larger region where economic activity has (partially) agglomerated is the more "competitive" market.
It is clear that if the larger region offers higher indirect utility (as in this example), individuals have an incentive to move to that region. More precisely, the example in Table 1 can be seen as a short-run snapshot where the population distribution is given. Now we move to the first stage, where individuals choose their location and anticipate the endogenous variables in the second stage.

FIGURE 1 HERE

In figure 1 we graphically illustrate the utility differential ($U_1 - U_2$) over the range of λ for different levels of trade costs τ and for given parameter values of θ, k, L and α. As can be seen, the symmetrical configuration ($\lambda = \frac{1}{2}$) where the two regions are exactly identical is always an equilibrium. When trade costs are low (panel 1a) this symmetrical equilibrium is unique and stable. The reason is the following: For given strength of the congestion force θ the benefits from spatial concentration are smaller the lower trade costs are. Firstly, even though the larger region always pays a wage premium this premium is smaller the lower τ is. Secondly, even though the larger region always has the lower cutoff m^d and, thus, higher average productivity, this productivity gap is also smaller the lower τ is. Put differently, at low trade costs the endogenous agglomeration force is relatively small, so that the dispersion force (commuting congestion) dominates. If trade costs are high (panel 1c) we have the opposite case that the agglomeration effect dominates the dispersion force, so that the model implies full agglomeration in either of the two regions. Finally, at intermediate trade costs (panel 1b) symmetry is locally stable, and there are two unstable equilibria with partial agglomeration and two stable equilibria with full agglomeration. In other words, our model generates a so-called "tomahawk bifurcation" similarly as in the seminal model by Krugman (1991), yet we find that agglomeration occurs at high and dispersion at low trade costs, similar as in

[10] We chose the wage in region 2 as the numeraire and normalize it to one in this constellation (($w_2 = 1$).

Helpman (1998).

FIGURE 2 HERE

In figure 2 we perform a related exercise. Here we illustrate the utility differential $(U_1 - U_2)$ for different levels of the congestion force θ and for a given value of trade costs τ (and the other parameters). It is clear that for a given strength of the agglomeration force (for given τ) agglomeration is more likely to be a stable equilibrium the weaker the congestion force is (the lower θ is). If agglomeration occurs and is stable in our model, then the individuals always earn a higher wage in the larger region and that region also features tougher firm selection.

These properties of our model can also be proven analytically. In proposition 2 we show how our model can be solved for the "break point", i.e., for the level of τ^b below which the symmetrical equilibrium $\lambda = 1/2$ is stable and above which it is unstable.

Proposition 2 *Assume that* $\tau_{11} = \tau_{22} = 1$; $\tau_{12} = \tau_{21} = \tau$; *and that* $\rho = 1$. *The unique break point* τ^b *is then given by*

$$(\tau^b)^k = \frac{(L\theta - 4)(2L\theta - 4 + k(3L\theta - 8)) + 4\kappa(k+1)(3L\theta - 4 + 4k(L\theta - 2))}{(2k+1)(L\theta - 4)(L\theta(k+2) + 4(\kappa(k+1) - 1))} \tag{34}$$

Proof. Follows from differentiating $(U_1 - U_2)$ with respect to λ taking into account the differential $d\omega/d\lambda$ that can be derived by applying the implicit function theorem on (33). Afterwards we impose symmetry ($\lambda = 1/2$) and solve the expression for τ. This yields the break point as given above. *MORE DETAILS TO BE ADDED* ■

It can be verified that the break point τ^b is increasing in θ and in L: The stronger the congestion force is, or the larger the overall population size (for given land size), the stronger is the required strength of the agglomeration force to generate agglomeration as a stable equilibrium outcome. A decrease in the commuting costs θ will, thus, have an agglomerating effect on the economy, whereas the effect of falling transport costs τ is dispersive. This result is consistent with the argument by Glaeser and Ponzetto (2007) that the "death of distance" does not necessarily make cities obsolete. The overall implications of falling spatial transaction costs for the spatial structure of the economy depends crucially on which type of cost decreases faster.

5 Conclusions

To be done

Acknowledgements. Kristian Behrens gratefully acknowledges financial support from the European Commission under the Marie Curie Fellowship MEIF-CT-2005-024266, from Université du Québec à Montréal (PAFARC), from CIRPÉE, and from FQRSC Québec. Yasusada Murata gratefully acknowledges financial support from the Japan Society for the Promotion of Science (17730165). Part of this paper was written while Kristian Behrens, Giordano Mion and Jens Südekum were visiting ARISH (Nihon University), and while Yasusada Murata and Jens Südekum were visiting CORE (Université catholique de Louvain) and UQAM. We gratefully acknowledge the hospitality of these institutions. The usual disclaimer applies.

References

[1] Antràs, P. and E. Helpman (2004) Global sourcing, *Journal of Political Economy* 112, 552-580.

[2] Aw, B., S. Chung and M. Roberts (2000) Productivity and turnover in the export market: Micro-level evidence from the Republic of Korea and Taiwan (China), *World Bank Economic Review* 14, 65-90.

[3] Axtell, R.L. (2001) Zipf distribution of U.S. firm sizes, *Science* 293, 1818-1820.

[4] Baldwin, R. and T. Okubo (2006) Heterogeneous firms, agglomeration and economic geography: spatial selection and sorting, *Journal of Economic Geography* 6, 323-346.

[5] Behrens, K. and Y. Murata (2007) General equilibrium models of monopolistic competition: a new approach, *Journal of Economic Theory* 136, 776-787.

[6] Behrens, K., G. Mion, Y. Murata and J. Suedekum(2008) Trade, wages and productivity, *IZA Discussion Papers* 3682, Bonn.

[7] Bernard, A.B. and J.B. Jensen (1999) Exceptional exporter performance: cause, effect, or both? *Journal of International Economics* 47, 1-25.

[8] Bernard, A.B., J. Eaton, J.B. Jensen and S. Kortum (2003) Plants and productivity in international trade, *American Economic Review* 93, 1268-1290.

[9] Bernard, A.B., J.B. Jensen, S.J. Redding and P.K. Schott (2008) Firm in international trade, *Journal of Economic Perspectives*, forthcoming.

[10] Bernard, A.B., S.J. Redding and P.K. Schott (2007) Comparative advantage and heterogeneous firms, *Review of Economic Studies* 73, 31-66.

[11] Corless, R.M., G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey and D.E. Knuth (1996) On the Lambert W function, *Advances in Computational Mathematics* 5, 329-359.

[12] Glaeser, E. and D.Mare (2001) Cities and skills *Journal of Labor Economics* 19, 316-342.

[13] Glaeser, E., G.Ponzetto (2007) Did the death of distance hurt Detroit and help New York? Mimeo, Harvard University.

[14] Head, K. and T.Mayer (2006) Regional Wage and Employment Responses to Market Potential in the EU, *Regional Science and Urban Economics* 36, 573-595.

[15] Helpman, E. (1998) The size of regions, In: D. Pines, E. Sadka, I. Zilcha (eds.), *Topics in public economics. Theoretical and empirical analysis*, Cambridge University Press, pp. 33-54.

[16] Helpman, E. (2006) Trade, FDI, and the organization of firms, *Journal of Economic Literature* 44, 589-630.

[17] Helpman, E., M.J. Melitz and S.R. Yeaple (2004) Export versus FDI with heterogeneous firms, *American Economic Review* 94, 300-316.

[18] Krugman, P.R. (1979) Increasing returns, monopolistic competition, and international trade *Journal of International Economics* 9, 469-479.

[19] Krugman, P.R. (1980) Scale economies, product differentiation and the pattern of trade, *American Economic Review* 70, 950-959.

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

[21] Murata, Y. and J.Thisse (2005) A simple model of economic geography la Helpman-Tabuchi, *Journal of Urban Economics* 58, 137-155.

[22] Melitz, M.J. (2003) The impact of trade on intra-industry reallocations and aggregate industry productivity, *Econometrica* 71, 1695-1725.

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

[24] Okubo, T. (2006) Trade liberalisation and agglomeration with firm heterogeneity - forward and backward linkages, Working Paper 16-2006, Graduate Institute of International Studies, Geneva.

[25] Ottaviano, G., T. Tabuchi and J. Thisse (2002) Agglomeration and Trade revisited, *International Economic Review* 43, 409-435.

[26] Pavcnik, N. (2002) Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants, *Review of Economic Studies* 69, 245-276.

[27] Syverson, C. (2004) Market structure and productivity: a concrete example, *Journal of Political Economy* 112, 1181-1222.

Appendix A: Proofs

A.1. Existence and uniqueness of the equilibrium cutoff m^d. We show that there exists a unique equilibrium cutoff m^d. To see this, applying the Leibniz integral rule to the left-hand side of (10) and using $W(\mathrm{e}) = 1$ to obtain

$$\frac{\mathrm{e}L}{\alpha(m^d)^2}\int_0^{m^d} m^2\left(W_m^{-2}-1\right)W_m'\mathrm{d}G(m) > 0,$$

where the sign comes from $W_m' > 0$ and $W_m^{-2} \geq 1$ for $0 \leq m \leq m^d$. Hence, the left-hand side of (10) is strictly increasing, which uniquely determines the equilibrium cutoff m^d because

$$\lim_{m^d\to 0}\int_0^{m^d} m\left(W_m^{-1}+W_m-2\right)\mathrm{d}G(m) = 0 \quad \text{and} \quad \lim_{m^d\to \infty}\int_0^{m^d} m\left(W_m^{-1}+W_m-2\right)\mathrm{d}G(m) = \infty. \tag{35}$$

A.2. Market size, the equilibrium cutoff and the mass of entrants. Differentiating (10) and using the Leibniz integral rule, we readily obtain

$$\frac{\partial m^d}{\partial L} = -\frac{\alpha F\left(m^d\right)^2}{\mathrm{e}L^2}\left[\int_0^{m^d} m^2\left(W_m^{-2}-1\right)W_m'\mathrm{d}G(m)\right]^{-1} < 0$$

because $W_m' > 0$ and $W_m^{-2} \geq 1$ for $0 \leq m \leq m^d$. Differentiating (11) with respect to L yields

$$\frac{\partial N^E}{\partial L} = \frac{F(N^E)^2}{L^2}\left\{1-\frac{\mathrm{e}L^3}{\alpha F\left(m^d\right)^2}\left[\int_0^{m^d} m^2 W_m'\mathrm{d}G(m)\right]\frac{\partial m^d}{\partial L}\right\} > 0,$$

where the sign comes from $\partial m^d/\partial L < 0$ as established in the foregoing.

A.3. Existence and uniqueness in the two-region case. Under our assumptions on trade costs, the RHS of (31) is non-negative if and only if $\underline{\omega} < \omega < \overline{\omega}$, where $\underline{\omega} \equiv (\tau_{22}/\tau_{12})^{k/(k+1)}\rho^{1/(k+1)}$ and $\overline{\omega} \equiv (\tau_{21}/\tau_{11})^{k/(k+1)}\rho^{1/(k+1)}$. Furthermore, the RHS is strictly decreasing in $\underline{\omega} < \omega < \overline{\omega}$ with $\lim_{\omega\to\underline{\omega}+}$ RHS $= \infty$ and $\lim_{\omega\to\overline{\omega}-}$ RHS $= 0$. The LHS of (31) is, on the contrary, strictly increasing in ω from 0 to ∞. Hence, there exists a unique equilibrium $\underline{\omega} < \omega^* < \overline{\omega}$.

A.4. Market size, trade frictions, and wages. (i) First, ω^* is increasing in L_1/L_2 as an increase in L_1/L_2 raises the RHS of (31) without affecting the LHS. This implies that if the two regions have equal productivity supports ($\rho = 1$) and face symmetric trade costs ($\tau_{12} = \tau_{21}$ and $\tau_{11} = \tau_{22}$), the larger region has the higher relative wage. (ii) Since $(\tau_{11}\tau_{22})^k < (\tau_{12}\tau_{21})^k$ holds by assumption, the RHS of (31) shifts up as ρ increases, which then also increases ω^*. This implies that if the two regions are of equal size ($L_1 = L_2$) and face symmetric trade costs ($\tau_{12} = \tau_{21}$ and $\tau_{11} = \tau_{22}$), the region with the better productivity support has the higher wage. (iii) Higher internal transport costs in one region reduce the relative wage of that region because

$$\frac{\partial(\text{RHS})}{\partial\tau_{11}} < 0 \quad \text{iff} \quad \omega^* > \underline{\omega} \quad \text{and} \quad \frac{\partial(\text{RHS})}{\partial\tau_{22}} > 0 \quad \text{iff} \quad \omega^* < \overline{\omega}.$$

(iv) Better access to the foreign market raises the domestic relative wage, whereas better access to the domestic market reduces the domestic relative wage because

$$\frac{\partial(\text{RHS})}{\partial\tau_{12}} < 0 \quad \text{iff} \quad \omega^* < \overline{\omega} \quad \text{and} \quad \frac{\partial(\text{RHS})}{\partial\tau_{21}} > 0 \quad \text{iff} \quad \omega^* > \underline{\omega}.$$

(v) Assuming that $\tau_{12} = \tau_{21} = \tau$ and that $\tau_{11} = \tau_{22} = t$, one can then verify that

$$\frac{\partial(\text{RHS})}{\partial\tau} = -\frac{k\rho}{\tau}\frac{L_1}{L_2}\left(\frac{t}{\tau}\right)^k \frac{\rho^2 - \omega^{2(k+1)}}{[\omega^{k+1} - \rho(t/\tau)^k]^2} \left\{\begin{array}{c} > \\ = \\ < \end{array}\right\} 0 \quad \text{for} \quad \left\{\begin{array}{c} \underline{\omega} < \rho^{\frac{1}{k+1}} < \omega^* < \overline{\omega} \\ \underline{\omega} < \omega^* = \rho^{\frac{1}{k+1}} < \overline{\omega} \\ \underline{\omega} < \omega^* < \rho^{\frac{1}{k+1}} < \overline{\omega} \end{array}\right\}. \quad (36)$$

Note that when regions are of equal size, but have different upper bounds ($\rho > 1$), the first case of (36) applies since $\omega^* > \rho^{1/(k+1)}$ in equilibrium. To see this, evaluate (31) at $\omega = \rho^{1/(k+1)}$ and recall that $\tau_{21} = \tau_{12} = \tau$ and $L_1 = L_2$. The LHS is equal to $\rho^{k/(k+1)}$, which falls short of the RHS given by ρ (since $\rho > 1$ and $k \geq 1$). Since the LHS is increasing and the RHS is decreasing, it must be that $\omega^* > \rho^{1/(k+1)}$. Hence, lower trade costs reduce the relative wage of the more productive region. Furthermore, assuming that regions have the same upper bounds but different sizes ($L_1 > L_2$), we obtain $\omega^* > \rho^{k/(k+1)} = 1$, so that first case of (36) applies again.

Appendix B: Integration of Lambert W functions

To derive closed-form solutions for various expressions throughout the paper requires us to compute integrals involving the Lambert W function. This can be done by using the change in variables suggested by Corless *et al.* (1996, p.341). Let

$$z \equiv W\left(\mathrm{e}\frac{m}{I}\right), \quad \text{so that} \quad \mathrm{e}\frac{m}{I} = z\mathrm{e}^z, \quad \text{where} \quad I = m_r^d, m_{rs}^x,$$

where r is in autarky. The change in variables then yields $\mathrm{d}m = (1+z)\mathrm{e}^{z-1}I\mathrm{d}z$, with the new integration bounds given by 0 and 1. Under our assumption of a Pareto distribution for productivity draws, the change in variables allows to rewrite many integrals in simplified form.

B.1. First, consider the following expression, which appear when integrating firms' outputs:

$$\int_0^I m\left[1-W\left(\mathrm{e}\frac{m}{I}\right)\right]\mathrm{d}G_r(m)=\kappa_1\left(m_r^{\max}\right)^{-k}I^{k+1},$$

where $\kappa_1\equiv k\mathrm{e}^{-(k+1)}\int_0^1\left(z\mathrm{e}^z\right)^k(1-z^2)\mathrm{e}^z\mathrm{d}z>0$ is a constant term which solely depends on the shape parameter k.

B.2. Second, the following expression appears when integrating firms' operating profits:

$$\int_0^I m\left[W\left(\mathrm{e}\frac{m}{I}\right)^{-1}+W\left(\mathrm{e}\frac{m}{I}\right)-2\right]\mathrm{d}G_r(m)=\kappa_2\left(m_r^{\max}\right)^{-k}I^{k+1}$$

where $\kappa_2\equiv k\mathrm{e}^{-(k+1)}\int_0^1\left(z\mathrm{e}^z\right)^k(1+z)\left(z^{-1}+z-2\right)\mathrm{e}^z\mathrm{d}z>0$ is also a constant term which solely depends on the shape parameter k.

B.3. Finally, the following expression appears when integrating firms' sales revenues:

$$\int_0^I m\left[W\left(\mathrm{e}\frac{m}{I}\right)^{-1}-1\right]\mathrm{d}G_r(m)=\kappa_3\left(m_r^{\max}\right)^{-k}I^{k+1},$$

where $\kappa_3\equiv k\mathrm{e}^{-(1+k)}\int_0^1\left(z\mathrm{e}^z\right)^k\left(z^{-1}-z\right)\mathrm{e}^z\mathrm{d}z$ is a constant term which is solely parametrized by the shape parameter k. Using the expressions of κ_1 and κ_2, one can verify that $\kappa_3\equiv\kappa_1+\kappa_2$.

Appendix C: Equilibrium in the open economy

Plugging (18) into (19), zero expected profits require that

$$\frac{1}{\alpha}\sum_s L_s\tau_{rs}\int_0^{m_{rs}^x} m\left[W\left(\mathrm{e}\frac{m}{m_{rs}^x}\right)^{-1}+W\left(\mathrm{e}\frac{m}{m_{rs}^x}\right)-2\right]\mathrm{d}G_r(m)=F_r. \tag{37}$$

As in the closed economy case, condition (37) depends solely on the cutoffs m_{rs}^x and is independent of the mass of entrants. Furthermore, using (18), the labor market clearing condition can be rewritten as follows:

$$N_r^E\left\{\frac{1}{\alpha}\sum_s L_s\tau_{rs}\int_0^{m_{rs}^x} m\left[1-W\left(\mathrm{e}\frac{m}{m_{rs}^x}\right)\right]\mathrm{d}G_r(m)+F_r\right\}=L_r. \tag{38}$$

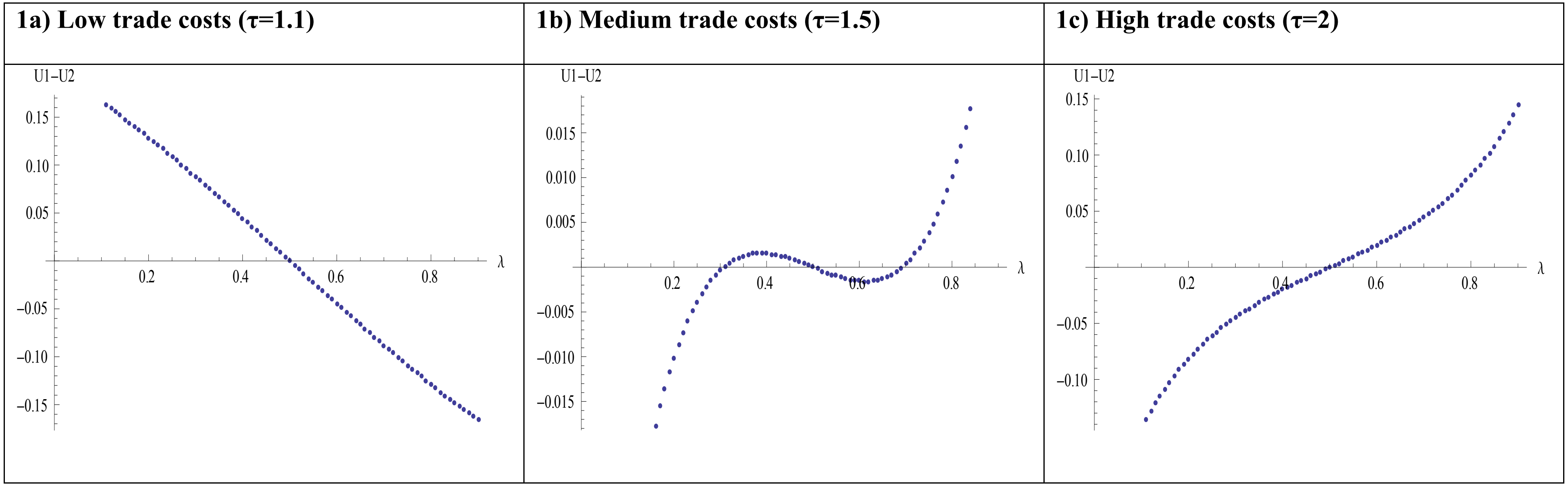
The equilibrium system given by equations (17), (37) and (38) has $K(K+2)$ unknowns (the K^2 cutoffs, K wages and K masses of entrants) and $K(K+1)$ equations (the $K(K-1)$ links between the domestic and the export cutoffs, K labor market clearing conditions, and

K zero expected profits conditions). To solve for the equilibrium requires to impose K trade balance conditions, which can be rewritten as follows:

$$N_r^E w_r \sum_{s\neq r} L_s \tau_{rs} \int_0^{m_{rs}^x} m \left[W\left(\mathrm{e}\, \frac{m}{m_{rs}^x} \right)^{-1} - 1 \right] \mathrm{d}G_r(m)$$
$$= L_r \sum_{s\neq r} N_s^E \tau_{sr} w_s \int_0^{m_{sr}^x} m \left[W\left(\mathrm{e}\, \frac{m}{m_{sr}^x} \right)^{-1} - 1 \right] \mathrm{d}G_s(m). \quad (39)$$

Using the region-specific Pareto distributions $G_r(m) = (m/m_r^{\max})^k$ yields after some algebra the expressions given in the main text.

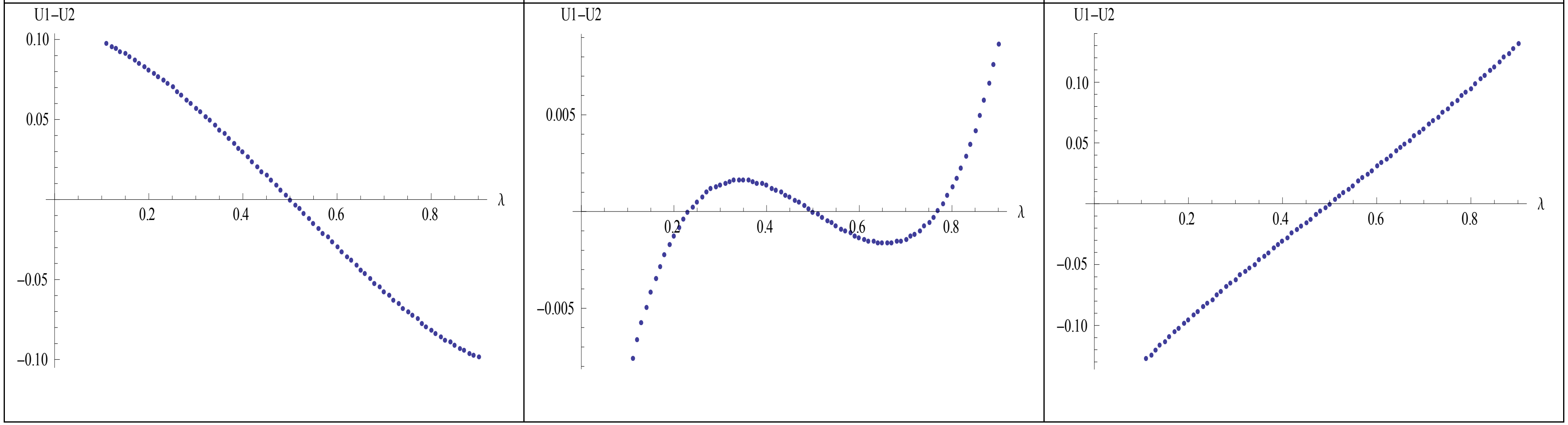
Figure 1: Regional utility differential for varying trade costs τ



parameter values: θ=0.035, L=10, α=1, k=2

Figure 2: Regional utility differential for varying congestion strength θ

1a) Strong congestion (θ=0.035)	1b) Medium congestion (θ=0.021)	1c) Weak congestion (θ=0.005)



parameter values: τ=1.2, L=10, α=1, k=2

Table 1: Characteristics of the larger region in case of endogenous agglomeration – An short-run example

Parameter constellation: $\tau = 3, \theta = 0.035, k = 3, L = 10, m_{MAX} = 1, \alpha = 1, F = 0.5$

Structure of the economy: $\lambda = 0.6 \quad \rightarrow \quad S_1 = 5.73, S_2 = 3.88 \qquad U_1 - U_2 = 0.078 > 0 \quad \rightarrow$ *agglomeration*!

Variable	**Region 1**	**Region 2**	**Results:** **Larger region has…**
Wage (using $w_2 = 1$)	$w_1 = 1.054$	$w_2 = 1$	… higher wage
Cutoff cost level (cutoff local price)	$m_1^d = 0.862$ $(p_1^d = 0.818)$	$m_2^d = 0.951$ $(p_2^d = 0.951)$	… tougher selection
# Entrants	$N_1^E = 14.32$	$N_2^E = 9.7$	… more firm entry
# Surviving firms (Survival probability)	$N_1^P = 9.4$ $(N_1^P / N_1^E = 0.65)$	$N_2^P = 8.6$ $(N_2^P / N_2^E = 0.89)$	… more surviving firms but lower survival probability
# Consumed varieties	$N_1^C = 9.47$	$N_1^C = 8.73$	… more consumption variety
Av. consumer price	$\bar{p}_1 = 0.719$	$\bar{p}_2 = 0.824$	… lower average price for consumers
Av. markup	$\alpha E_1 / N_1^C = 0.106$	$\alpha E_2 / N_2^C = 0.111$	… lower average mark-up