Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Urban Economics 59 (2006) 300–316
JOURNAL OF Urban Economics
www.elsevier.com/locate/jue

Timeliness and agglomeration ☆

James Harrigan [a,b], Anthony J. Venables [c,d,*]

[a] *International Research Department, Federal Reserve Bank of New York, New York, NY 10045, USA*
[b] *NBER*
[c] *Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, UK*
[d] *CEPR*

Received 15 November 2004; revised 6 October 2005

Available online 2 February 2006

Abstract

An important element of the cost of distance is time taken in delivering final and intermediate goods. We argue that time costs are qualitatively different from direct monetary costs such as freight charges. The difference arises because of uncertainty. Unsynchronised deliveries can disrupt production, and delivery time can force producers to order components before demand and cost uncertainties are resolved. Using several related models we show that this generates a hitherto unexplored mechanism for clustering. If final assembly takes place in two locations and component production has increasing returns to scale, then component production will tend to cluster around just one of the assembly plants.

JEL classification: F1; L0

Keywords: Just-in-time; Clustering; Location; Trade

☆ Produced as part of the Globalisation Programme of the UK ESRC funded Centre for Economic Performance at the LSE. The views expressed in this paper are those of the authors and do not necessarily reflect the position of the Federal Reserve Bank of New York or the Federal Reserve System.

* Corresponding author.
E-mail addresses: james.harrigan@ny.frb.org (J. Harrigan), a.j.venables@lse.ac.uk (A.J. Venables).
URLs: http://www.newyorkfed.org/research/economists/harrigan/index.html (J. Harrigan), http://econ.lse.ac.uk/staff/ajv/ (A.J. Venables).

0094-1190/$ – see front matter
doi:10.1016/j.jue.2005.10.004

1. Introduction

People pay a lot of money to save time. A modern economy is inconceivable without air travel and air shipment, which are ways of saving time at the expense of money. For workers in urban areas the main component of commuting costs is time. For international trade in manufactured goods estimates of the costs of the time-in-transit range as high as 0.5% of the value of goods shipped, *per day* (Hummels [10]), while Anderson and van Wincoop [1] suggest that time in transit amounts to an average 9% tariff equivalent for US trade. Protagonists of 'just-in-time' manufacturing techniques emphasise the importance of organising and locating production to ensure timely delivery of parts and components.

Surprisingly, these observations have had little impact on the economic analysis of location decisions.[1] Economists have worked with an aggregate of 'transport costs' or 'trade costs' to capture the penalty of distance, while simply remarking that these costs are a shorthand for a wide range of penalties (e.g. Fujita et al. [7]). These include freight and other monetary transaction costs; lack of information about markets and suppliers and about local institutions and regulations; difficulty in monitoring contracts; the impossibility of face-to-face contact and communication; and the fact that distance introduces delay into completion of trades. It is unlikely that summarising these penalties as a single value of 'trade costs' is adequate for understanding their effects. The objective of this paper is to contribute to the process of unpacking and evaluating the different elements of these costs.[2]

We focus on the costs associated with delivery times and argue that timeliness is not only a quantitatively important aspect of proximity, but also matters qualitatively, creating an incentive for clustering of activities. The context is the time taken between initiating a project and completing it and making delivery to the consumer. When the stages of the value chain (e.g. component manufacture and final assembly, or production and final sale) are physically separated it takes more time to complete a project, and we argue that delay matters for several reasons.

One reason is discounting and analogous factors, such as the physical depreciation or technical obsolescence that component parts may be subject to during shipment. These costs will not be the focus of our attention, although they may be large—computer chips become obsolescent very rapidly, and fish gets smelly after a few days. Other reasons why delay matters are intimately connected with uncertainty. One set of arguments has to do with the synchronisation of activities; production cannot be completed until all the parts have arrived, so uncertain arrival times of components can have a cost that is quite disproportionate to the cost of any single component. Other arguments arise since uncertainty—about demand or costs—makes it profitable to postpone production until as much uncertainty as possible has been resolved. Delivery delay brings forward the date at which production decisions have to be made, orders placed, and expenditures incurred, thereby increasing the uncertainty borne by a firm.

Of course, saving time is always going to be beneficial, just as is saving freight charges. To make the point that there may be qualitative (as well as a quantitative) implications of timeliness, we develop all our models in a very particular framework that enables us to assess the profitability of clustering activities together. The framework is one in which there are two locations, each of which has an assembly plant or retailer supplying local final demand. The assembly process uses

[1] For instance, Fujita and Thisse's [7] lucid and authoritative recent book does not have "time" in the index. There is some modelling of the issues in Evans and Harrigan [5], Harrigan [8], and Venables [19].

[2] Previous attempts to evaluate non-monetary trade costs include study of the benefits of face-to-face contact, see Leamer and Storper [16] and Storper and Venables [17].

a number of component parts, and increasing returns in production of each of these components are sufficiently great that each is produced in a single plant. Where do the component producers locate? Clustered around one of the assembly plants, or divided between the two locations? We show that the demand for timeliness in delivery creates a force for clustering of plants around a single assembler/retailer. This is a previously unexplored agglomeration mechanism.

We develop this argument in a series of models. Section 4 outlines a benchmark case in which there are monetary trade costs, but delivery is instantaneous and component producers consequently have no incentive to cluster. In Section 5 we look at the issues raised by the synchronisation of delivery of components, and show that uncertain delivery times will cause clustering of component producers. Section 6 shows how uncertainty about demand and/or about production costs are also forces for clustering. Before developing these models we briefly connect our approach to the extensive management literature on just-in-time (JIT) production.

2. Just-in-time

Although timely delivery has received little attention in the economics literature, it has been central to just-in-time (JIT) management techniques. The JIT approach was pioneered by Toyota Motors in the 1950s. Its main features are that components are delivered in small but frequent batches, that minimal stocks are held, and that 'quality control is built in.' The perceived advantages are a reduction in the cost of holding stock, rapid response to customer orders, and the ability to rapidly detect and fix or replace defective components. Effective implementation of JIT is thought to require long term supplier/customer relationships and, where possible, proximity.

In the management literature on just-in-time production it has been suggested that the spread of JIT systems might lead to a geographical reconcentration of supplier firms and customers (e.g. Dicken [4]). The importance of proximity is illustrated by the example of General Electric's appliances division in their attempt to implement JIT in the 1980s and 1990s. General Electric was hampered by the fact that some suppliers were several thousand kilometres away from General Electric plants, this causing a 1993 decision to increase inventory levels (Jones, George and Hill [11]). The US auto-industry has been extensively studied, although identifying the effects of JIT on supplier location is a tricky empirical question. Assemblers tend to locate where suppliers are already located, and in addition there are non-JIT reasons why suppliers may want to be near assemblers (such as minimizing transport costs irrespective of timeliness considerations). Klier [12] assembles a comprehensive dataset on assemblers and suppliers and shows that, since the advent of JIT, new supplier plants are more likely to locate near their assembly plant customers than they were before the advent of JIT. Klier also finds that proximity generally means "within a days drive," rather than right next door, which implies that the agglomeration force of JIT operates at the regional rather than the urban level. This is consistent with the results of Rosenthal and Strange [14], who find "... shipping-oriented attributes (manufactured inputs, resources, perishability) influencing agglomeration at the state level ..." (p. 193).

Our goal in this paper is to develop some simple models that capture some of the features referred to in this literature. As we will show, these features create forces for the spatial concentration of activity.

3. A family of models

We develop our ideas in a family of models, each based on two a priori identical locations A and B, where final assembly occurs and demand is met. The two locations could be separate

cities, regions or countries. A key assumption is that the final assembled product is non-tradeable, so assembly must be undertaken in both places. The purpose of this unrealistic assumption is to clarify how our framework differs from existing models of agglomeration incentives, and the message of the paper is unaffected by dispensing with it. We refer to the final stage as assembly but the idea is more general: "assemblers" could be service firms who require a variety of manufactured or service inputs, or retailers who sell a variety of products.

Assembly in A and B has constant returns to scale and uses labour and a fixed number N of types of components. Components are tradeable, although trade typically takes time.

Production of each type of component incurs a plant level fixed cost and then has constant marginal cost, and we assume that the fixed cost is large enough to ensure that each component is only produced in one location, either A or B. Our primary question is to ask where production of each type of component takes place.

The objective is to maximise the combined profit of assemblers and component producers. This objective can be rationalised either by assuming that a single firm controls all activities, or by assuming that both assembly plants are controlled by one firm and all components produced by another. In this case Nash bargaining by these two parties over the location of components' production and division of the surplus would lead to the efficient (joint profit maximising) outcome. It would be interesting to look at more general non-cooperative outcomes. However, in most of the models that we develop all components are necessary to production of the final product, raising the question of how surplus is split between independent assemblers and component producers. The theory of (non-cooperative) bargaining offers no answer to this when there is more than one supplier (see Sutton [18], Binmore and Dasgupta [2]). However, it does lead us to expect that the outcome will be efficient, maximising the combined returns to all parties.

Table 1 presents an overview of the models we develop. Columns of the table correspond to the order in which decisions are taken. In all cases we seek to determine, at the first stage, the number of component producers in each location, N_A and N_B, with $N_A + N_B = N$. Production requires inputs of each component, and x_A, x_B denote chosen input quantities; all components enter production symmetrically, so we only distinguish inputs according to whether they are produced in A or B. In our first model there is no uncertainty, so quantities of inputs are chosen and then production and sales take place. In the second model all input decisions are taken at the same time but there is uncertainty concerning the delivery time for components from remote suppliers. In the final model there is uncertainty about the level of final demand or assembly costs. Timing is such that decisions concerning locally produced inputs can be taken after uncertainty is resolved. However, decisions about remote inputs have to be taken under uncertainty. Thus, for example, the assembler in A has to choose quantities x_B of inputs from each of the N_B remote suppliers before uncertainty is revealed, while quantities x_A from local suppliers can be chosen later (and conversely for the assembler in B).

Table 1
Timing in a family of models

Model	Input quantities	Nature	Input quantities
Timeless (Section 4)	x_A, x_B	No uncertainty	
Synchronisation (Section 5)	x_A, x_B	Late parts (if any) from B	
Demand or cost uncertainty (Section 6)	x_B	Level of demand/cost	x_A

Notes. The table summarises key features of the models developed in the paper. Time runs from left to right, so reading across a row gives the sequence of decisions and information in each model, from the standpoint of the assembler in A.

We will show how the presence of these uncertainties can make it efficient to cluster all component producers around one of the assembly plants, and in equilibrium all component producers will do so (without loss of generality, suppose that the cluster occurs at location A). The economic reason is that it is better to concentrate the costs associated with uncertainty on a single assembler than for both assemblers to incur them. An implication is that the assembly plant in A that is blessed with a nearby cluster of suppliers will have a competitive advantage over the assembler in B. If there were low-cost trade in final goods A's advantage might lead to the assembler in B shutting down. In such a case there would be complete clustering of intermediate and final goods production in A, and the assembler in A would supply customers in both A and B. Such clusters are quite familiar from standard economic geography models. By assuming that final goods are non-tradeable we shut down this standard mechanism for agglomeration, thereby highlighting what is novel in our framework.

4. A timeless benchmark model

Before turning to models with uncertainty, we look briefly at the benchmark 'timeless' case, based on ingredients from a standard economic geography model. In this benchmark model assemblers in A and B each produce a unit of output using N symmetric inputs in a CES production function with elasticity of substitution σ. The value of producing one unit of final output in location A is p, the exogenously given price of final output net of any assembly costs, minus the costs of producing and shipping components,

$$V_A = p - \left[N_A r_A^{1-\sigma} + N_B(\tau r_B)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}. \tag{1}$$

The cost function is expressed with inputs divided between the N_A components sourced locally with unit production costs r_A, and the remaining N_B that come from B, the 'remote' location, with unit cost r_B and shipping cost factor $\tau > 1$. Notice that, since we are looking for efficient outcomes, we use the unit production costs of components, r_i, which may not be the same as the prices at which they are traded. Furthermore, we will henceforth refer to V_A as the profits of assembly in A, noting that it is both the profits of the assembler and profits (before fixed costs) earned by component producers on supply of parts to A. A similar equation gives profits in B.

What values of N_A and N_B maximise the total profit of assemblers in A and B, $V_A + V_B$? The total number of component suppliers is fixed at N, so that $N_A = N - N_B$, and we let input costs be the same in each location. Making these substitutions in (1) and taking the derivative of V_A with respect to N_B gives

$$\frac{\partial V_A}{\partial N_B} = \frac{r(\tau^{1-\sigma}-1)}{\sigma-1}\left[N + N_B\left(\tau^{1-\sigma}-1\right)\right]^{\frac{\sigma}{1-\sigma}} < 0, \tag{2}$$

$$\frac{\partial^2 V_A}{\partial N_B^2} = \frac{-r\sigma(\tau^{1-\sigma}-1)^2}{(1-\sigma)^2}\left[N + N_B\left(\tau^{1-\sigma}-1\right)\right]^{\frac{1}{\sigma-1}} < 0. \tag{3}$$

These derivatives establish that V_A is decreasing and concave in N_B: shifting assembler locations from A to B has an increasingly negative effect on the returns to assembling in A. The opposite is true for the returns from assembly in B. The point here is that the *increasing* marginal cost of remote suppliers implies that the sum $V_A + V_B$ is maximised when half of the suppliers locate in each region. Equivalently, there is a decreasing marginal value to proximity. The point is illustrated in Fig. 1, in which the number of component suppliers located in B is on the horizontal

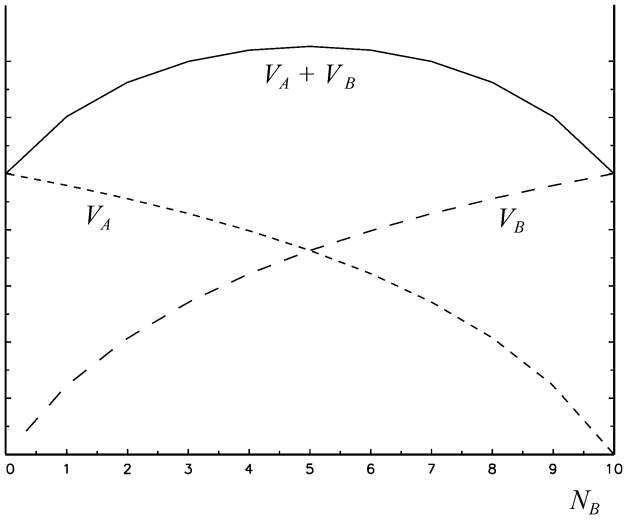

Fig. 1. CES assembly ($r_A = r_B = 1$, $\sigma = 5$, $p = 1.5$, $\tau = 1.5$, $N = 10$).

axis.[3] Curves give profits in each place, and their sum, $V_A + V_B$, is maximised when $N_A = N_B$. A lower elasticity of substitution, σ, gives less curvature and a flatter $V_A + V_B$ schedule, but only in the limit, when σ is zero, do the curves become linear, and their sum horizontal.

This result does not turn on a CES cost function. Quite generally, if the assembler did not adjust its input quantities as N_A and N_B changed, then V_A would be the straight line between the values of V_A at $N_A = 0$ and $N_B = 0$. The possibility of adjustment means that V_A lies on or above this straight line, giving the convexity illustrated. (A formal proof is given in Appendix A.)

We conclude that in this benchmark case there is no incentive for clustering of suppliers. In fact the opposite is true: if there is any substitutability in inputs, then the efficient distribution of suppliers is to have half located in each assembly location (with zero substitutability, any division of suppliers is equally efficient). Finally, note that if final goods are tradeable at moderate cost then the benchmark becomes a standard "core–periphery" economic geography model, and there will be a tendency for all production of components and final goods to be clustered in a single location (see, for example, of Fujita et al. [6, Chapter 14]). With this benchmark in mind we now turn to models where remote supply takes time.

5. Synchronisation

Our first model of timeliness turns on uncertainty about delivery time, and the consequent risk that production may be delayed by the late arrival of components from a distant supplier. We model this by supposing that each assembly firm seeks to produce a unit of output for delivery at a particular date. Assembly uses labour to combine N different component parts into final output using a Leontief production function with unit coefficients. Of course, production cannot

[3] This and other figures are generated by simulation of the models. Parameter values are given in every figure.

be completed until all the parts needed have arrived.[4] For the moment, we assume that holding stocks of components is infeasible or prohibitively costly. This might be because of very high storage or depreciation costs, or simply because the exact specification of the product is unknown prior to the decision to produce.

Transport of components between locations is costless, but timely delivery of parts can only be guaranteed if the assembler and parts supplier are located in the same region. The probability of timely delivery is $q < 1$ if supplier and assembler are located in different regions. Assuming that delivery of each part is i.i.d. across suppliers and assemblers, for assemblers located in A,

$$\Pr(\textit{all parts arrive on time}) = q^{N_B},$$
$$\Pr(\textit{at least one part arrives late}) = 1 - q^{N_B},$$

where as before N_B is the number of parts suppliers located in B, $N_A + N_B = N$. Clearly, Pr(*all parts arrive on time*) is decreasing and (importantly, as it turns out) convex in N_B:

$$\frac{\partial q^{N_B}}{\partial N_B} = q^{N_B}\ln q < 0, \qquad \frac{\partial^2 q^{N_B}}{\partial N_B^2} = q^{N_B}[\ln q]^2 > 0. \tag{4}$$

This means that each part which changes from being supplied locally to remotely decreases the probability that all parts arrive on time, but does so at a diminishing rate. The intuition for this is straightforward: if one part is delayed, it does not matter if a second part is also delayed.

There are several reasons why delays in completing assembly might be bad for profits. One is demand decay. Many goods and services have demand which peaks at a certain time and the price that the assembler can get for the final product falls unless it is delivered on time. Another is that some assembly costs have to be met whether production occurs or not. For example, if labour must be hired to assemble parts, then wages must be paid regardless of whether all parts have arrived. Think of labour as a cost incurred before the outcome of the delivery process is known, so that if there are delays, labour must be hired again once all parts arrive.

To capture these arguments, let final demand be characterised by a reservation price which is p on the day that demand is realized and $p(1-\delta)$ one day later, $\delta \in (0,1)$. Profits if all parts are delivered on time are therefore

$$v_A^0 = p - \beta w_A - N_A r_A - N_B r_B \tag{5}$$

where β is the daily unit labour requirement for assembly and w_A is the wage. If parts are delivered one day late, the reservation price falls and labour must again be hired, so profits are

$$v_A^1 = p(1-\delta) - 2\beta w_A - N_A r_A - N_B r_B. \tag{6}$$

The difference between profits on day 0 and on day 1, $\delta p + \beta w_A$, is the penalty paid by firms who suffer late delivery of parts. Expected profits are just profits if there is no delay minus the expected cost of delay,

$$V_A = v_A^0 - \left(q - q_B^N\right)(\delta p + \beta w_A). \tag{7}$$

If there are no cost differences between the two locations, then (5) and (7) imply that expected profits in A are decreasing and convex in N_B: the hit to expected profits of sourcing an additional part from far away gets smaller as the number of them increases.

[4] This production function is formally quite similar to Kremer's [13] O-ring technology. Our results go through with more general technologies providing the elasticity of substitution between parts is less than unity, so each is necessary for production.

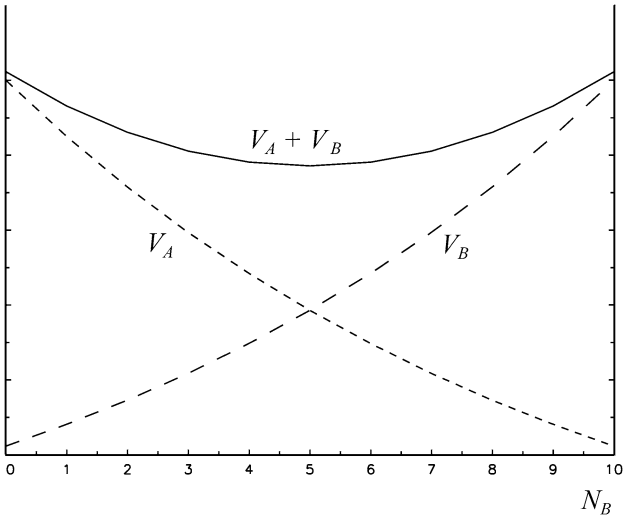

Fig. 2. Arrival uncertainty ($v_A^0 = 1$, $\delta p + \beta w_A = 1.5$, $q = 0.9$).

Symmetric results apply to expected profits in B, which has the important implication that total expected profits are maximised at $N_B = 0$ and at $N_B = N$. This is illustrated in Fig. 2. In contrast to the baseline model of the previous section, total expected profits are *minimized* at $N_B = N_A = N/2$: with such a division of production, neither suppliers in A nor in B get the benefit of reliable deliveries. This illustrates the increasing marginal value of timeliness: if almost all parts have guaranteed on-time delivery, an increase in share of timely parts has a bigger effect on expected profits than if most parts are already subject to erratic delays. As a result, there is an economic incentive for agglomeration of all suppliers in either A or B.

The increasing marginal value of timeliness just illustrated, and the associated agglomeration incentives, have a close analogy with Kremer's [13] O-ring model. In Kremer's model, there is an increasing marginal value of quality, and so it is efficient for all the high-quality workers to work together.

The point of this simple case is that although the locations are *ex ante* symmetric, the efficient location of component producers is asymmetric. It is best to have one assembler operating in a cluster of all the component suppliers and producing without delay, while the other bears the full cost of the uncertainties associated with delivery delay.

The difference between locations also shows up as a productivity difference. One of the key facts about agglomeration is that localized industries have higher measured productivity (see Rosenthal and Strange [15] for a review of the evidence). The model offers an explanation for this: localized activities benefit from timeliness, which reduces or eliminates periods when production is interrupted by delayed delivery. If all suppliers locate in A, then assemblers in A never have to pay labour twice, while assemblers in B have to pay labour a second time with probability $1 - q^N$. Since output is the same in each location, relative productivity in A is given by the ratio of expected unit costs:

$$TFP_{AB} = \frac{(2 - q^N)\beta w + rN}{\beta w + rN} > 1. \tag{8}$$

This TFP advantage for assemblers in A is increasing in the probability that at least one part is delayed and in the importance of assembly labour in total costs. It is also increasing in the total number of parts, which might be thought of as complexity.[5] This is intuitive, since the greater the number of parts the greater the chance of a delay in having all parts arrive. This result suggests that parts used in more complex activities have a greater incentive to cluster than do parts used in simpler activities.

The model just described includes a Leontief production structure, where a single delayed part introduces a discrete additional cost in terms of wasted labour and reduced final output price. In this way the model is not precisely comparable to the benchmark timeless model, which allowed for substitutability in inputs. However, the absence of substitutability is not necessary to the results obtained in this section. To demonstrate, replace the Leontief production structure assumed so far by a CES function of N inputs,

$$y = \left[\sum_{i=1}^{N} x_i^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)} \tag{9}$$

where σ is the elasticity of substitution and y is output. Let $\widehat{N}$ hat denote the number of parts that do arrive, with $\widehat{N} \geqslant N_A$ since parts from A are never delayed. Output when $\widehat{N} < N$ parts are available is

$$y = \left[\widehat{N} x^{(\sigma-1)/\sigma} + \left(N - \widehat{N}\right) 0^{(\sigma-1)/\sigma} \right]^{\sigma/(\sigma-1)}.$$

With this production function, when some inputs are unavailable output is zero if $\sigma \leqslant 1$.[6] In other words, every input is essential for production when the elasticity of substitution is low. This implies that the model with substitutability is identical to the Leontief case when substitutability is low, since the key feature of the Leontief case is that production can not proceed until all parts are delivered.

When $\sigma > 1$, the firm faces a choice when some parts fail to arrive on time, since production is feasible with fewer than N parts. If forging ahead in the face of delivery failures is more profitable than waiting, then the resulting expected profit function V_A has ambiguous curvature: it is certainly declining in N_B, but it might or might not be convex.[7] If the function is concave in N_B over some range, then the results illustrated in Fig. 2 no longer apply, and there might not be incentives for full clustering of parts suppliers. We do not regard this case as particularly relevant, however, since the essence of a production process where timely delivery is important is that each part is crucial. High substitutability between inputs seems more relevant over longer horizons.

A second difference between the benchmark model and the synchronisation model of this section is that here we abstract from transport costs on parts deliveries. To illustrate why this assumption economizes on notation without affecting our results, suppose that input costs are

[5] To show this, we calculate the derivative of TFP with respect to N, holding labour's share in cost fixed (this requires an offsetting drop in r as N increases so that rN is constant, that is, $N\,\mathrm{d}r + r\,\mathrm{d}N = 0$). The result is

$$\frac{\partial TFP_{AB}}{\partial N} = \frac{-\beta w}{\beta w + rN} q^N \log q > 0.$$

[6] Technically, when $\sigma \leqslant 1$ the limit of the production function as any part quantity goes to zero is zero.

[7] Expected profits depend on the binomial probability of late arrivals, and this density is not monotonic in N_B. As a result V_A might have a concave portion before becoming convex. Details of this analysis are available from the authors on request.

the same in each location but that iceberg transport costs τ need to be paid on parts delivered from the remote location. In this case Eq. (7) becomes

$$V_A = p - \beta w_A - Nr - \left(1 - q^{N_B}\right)(\delta p + \beta w_A) - N_B r(\tau - 1). \tag{10}$$

Like Eq. (7), Eq. (10) is decreasing and convex in N_B, as is the symmetric expression for V_B. This is all that is required for the mechanism illustrated in Fig. 2. Lastly, as noted in the previous section, relaxing our assumption that final goods are not tradeable would only enhance the incentives for agglomeration, since final goods producer in the cluster will have higher productivity as shown in Eq. (8).

6. Demand or cost uncertainty

We now consider the implications of a different type of uncertainty: unpredictable fluctuations in the level of demand. When demand is variable firms will want to wait until the uncertainty is resolved before deciding how much to produce. The key assumption is that, because of the time that it takes for inputs to be delivered from remote locations, firms must place orders from remote suppliers before finding out the level of demand. Therefore, a firm's response to information about the level of demand or costs depends on the location of component suppliers.[8] We show that, once again, it is efficient for component producers to cluster in one location. The assembler in this location responds flexibly to realizations of demand, while the other assembler is inflexible, being locked into decisions made on the basis of expected demand. We develop the model with demand uncertainty, although we show at the end of the section that assembly cost uncertainty has identical effects.

Demand for the output of each assembler can be high or low, represented by a linear inverse demand curve in which the intercept depends on the state of nature, so

$$p_i = \alpha^s - \beta y_i, \quad i = A, B, \quad s = H, L, \quad \alpha^H > \alpha^L, \tag{11}$$

where p_i is price and y_i is quantity of final product in region i, and superscripts denote the state of nature. High demand occurs with probability ρ. Whether high or low, demand is fleeting, and falls to zero if not met immediately.

As before, the production function has fixed unit input coefficients for each component, and we ignore labour costs in assembly. The assembler in region A faces the following sequence of decisions. First, she has to choose the quantity x_B of components to order from each of the N_B remote suppliers. These have to be ordered before the state of nature is revealed if they are to arrive in time for production. The state of nature is then revealed, and firms choose quantities of components x_A^s from each of the local suppliers. Finally, delivery of all components takes place and production occurs. This is summarised by the following time line:

$$\begin{aligned} \text{Choose } x_B \quad &\rightarrow \quad \alpha^s \text{ revealed} \quad \rightarrow \quad \text{Choose } x_A^s \\ &\rightarrow \quad \text{Produce } y_A^s = \min\left[\ldots x_B \ldots, \ldots x_A^s \ldots\right]. \end{aligned}$$

Output is determined by the component with the minimum delivered quantity because the production function has fixed coefficients. As before, we want to know the dependence of profits on

[8] In this way, we build on the work of Evans and Harrigan [5], who examined a model of "lean retailing" and its implications for international specialization. This section goes beyond their model in focusing on the location of multiple input suppliers.

the location of component producers, and to derive this we have to obtain the profit maximising input choices of assemblers as functions of suppliers' location.

An assembler faces two sets of decisions—quantities of locally produced and of remotely produced components—and we solve these problems in turn. The assembler in A has second stage choice problem (once the state of nature, $s = H, L$, is known) to choose quantities of locally produced components, x_A^s, to maximise v_A^s, defined as

$$\begin{aligned} &v_A^s = x_A^s\left(\alpha^s - \beta x_A^s\right) - N_A r_A x_A^s, \\ &\text{s.t.} \quad x_A^s \leqslant x_B, \quad s = H, L. \end{aligned} \tag{12}$$

The maximand is revenue (where we have used the production function and the inverse demand curve) minus the costs of locally supplied inputs. The constraint reflects the fact that the assembler will never choose more local components than the quantity set by the supply of components coming from region B, because of the fixed coefficient technology. We solve this problem by maximising the Lagrangian

$$L_A^s = x_A^s\left(\alpha^s - \beta x_A^s\right) - N_A r_A x_A^s + \lambda^s\left[x_B - x_A^s\right], \quad s = H, L. \tag{13}$$

The first-order condition with respect to x_A^s implies,

$$\lambda^s = \alpha^s - 2\beta x_A^s - N_A r_A, \quad s = H, L. \tag{14}$$

There are two qualitatively different outcomes, depending on parameters including the level of demand. In one, production is always constrained by the quantity of components coming from the remote supplier, so $x_A^s = x_B$ and $\lambda^s > 0$. In the other this constraint does not bind when demand is low so $\lambda^L = 0$ and x_A^L is solved from (14); some components ordered from B are unused and freely disposed of when demand is low.

The assembler's first stage problem is to choose x_B, before the state of nature is known, to maximise expected profits across states H and L,

$$V_A = \rho v_A^H + (1-\rho) v_A^L - N_B r_B x_B. \tag{15}$$

Varying x_B changes costs directly, and also changes v_A^H and v_A^L via the inequality constraint in (12). The first order condition for this problem is

$$\partial V_A / \partial x_B = \rho \lambda^H + (1-\rho)\lambda^L - N_B r_B = 0 \tag{16}$$

since the Lagrange multiplier measures the value of a unit relaxation of the constraint.

As noted above, there are two cases to study. One we call the no-flexibility case, in which production in both states is constrained by the quantity of components supplied by remote producers and output is the same in both states, independent of the realization of demand. The other is the flexibility case in which sufficient quantities of remote components are purchased such that production is constrained by the quantities of these components only in the high demand case; if demand is low then not all these components are used and there is free disposal of unused components.[9] Which regime applies depends on the values of N_A and N_B, amongst other things. We look first at the no-flexibility case, then turn to the flexibility case and the boundary between the regimes.

[9] Obviously, it is not profitable to discard components in both the high and the low state. The assumption of free disposal could be replaced by costly stock holding into a future period.

No-flexibility: In this case production in both states of nature is determined by quantities ordered from remote suppliers, so $x_A^L = x_A^H = x_B$ and $\lambda^H > 0$, $\lambda^L > 0$. Solution of the first-order conditions (14) and (16) gives

$$\begin{aligned} x_A^L = x_A^H = x_B &= \left[\rho\alpha^H + (1-\rho)\alpha^L - N_A r_A - N_B r_B\right]/2\beta, \\ \lambda^L &= N_B r_B - \rho\left(\alpha^H - \alpha^L\right), \\ \lambda^H &= N_B r_B + (1-\rho)\left(\alpha^H - \alpha^L\right). \end{aligned} \tag{17}$$

The first equation gives purchases of components and hence also the level of output. This is the same in both states, so demand variability goes entirely into the price. Expected profits, V_A, can be computed using (17) in (12) and (15). For present purposes, the important point to notice is that if $r_A = r_B$ then output and sales do not depend on the location of assemblers (the division of N between N_A and N_B, see Eq. (17)), so neither do profits. In the interior of this regime having more local component suppliers does not induce the assembler to change behaviour, and profits are independent of the location of component producers.

Flexibility: In this case production varies with the state of nature. If demand is high then all remote components are used, so $x_A^H = x_B$ and $\lambda^H > 0$. However, if demand is low then not all these components are used, so $x_A^L < x_B$ and $\lambda^L = 0$. Solution of first-order conditions (14) and (16) gives,

$$\begin{aligned} x_A^H = x_B &= \left[\alpha^H - N_A r_A - N_B r_B/\rho\right]/2\beta, \quad \lambda^H = N_B r_B/\rho, \\ x_A^L &= \left[\alpha^L - N_A r_A\right]/2\beta < x_B, \quad \lambda^L = 0. \end{aligned} \tag{18}$$

To establish how profits in this case depend on the location of component producers we proceed as follows. The effects of varying N_B on profits are given by differentiating (15). Using (12), (14) and (18) with $r_A = r_B = r$ gives (see Appendix B for derivation):

$$\frac{\mathrm{d}V_A}{\mathrm{d}N_B} = \rho\frac{\mathrm{d}v_A^H}{\mathrm{d}N_B} + (1-\rho)\frac{\mathrm{d}v_A^L}{\mathrm{d}N_B} - N_B r\frac{\mathrm{d}x_B}{\mathrm{d}N_B} - r x_B = (1-\rho)r\left[x_A^L - x_B\right] < 0. \tag{19}$$

The loss of profits due to a marginal increase in N_B is simply the expected cost of the quantity of this component that remains unused in the low state, $x_A^L - x_B$. Using values of x_A^L and x^B from (19) we further derive,

$$\frac{\mathrm{d}V_A}{\mathrm{d}N_B} = \frac{r(1-\rho)}{2\beta}\left[\alpha^L - \alpha^H + \frac{N_B r}{\rho}\right] < 0, \qquad \frac{\mathrm{d}^2 V_A}{\mathrm{d}N_B^2} = \frac{r^2(1-\rho)}{2\beta\rho} > 0. \tag{20}$$

This establishes that, in the flexibility case, V_A is decreasing and convex in N_B. The intuition is that if the assembler in A did not adjust its production plan in response to changes in N_B then V_A would decline linearly, as more types of component are discarded in the low state. Adjustment raises profits, giving the convexity. The implication is that when parameters are such that firms behave flexibly (that is, sell a different amount depending on the state of demand), there is a force for clustering of component suppliers around one of the final assemblers.

The final piece of analysis is to establish the boundary between the flexible and non-flexible cases. This is when N_B takes value $N_B = \rho(\alpha^H - \alpha^L)/r_B$. This can be derived as the point at which $\lambda^L = 0$ in the no-flexibility case, Eqs. (17). Equivalently, in the flexibility case the condition for $x_B > x_A^L$ is that $N_B < \rho(\alpha^H - \alpha^L)/r_B$, Eqs. (18).

The complete picture is illustrated in Fig. 3. The horizontal axis gives N_B, and the vertical axis gives levels of production and profits of the location A assembler. The graph for the assembler

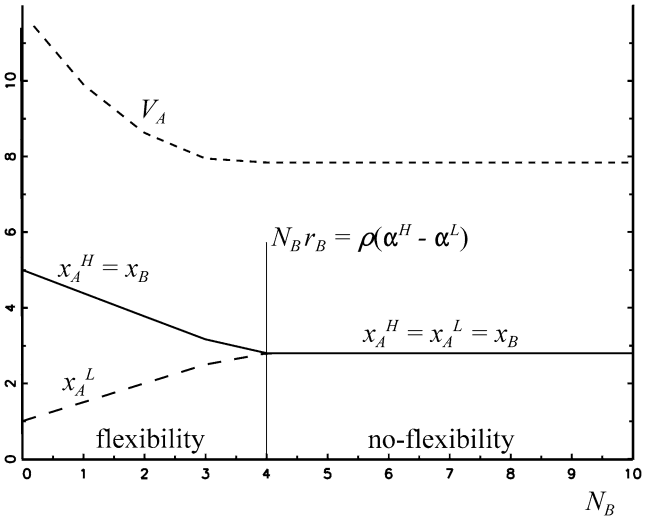

Fig. 3. Demand level uncertainty ($r_A = r_B = 1$, $\rho = 0.45$, $\alpha^H = 20$, $\alpha^L = 12$, $\beta = 1$).

in B would be perfectly symmetric, so we leave it out to avoid clutter. The no-flexibility regime is where $N_B r_B > \rho(\alpha^H - \alpha^L)$; a sufficiently large number of components come from remote suppliers that it is very costly to leave some of each of them unused when the low state occurs. By contrast, when $N_B r_B < \rho(\alpha^H - \alpha^L)$ then only a small share of component types face the risk of being left unutilised and discarded. It is therefore worthwhile to order a larger quantity of each type of remote component, x_B, output becomes state contingent, and the flexibility case applies. In the flexibility case, output in the high state is a decreasing function of N_B while low state output is an increasing function. Intuitively, lower N_B raises the quantity of each remote input, x_B, since the expected number discarded is reduced. However, if the low state occurs then a smaller fraction of component types have zero shadow price (are discarded at the margin) so that marginal cost is higher and output lower.

Notice that there are now two distinct arguments creating convexity of profits, V_A, with respect to N_B. One is that, within the flexibility regime, profits are convex, as discussed above. The other arises because of the kink in V_A due to the change in regimes. Intuitively, having more local suppliers is of no value until some threshold is passed—only then is it worth adjusting production to exploit the benefits of rapid delivery times. Because of the convexity in V_A and the symmetric convexity in V_B, the sum $V_A + V_B$ is also convex (though not drawn in the figure). The implication is, once again, that there is an incentive for all input suppliers to cluster in one location. In such a situation one of the assemblers becomes completely flexible, ordering all its inputs from local suppliers once the level of demand is known. The other is inflexible, as all its inputs take time to be delivered and must be ordered before the state of nature is known.

Several other remarks are worth making on this model. First, price variability is lower in the location with the cluster of activity, as quantities respond to demand shocks. With linear demands the expected price is the same in A and B, $Ep_i = \rho p_i^H + (1-\rho)p_i^L = (\bar{\alpha} + Nr)/2$, as is the expected quantity sold, $Ey_i = \rho y_i^H + (1-\rho)y_i^L = (\bar{\alpha} - Nr)/2\beta$. However, since the region with the cluster produces more in the higher price state, the average value of output produced $E(p_i y_i)/Ey_i$ is higher in the region with the cluster. Once again, the location with the cluster has higher measured productivity.

Notice that this structure is isomorphic to a model in which shocks are on the cost side, rather than the demand side. Suppose that revenue $x_A^s(\alpha^s - \beta x_A^s)$ Eq. (12) were to be replaced by revenue net of labour costs, $\bar{p}x_A^s - (c^s + bx_A^s)x^s$ where $\bar{p}$ is an exogenously given price, and c^s and b are technology coefficients, giving the level and slope of average costs. If c^s is state dependent, then this model is evidently identical to the one above, with parameter α^s replaced by parameter $\bar{p} - c^s$. Uncertainty—in either costs or demand—means that profits are higher if input decisions can be postponed. The argument of this section shows that it also generates convexity of profits with respect to the location of component suppliers, implying that this uncertainty gives rise to clustering.

As in the previous section, allowing for final goods to be traded will strengthen the incentive for clustering, and clustering will be efficient. One location would produce all the output using parts ordered from nearby suppliers after demand is revealed.

In the synchronisation model of Section 5 we showed that zero input substitutability was not key to the result, the model holding for elasticities of substitution less than one and in some interval above one. How general are the results of this section? An important mechanism is the disincentive to order parts from remote suppliers, because of the possibility that the low state of nature occurs. With zero substitutability in production any over-ordered parts are discarded, while some degree of substitutability means that they will be used, although perhaps with very low marginal value.

With some substitutability in production the economic model is simpler to write down (although less easy to solve analytically). The firm chooses $\{x_A^H, x_A^L, x_B\}$ to maximise expected profits:

$$V_A = \pi v_A^H + (1-\rho)\pi_A^L$$

where state-contingent profits are

$$v_A^H = R\big(\alpha^H, y\big(N_A, x_A^H, N_B, x_B\big)\big) - r_A N_A x_A^H - r_B N_B x_B,$$
$$v_A^L = R\big(\alpha^L, y\big(N_A, x_A^L, N_B, x_B\big)\big) - r_A N_A x_A^L - r_B N_B x_B,$$

in which the function R is revenue, depending on the demand shift parameter α^i and output, y.

In principle this is a simple expected profit maximisation problem, but in practice it is not possible to solve the first-order conditions analytically for the optimal input choices. We have numerically explored a number of cases. If demand is linear (as in Eq. (11)) and the production function is CES, then we find that convexity of the expected profit function in N_B is a fragile result: it is possible to construct examples where an elasticity of substitution as low as $\sigma = 0.1$ is sufficient to generate a concave function and thus overturn our conclusions about agglomeration incentives.[10] Alternatively, if demand is iso-elastic and the demand shift parameter multiplicative, then the results of this section hold over a wide range of values of σ, and certainly for all values up to and including unity.

While it is useful to understand the limits of our result, we regard the Leontief case as the most relevant for understanding the implications of timeliness for agglomeration. For most production processes substitutability is near zero in the short run: it is not possible to substitute any number of tires for carburetors in short supply. In such cases the incentive to wait until demand is realized before making production plans generates an incentive for clustering.

[10] Our numerical results are available on request.

7. Summarising the models

In Section 4 we presented a standard economic geography model with orthodox transportation costs but no role for timeliness. In that model, there is no incentive for suppliers to agglomerate with one of the ex ante identical assemblers; in fact, the contrary holds, with efficiency tending to lead to a 50–50 split of suppliers in each assembly location.

The models of Sections 5 and 6 give the opposite conclusion. In each model, orthodox transport costs are absent, but a value for timeliness is introduced. In Section 5, the value of timeliness comes from reducing the chance of costly production delays. The model of Section 6 shows how uncertainty about the level of demand or costs creates a demand for timely delivery.

In these two models, the demand for timely delivery creates a convexity in profits as a function of the location of suppliers, which can be thought of as an increasing marginal value of timeliness. Because of this convexity, there is in each model a force for agglomeration. All the suppliers will tend to locate in the same region as one of the ex ante identical assemblers, giving that assembler the full benefit of flexibility while the other assembler makes do with non-timely delivery. This corner solution yields higher profits than are earned if both assemblers face long delivery times on some fraction of their inputs.

The corner solution outcome is strengthened if there is low-cost shipment of final goods. In both models, the location where the suppliers cluster (say, location A) will have higher measured productivity. The only reason that the assembler in B does not move her operation to A is that we have assumed that final goods are not tradeable. If we remove this assumption and allow final goods to be shipped from A to B at low cost, then the incentive for clustering is even stronger.

In both models we have assumed that inventories are prohibitively expensive. This assumption is crucial, since low-cost inventories are an obvious substitute for timely delivery. In an earlier version of this paper (Harrigan and Venables [9]), we studied a model of the tradeoff between inventories and timely delivery, and showed how an incentive for clustering can arise even when inventories are available. In that model firms face consumers who are fickle and picky, and the more unpredictable consumers are the more expensive it is to hold enough inventories ex ante to cater to their every ex post whim. The resultant option value of waiting for the resolution of uncertainty can create an incentive for clustering similar to that discussed in Section 5.

8. Policy implications

Governments are perennially interested in regional economic development, and subsidies have often been used (and even more often proposed) as a means of sustaining regional economies. In particular, subsidies to manufacturing assembly plants have been justified in the hope that their presence in a region will trigger agglomerations of related activities. The baseline model of Section 3 offers some theoretical support for such a subsidy: starting from a world with one assembly plant with all suppliers located nearby, establishment of a second assembly plant elsewhere creates an incentive for some suppliers to move near the new plant. This is because of the decreasing marginal value of proximity in such a model: the first supplier that moves to the location of the new assembler will generate greater value as a result.

In contrast, our models of timeliness deliver the opposite conclusion. Because of the increasing marginal value of timeliness (and hence proximity), there is no incentive for any supplier to move to the location of a new assembly plant. If these models apply, we would expect new assembly plants that locate far from existing plants (for whatever reason) to not be followed by their suppliers. As shown by Klier [12], this is what has happened in the US auto industry: assem-

bly plants established far from the "auto corridor" as a result of government subsidies (BMW in South Carolina, Mercedes Benz in Alabama) or private incentives (NUMMI in California) have not been followed by a substantial migration of suppliers.[11]

9. Conclusions

Just-in-time production methods have been researched extensively in the management literature, but have received almost no attention in economics.[12] This paper has taken a step towards redressing this imbalance. In an uncertain environment the benefits of securing timely delivery of components alters the efficient spatial organisation of production. In a situation in which conventionally modelled monetary trade costs would lead to dispersed location of component suppliers, delay or uncertainty in delivery times cause clustering. The efficient organisation of production requires the concentration of all component plants next to just one of several assembly plants.

This is a new mechanism for agglomeration. In this paper we have developed the idea in a simple framework in which the final product is non-tradeable, it is prohibitively costly to have multiple plants producing the same component, and factor prices are fixed. Each of these assumptions could be relaxed, embedding the mechanism in a wider economic environment.

Acknowledgments

Thanks to Niko Matouschek and participants in seminars at the LSE, University of Nottingham, City University of Hong Kong, Hitotsubashi University, IMF, Federal Reserve Bank of New York, AEA (San Diego), and NBER for helpful comments.

Appendix A

Consider any symmetric unit cost function $c(\ldots r\ldots;\ \ldots r+\mathrm{d}r\ldots)$ in which inputs are partitioned into a group (A) available at price r, the remainder available at price $r+\mathrm{d}r$ (group B). Quantities demanded in each group are x_A, x_B, $x_A > x_B$. The increase in costs when a product moves from group A to group B is $(\partial c/\partial r)\,\mathrm{d}r = x_A\,\mathrm{d}r$ (by Shepherd's lemma). As more products enter group B so x_A must increase (in order that input levels are sufficient to produce the unit of output),[13] meaning that the cost of moving inputs from group A to group B is increasing. This increasing marginal cost gives the convexity of the cost function with respect to N_B and the consequent concavity of profits.

Appendix B. Derivation of Eq. (19)

Total differentiation of (15) gives

$$\frac{\mathrm{d}V_A}{\mathrm{d}N_B} = \rho\frac{\mathrm{d}v_A^H}{\mathrm{d}N_B} + (1-\rho)\frac{\mathrm{d}v_A^L}{\mathrm{d}N_B} - N_B r\frac{\mathrm{d}x_B}{\mathrm{d}N_B} - rx_B.$$

[11] The "auto corridor" is the region in the middle of the country where most auto production is concentrated. It includes seven contiguous states: Michigan, Ohio, Indiana, Illinois, Wisconsin, Kentucky, and Tennessee.

[12] See Cremer [3] for a rare exception.

[13] The cross-partial derivatives of a symmetric unit cost function are positive, so raising the price of some inputs increases demand for other.

By (12) and (13), this is

$$\frac{\mathrm{d}V_A}{\mathrm{d}N_B} = \rho\left[\lambda^H \frac{\mathrm{d}x_A^H}{\mathrm{d}N_B} + rx_A^H\right] + (1-\rho)\left[\lambda^L \frac{\mathrm{d}x_A^L}{\mathrm{d}N_B} + rx_A^L\right] - N_B r \frac{\mathrm{d}x_B}{\mathrm{d}N_B} - rx_B.$$

Using values of λ and of x_A^L and x^B from (18) gives Eq. (19).

References

[1] J. Anderson, E. van Wincoop, Trade costs, Journal of Economic Literature 42 (2004) 691–751.
[2] K.G. Binmore, P. Dasgupta, Nash bargaining III, in: K.G. Binmore, P. Dasgupta (Eds.), The Economics of Bargaining, Blackwell Sci., Oxford, 1987.
[3] J. Cremer, Towards an economic theory of incentives in just-in-time manufacturing, European Economic Review 39 (1995) 432–439.
[4] P. Dicken, Global Shift: Transforming the World Economy, Chapman & Hall, London, 1998.
[5] C. Evans, J. Harrigan, Distance, time, and specialization: Lean retailing in general equilibrium, American Economic Review 95 (1) (2005) 292–313.
[6] M. Fujita, P. Krugman, A.J. Venables, The Spatial Economy: Cities, Region and International Trade, MIT Press, Cambridge, MA, 1999.
[7] M. Fujita, J.-F. Thisse, The Economics of Agglomeration: Cities, Industrial Location and Regional Growth, Cambridge Univ. Press, Cambridge, UK, 2001.
[8] J. Harrigan, Airplanes and comparative advantage, 2005, in preparation.
[9] J. Harrigan, A.J. Venables, Timeliness, trade, and agglomeration, Working paper 10404, NBER, 2004.
[10] D. Hummels, Time as a trade barrier, Mimeo, Purdue University, 2001.
[11] G.R. Jones, J.M. George, C.W. L Hill, Contemporary Management, McGraw–Hill, Boston, 2000.
[12] T. Klier, Thomas, Agglomeration in the US auto supplier industry, Federal Reserve Bank of Chicago Economic Perspectives Q I (1999) 18–34.
[13] M. Kremer, The O-ring theory of economic development, Quarterly Journal of Economics 108 (3) (1993) 551–575.
[14] S.S. Rosenthal, W.C. Strange, The determinants of agglomeration, Journal of Urban Economics 50 (2001) 191–229.
[15] S.S. Rosenthal, W.C. Strange, Evidence on the nature and sources of agglomeration economics, in: J.V. Henderson, J.F. Thisse (Eds.), Handbook of Urban and Regional Economics, vol. 4: Cities and Geography, North-Holland, Amsterdam, 2004.
[16] M. Storper, E. Leamer, The economic geography of the Internet age, Journal of International Business Studies 32 (4) (2001) 641–665.
[17] M. Storper, A.J. Venables, Buzz: Face to face contact and the urban economy, Journal of Economic Geography 4 (2004) 351–370.
[18] J. Sutton, Non-cooperative bargaining theory: An introduction, Review of Economic Studies LIII (1986) 709–724.
[19] A.J. Venables, Geography and international inequalities: The impact of new technologies, in: B. Pleskovic, N.H. Stern (Eds.), Annual World Bank Conference on Development Economics, 2001–2002.