A Unified Theory of Cities[1]

Jacques-François Thisse

Matthew A. Turner

Philip Ushchev

March 5, 2021
Preliminary

Abstract: How do people arrange themselves when they are free to choose work and residence locations, when commuting is costly, and when increasing returns may affect production? We consider this problem in a framework with discrete locations and households with heterogenous preferences over workplace-residence pairs. We provide a general characterization of equilibrium throughout the parameter space. The introduction of preference heterogeneity into an otherwise conventional urban model fundamentally changes equilibrium behavior. Stronger increasing returns to scale need not concentrate economic activity and lower commuting costs need not disperse it. The qualitative behavior of the model as returns to scale increase accords with changes in the patterns of urbanization observed in the Western world between the pre-industrial period and the present.

JEL: R0

[1]We gratefully acknowledge helpful conversations with Dan Bogart about the history of cities and agglomeration effects. Turner is grateful for the support of a Kenen fellowship at Princeton University during part of the time this research was conducted. Thisse and Ushchev acknowledge the support of the HSE University Basic Research Program.

1 Introduction

Two centuries ago, one human out of ten lived in a city, incomes were a fraction of their present level, and commuting occurred on foot. Today, more than half of the world's population is urbanized, and commuting by foot is a curiosity in much of the world. The process of urbanization is widely understood as a contest between increasing returns to scale in production and the cost of commuting. Returns to scale in production leads to highly concentrated employment while commuting allows people to live at lower densities than those at which they work: as returns to scale increase and commuting costs fall, cities grow.

We investigate how the organization cities is related to the strength of returns to scale and to the cost of commuting. Our problem is to understand how people arrange themselves when they are free to choose work and residence locations, when commuting is costly, and when some form of increasing returns may affect production. This is the fundamental problem of urban economics. We consider this problem in a framework with discrete locations and households with heterogenous preferences over workplace-residence pairs. We provide a general characterization of equilibrium throughout the parameter space. In particular, we consider the possibility of arbitrarily strong increasing returns to scale that has long resisted analysis.

The contest between increasing returns to scale and commuting costs does not play out as the conventional wisdom suggests. Qualitative features of an equilibrium city depend sensitively on the strength of returns to scale in production. When increasing returns to scale are weak, equilibrium is unique, economic activity is centralized, and stronger returns to scale concentrate economic activity. When increasing returns to scale are moderate, multiple stable equilibria may occur and stronger returns to scale disperse economic activity. When returns to scale are strong, equilibrium is unique and stronger returns to scale continue to disperse economic activity. That is, the conventional intuition that returns to scale is an agglomeration force does not hold for sufficiently strong increasing returns to scale. Our findings about commuting costs are equally surprising. We find that economic activity disperses both when commuting costs are high and when they are low. It is only at intermediate levels of commuting costs that highly concentrated economic activity can arise. That is, when the location of production and residence is endogenous, the standard intuition that decreases in transportation costs must disperse economic activity need not apply.

Our investigation is important for a number of reasons. First, we address the most basic problem of urban economics. While much progress has been made on this problem, existing work often relies on strong or implausible simplifying assumptions and arbitrary restrictions on the strength of the dispersion and agglomeration forces. For example, the basic monocentric city model developed by Alonso (1964), Mills (1967) and Muth (1969) (AMM) restricts attention to the case where agents only choose residential location, as their workplace is set exogenously, while the recent development of quantitative models relies heavily on assumptions that prohibit multiple equilibria (Redding and Rossi-Hansberg, 2017). We advance our understanding of this problem by applying the machinery of the recent quantitative models to a simple discrete geography. This simplifying assumption permits an otherwise general theory of the economics of cities.

Second, existing work on the economics of cities can be usefully divided into two frameworks. The first is the urban economics literature. This literature assumes space is a continuum and, for the most part, that households are homogenous. The first general statement of our problem is due to

Ogawa and Fujita (1980), who provide a partial characterization of equilibria under strong simplifying assumptions. Lucas (2001) and Lucas and Rossi-Hansberg (2002) extend their results, but still do not completely characterize equilibria. The second framework, now known as 'quantitative spatial models' or QSM, differs in two important ways. First, they are based on a discrete choice model where households have heterogenous preferences over workplace-residence pairs. This assumption has no analog in the older literature and, to the extent that preference heterogeneity is a feature (at least) of developed world economies, this is an important innovation. Second, although quantitative spatial models typically permit closed form solutions, they are usually applied to complicated geographies and can only be evaluated numerically. This has two implications. First, numerical analysis is only feasible in regions of the parameter space where equilibrium is unique. Second, because most results are numerical, it is not clear whether basic comparative statics derived in the context of the older continuous space homogenous agents models, e.g., commuting costs capitalized into land rent, apply to models based on heterogenous agents and discrete space. We help to unify these two literatures by characterizing equilibria in a model where heterogenous agents choose among discrete locations for work and residence and where arbitrary returns to scale may affect productivity. We find that preference heterogeneity fundamentally changes the behavior of equilibrium cities.

Finally, for the cities of the Western world, we present a stylized 500 year the history of urban form and increasing returns to scale. We then compare the qualitative features of this history with the behavior of our model as increasing returns to scale change. We find a qualitative relationship between the behavior we see in our model and history. Thus, our model appears to offer a simple theory to predict qualitative behavior of urban geography in the Western world from the pre-industrial period to the present.

2 Literature

We can usefully partition the literature into an older urban economics literature and a more recent literature on quantitative spatial models. Most papers in the urban economics literature rest on the following assumptions. Households are homogenous or there are at most a small number of types. Space is continuous and uniform, whether on a line or in a plane, and equilibrium cities are generally symmetric around a single exogenously selected point. Commuting consumes real resources, either time or the numéraire consumption good. The simplest, and most influential model in this literature is the monocentric city model (Fujita, 1989). This workhorse model rests on the assumption that households choose only their location of residence, the location of work is fixed exogenously at the center, although the model is otherwise quite general. In particular, firms may substitute between land and labor, and households between land and consumption.

Beckmann (1976) is one of the earliest models that explains the endogenous formation of a city center. He assumes that the utility of an individual depends on the average distance to all individuals with whom this person interacts as well as on the amount of land she buys on the market. Under such preferences, the city exhibits a bell-shaped population density distribution supported by a similarly shaped land rent curve. Thus, the city emerges here as a social magnet. Alternatively, in spirit of Armington (1969) (where each variety of a product is differentiated by the place where it is produced) one might think of the density of social interaction between two locations as the volume of trade between these two locations whose size decreases with distance.

In this case, Beckmann's model may be viewed as a reduced form of Allen and Arkolakis (2014).

Ogawa and Fujita (1980) consider a simple setting where firms choose only their location and households choose only their places of work and residence. They introduce the idea of a linear potential function in that firms benefit from spillovers with every location, while distant spillovers are weaker than those nearby. This assumption, now conventional, requires that the productivity of any given location responds to a distance weighted mean of employment at all locations. This creates an agglomeration externality as productivity depends on the density of nearby employment, while land scarcity acts as a dispersion force. As the benefits of spillovers increases relative to the cost of commuting, they observe first a uniform, then a monocentric, and finally a duocentric equilibrium.

Fujita and Ogawa (1982) build on this initial paper by considering a negative exponential decay function. When the spatial decay parameter is very small, numerical results indicate that the three possible configurations described above still occur. However, when the spatial decay parameter takes higher values, the following may occur: (i) there exist patterns exhibiting several centers; (ii) there is multiplicity of equilibria; and (iii) the transition from one equilibrium to another may be catastrophic. Lucas (2001) establishes general existence results in a model where agglomeration economies are not too strong and where firms and households are allowed to choose locations, subject to firms being restricted to a central business district and households to the surrounding region. Lucas and Rossi-Hansberg (2002) revisits the problem posed by Fujita and Ogawa (1982), but allow firms and households to make the same substitutions between labor and land and consumption and land as in a typical AMM model. They consider both constant and increasing returns to scale. They establish general existence and uniqueness results, but otherwise rely on numerical methods and restrict attention to 'weak enough' increasing returns that the multiple equilibria observed in Fujita and Ogawa (1982) do not arise.

Recently, a second class of models, 'quantitative spatial models' (QSM), has been brought to bear on problems of urban economics (Redding and Rossi-Hansberg, 2017). The fundamentals of these quantitative models are different from the older urban economics literature. In the QSM literature, model cities consist of discrete sets of locations rather than continuous spaces, and they describe realistic rather than highly stylized geographies. More importantly, this literature considers heterogenous rather than homogenous agents. It also treats commuting costs as purely psychic, unlike the older urban economics literature where commuting typically consumes real resources. The emphasis of QSM literature is also different from the older literature. Where the older literature tends to focus on analytic solutions and qualitative results, the QSM literature focuses on the numerical evaluation of particular comparative statics in models that describe particular real world locations. Examples include the fall of the Berlin wall (Ahlfeldt *et al.*, 2015) and the construction of the Transmilenio Bus Rapid Transit System in Bogota (Tsivanidis, 2019).

The QSM literature draws on a long history of scholarship on discrete choice models dating back to Luce (1959) and McFadden (1974), and also on the well-established literature that applies discrete choice models to transportation, location and trade problems (Anas, 1983; de Palma *et al.*, 1985; Eaton and Kortum, 2002). With that said, much of the recent work closely follows Ahlfeldt *et al.* (2015).[2] In this model, households have preferences over housing and consumption, as in the older urban economics literature, and commute from home to work. Space is discrete

[2]For example: Severen (2019); Monte *et al.* (2018); Dingle and Tintlenot (2020); Tsivanidis (2019); Couture *et al.* (2020); Balboni *et al.* (2020), Heblich *et al.* (2020), Herzog (2020) and Allen *et al.* (2016).

and is described a matrix of pairwise commuting costs. These matrices are typically constructed to describe commuting costs between pairs of neighborhoods in the empirical application of interest. These commuting costs are purely psychic and commuting does not consume real resources. In the spirit of discrete choice models, households have heterogenous preferences over work-residence pairs and each household selects a unique pair. Locations are heterogenous in their amenities and productivity, and the possibility of endogenous agglomeration economies is sometimes considered. As a consequence of this realism, analytical results are limited. It is possible to derive analytic expressions characterizing equilibrium. There are also well known existence results, and uniqueness has been established for the case when increasing returns are small enough to preclude multiplicity. Beyond this, most of what is known about these models results from the numerical evaluation of particular, empirically founded, comparative statics.

Our analysis is based on a hybrid of the models considered in these two literatures. Most of the main features of our model we borrow from the quantitative spatial literature. We have heterogenous agents, discrete space, and purely psychic commute costs. However, we restrict attention to a simple linear city of homogenous locations in the spirit of the urban economics literature. In effect, we are using the new quantitative spatial model toolbox to analyze the problem of spatial equilibrium in the stylized geographies of the older urban economics literature. As we will see, this will lead us to a better understanding of both classes of models and their differences. Indeed, the extent of household heterogeneity in our statement of the QSM model is parameterized by a single variable, and so we are able to investigate what happens as the heterogenous households of the QSM literature approach the homogeneity of the older urban economics literature. Our hope is that the study of simple, but still rich, settings like ours will serve as reference structures that will shed new light on the various effects taken into account in QSM which interact in ways that are not always transparent.

Anas (1990) precedes us in efforts to unify the discrete choice framework with standard models of urban economics. Anas (1990) combines the monocentric city model with a finite number of colinear locations and a population of consumers who have heterogeneous tastes for housing and a composite good, described by a multinomial logit model. Anas shows that more taste heterogeneity flattens the land rent and population density gradients. Hence, the city radius expands, residential and population densities around the center decrease whereas they rise near the city edge. To put it differently, a more heterogeneous population makes the urban structure more decentralized. We will see what this result becomes when the location of production is also endogenous.

3 A discrete city with heterogenous households

A city consists of a finite set of locations $\mathcal{I}$ with $|\mathcal{I}| = I$ and each location is endowed with one unit of land. The city is populated by a continuum $[0, 1]$ of households indexed by ν and by a competitive production sector whose size is endogenous. All households choose a residence $i \in \mathcal{I}$, a workplace $j \in \mathcal{I}$ and their consumption of housing and a tradable produced good.

Each household $\nu \in [0, 1]$ has a type $\mathbf{z}(\nu) \equiv (z_{ij}(\nu)) \in \mathbb{R}_+^{I \times I}$. Thus, a household's type is a vector of non-negative real numbers, one for each possible workplace-residence pair ij. The mapping $\mathbf{z}(\nu) : [0, 1] \to \mathbb{R}_+^{I \times I}$ is such that the distribution of types is the product measure of I^2 identical

Fréchet distributions:

$$F(\mathbf{z}) \equiv \exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{I} z_{ij}^{-\varepsilon}\right). \tag{1}$$

Households have heterogenous preferences over workplace-residence pairs, and household types will parameterize these workplace-residence preferences. Thus, types describe preferences while $\varepsilon > 0$ describes the heterogeneity of preferences. An increasing ε reduces preference heterogeneity and conversely.

Households must commute between workplace and residence. Commuting from i to j involves an iceberg cost $\tau_{ij} \geq 1$. This cost is the same for all households and $\tau_{ij} = 1$ if and only if $i = j$. The commuting household utility, but does not directly impact the household budget. That is, the commuting cost is entirely psychic.[3]

A household that lives at i and works at j has utility

$$U_{ij}(\nu) = \frac{z_{ij}(\nu)}{\beta^{\beta}(1-\beta)^{1-\beta}} \frac{H_{ij}^{\beta} C_{ij}^{1-\beta}}{\tau_{ij}}. \tag{2}$$

where H_{ij} is household housing consumption and C_{ij} is household consumption of a homogeneous and costlessly tradable numéraire good. To simplify the model, 'housing' consists entirely of land.

Given the choice of workplace and residence, ij, the household budget constraint is,

$$W_j = C_{ij} + R_i H_{ij}, \tag{3}$$

where W_j is the wage paid at location j and R_i the land rent at i. The resulting indirect utility function is

$$V_{ij}(\nu) = z_{ij}(\nu) \frac{W_j}{\tau_{ij} R_i^{\beta}}. \tag{4}$$

Summing up, we have a discrete choice model in which households choose a single workplace-residence pair ij from among a finite number of indivisible alternatives in order to maximize their indirect utility (4).

Let

$$S_{ij} = \left\{ \mathbf{z} \in \mathbb{R}_{+}^{I \times I} \,;\, V_{ij}(\mathbf{z}) = \max_{r,s \in \mathcal{I}} V_{rs}(\mathbf{z}) \right\}$$

be the set of types $\mathbf{z}$ such that ij is (weakly) preferred to all other location pairs rs. Then, using (1) and (4), the share s_{ij} of households who choose the location pair ij equals

$$s_{ij} = \mu\left(\mathbf{z}^{-1}(S_{ij})\right) = \frac{\left[W_j/(\tau_{ij} R_i^{\beta})\right]^{\varepsilon}}{\sum_{r \in \mathcal{I}} \sum_{s \in \mathcal{I}} \left[W_s/(\tau_{rs} R_r^{\beta})\right]^{\varepsilon}}, \tag{5}$$

[3] As Redding (2020) points out, with a linear homogenous utility function this formulation is equivalent to one in which the budget is $W_j/z_{ij} = C_{ij} + R_i H_{ij}$ and z_{ij} does not appear directly in the utility function. This equivalence also requires that the taste parameters do not enter the market clearing condition for land. Although this specification is widely used in the literature (see, e.g., Ahlfedt *et al.*, 2015; Monte *et al.*, 2018) we prefer our formulation because it makes it easier to see that individual heterogeneity affects tastes and not productivity.

where the last equality stems from the Fréchet distribution assumption and μ is the Lebesgue measure over $[0,1]$.[4] Our model is static and so all choices occur simultaneously.

Because the $V_{ij}(\nu)$ are Fréchet distributed, the average utility $\bar{V}$ across all households equals:

$$\bar{V} \equiv \int_0^1 \max_{i,j=1,\dots I} V_{ij}(\mathbf{z}(\nu))\mathrm{d}\nu = \Gamma\left(\frac{\varepsilon-1}{\varepsilon}\right)\left\{\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}}\left[W_j/\left(\tau_{ij}R_i^{\beta}\right)\right]^{\varepsilon}\right\}^{1/\varepsilon},$$

where $\Gamma(\cdot)$ is the gamma function.[5] Households that share the same type choose the same location pair[6] ij and reach the same equilibrium utility level, while households who make the same choice may have different types and do not have the same equilibrium utility level. Likewise, households that choose different location pairs do not generally enjoy the same equilibrium utility level. In general, equilibrium utility varies with type.

In the QSM based literature, locations are typically endowed with both employment and residential 'amenities' that scale the payoffs from work and residence in each location. While it is straightforward to incorporate these terms into our model, we omit them for two reasons. First, it eases notation. More importantly, it restricts our QSM based model to a stylized, featureless landscape of the sort considered in the urban economics literature, and thereby facilitates the comparison of our results to those in this literature.

We have described an explicitly static and deterministic model of heterogenous households. Existing formulations of this model are less clear about this issue. This can lead to questions about households' expectations and the extent to which those expectations coincide with realized outcomes.[7] Such questions are usually resolved by informal appeals to the law of large numbers, despite the difficulty of formulating the law of large numbers for a continuum (Judd, 1985; Feldman and Gilles, 1985; Uhlig, 1996). By insisting on a deterministic model of heterogenous households, we avoid these (admittedly subtle) issues. There is no need for probabilities and expectations; our outcomes are shares and averages.

We can rewrite (5) as,

$$s_{ij} = \kappa R_i^{-\beta\varepsilon} W_j^{\varepsilon} \tau_{ij}^{-\varepsilon}, \tag{6}$$

for

$$\kappa \equiv \left[\Gamma\left(\frac{\varepsilon-1}{\varepsilon}\right)\right]^{\varepsilon} \bar{V}^{-\varepsilon}. \tag{7}$$

Let M_i and L_i be the mass of residents and households at i. Labor and land market clearing requires

$$\sum_{i\in\mathcal{I}} L_i = \sum_{j\in\mathcal{I}} M_j = 1, \tag{8}$$

where the residential population at i is

$$M_i \equiv \sum_{j\in\mathcal{I}} = \kappa R_i^{-\beta\varepsilon} \sum_{j\in\mathcal{I}} W_j^{\varepsilon} \tau_{ij}^{-\varepsilon}, \tag{9}$$

[4] Note that there might be types which belong to more than one S_{ij} as they are indifferent between their multiple favorite choices. However, the set of such types always has a zero measure, so that they do not affect the type distribution (1).

[5] For the average utility to be finite, we require $\varepsilon > 1$.

[6] Except for the measure zero set of types that is indifferent between alternatives.

[7] These issues are stated clearly in Dingle and Tintlenot (2020).

while the labor force at j is

$$L_j \equiv \sum_{i \in \mathcal{I}} s_{ij} = \kappa W_j^{\varepsilon} \sum_{i \in \mathcal{I}} R_i^{-\beta\varepsilon} \tau_{ij}^{-\varepsilon}. \tag{10}$$

Because each location i is endowed with one unit of land, land market clearing requires

$$H_i + N_i = 1, \tag{11}$$

where H_i is the amount of residential land and N_i is the amount of commercial land at location i. By neglecting complementary slackness conditions in (8)-(11), we implicitly restrict attention to interior equilibria. To see that this is without loss of generality, suppose that $M_k = 0$ for location k. Since at least one wage must be positive, the sum in the right-hand side of (9) is positive. Therefore, for (9) to be satisfied, the land rent R_k must be infinite. Thus, restricting attention to interior equilibria is equivalent to ruling out infinite prices.

Applying Roy's identity to (4), we have

$$H_i \equiv \sum_{j \in \mathcal{I}} s_{ij} H_{ij} = \sum_{j \in \mathcal{I}} s_{ij} \frac{\beta W_j}{R_i}. \tag{12}$$

Substituting from (6) gives

$$R_i = \left(\frac{\beta\kappa}{H_i} \sum_{j \in \mathcal{I}} W_j^{1+\varepsilon} \tau_{ij}^{-\varepsilon} \right)^{\frac{1}{1+\beta\varepsilon}}. \tag{13}$$

This is the bid rent that households are willing to pay to reside at i, given $(W_1, ..., W_I, H_i)$. Land rent accrues to absentee landlords who play no further role in the model.

Assume that the numéraire is produced under perfect competition while the production function at location j is

$$Y_j = A_j L_j^{\alpha} N_j^{1-\alpha}, \tag{14}$$

where A_j is location-specific TFP and $0 < \alpha < 1$. It is well established that knowledge and information spillovers are localized because they arise mainly through face-to-face contacts (Arzaghi and Henderson, 2008; Ahlfeldt *et al.*, 2015; Liu *et al.*, 2018; Battiston *et al.*, 2021). Hence, we assume that A_j depends only on the level of employment at j,

$$A_i = L_i^{\gamma}, \tag{15}$$

for $\gamma \geq 0$.

If location j hosts a positive share of the production sector, the first-order conditions for the production sector yields the equilibrium wage and land rent as functions of land-labor ratio:

$$W_j = \alpha A_j \left(\frac{N_j}{L_j} \right)^{1-\alpha}, \tag{16}$$

$$R_j = (1-\alpha) A_j \left(\frac{L_j}{N_j} \right)^{\alpha}. \tag{17}$$

Since we consider interior equilibria, these expressions imply that W_j and R_j are positive for all j.

Dividing (16) by (17) and simplifying,

$$\frac{W_j}{R_j} = \frac{\alpha}{1-\alpha}\frac{N_j}{L_j}. \tag{18}$$

When $\gamma = 0$, note that

$$R_j = (1-\alpha)\left(\frac{N_j}{L_j}\right)^{-\alpha} = (1-\alpha)\left(\frac{W_j}{\alpha}\right)^{-\alpha/(1-\alpha)}. \tag{19}$$

This expression implies that R_j and W_j move in opposite directions under constant returns: a higher land rent at j is equivalent to a lower wage at this location. Moreover, the land rent at j is positive and finite if and only if W_j is positive and finite.

Our city is described by I-vectors of real numbers. To describe them define a *spatial pattern* to be an element of $\mathbb{R}_+^I$, a vector enumerating a non-negative real number for each location in $\mathcal{I}$. The spatial patterns that describe our model city are: the *residential pattern* $\mathbf{M} \equiv (M_1, M_2, \ldots, M_I)$; the *employment pattern* $\mathbf{L} \equiv (L_1, L_2, \ldots, L_I)$; the *housing pattern* $\mathbf{H} \equiv (H_1, H_2, \ldots, H_I)$; the *commercial pattern* $\mathbf{N} \equiv (N_1, N_2, \ldots, N_I)$; the *wage pattern* $\mathbf{W} \equiv (W_1, W_2, \ldots, W_I)$; and the *land rent pattern* $\mathbf{R} \equiv (R_1, R_2, \ldots, R_I)$. A spatial pattern is *interior* if none of its elements is zero.

We are now able to provide a precise definition of equilibrium for our discrete city with heterogenous agents.

Definition 1 *A spatial equilibrium is a set of patterns,* $\{\mathbf{M}^*, \mathbf{L}^*, \mathbf{H}^*, \mathbf{N}^*, \mathbf{W}^*, \mathbf{R}^*\}$ *such that*

i. All households make utility-maximizing choices of workplace, residence, housing and consumption to satisfy equation (5),

ii. The tradable good is produced by a competitive sector and the first-order conditions (16) and (17) are satisfied,

iii. All households live and work somewhere so that equation (8) holds, and

iv. Land markets clear, so that equation (11) holds.

4 A linear city with three locations

Our discrete city with heterogenous households is based on assumptions and functional forms common in the QSM literature. Because the object of the QSM literature is (usually) the numerical evaluation of particular empirically founded comparative statics, this means that model is more complex than we require to investigate analytic comparative statics or to compare with the more stylized urban economics literature. Therefore, we let $\mathcal{I} \equiv \{-1, 0, 1\}$ and restrict our discrete city with heterogenous agents to a geography consisting of three identical locations evenly spaced along a line.

This geography is the simplest in which to examine when activities concentrate in the center or disperse to the periphery. Thus, it is the simplest geography in which we can consider when a city forms a central district where activity is concentrated, and when economic activity remains dispersed. It is also qualitatively similar to the geography of the linear monocentric city that forms the basis for much of the urban economics literature.

For tractability and to ease comparison with the urban economics literature, we focus on *symmetric spatial patterns* $x = (x_{-1}, x_0, x_1)$ where $x_1 = x_{-1}$, and usually write such patterns as (x_1, x_0, x_1) rather than (x_{-1}, x_0, x_1). We say that a symmetric spatial pattern is *bell-shaped*, *flat*, or *U-shaped* as x_0 is greater than, equal to, or less than x_1. Say that a symmetric spatial pattern (x_1, x_0, x_1) is *more centralized* than a symmetric spatial pattern (y_1, y_0, y_1) if and only if $\frac{x_0}{x_1} > \frac{y_0}{y_1}$.

The restriction to three locations and symmetric patterns, allows us to focus our attention on the symmetric, three element versions of the spatial patterns listed above: (M_1, M_0, M_1), (L_1, L_0, L_1), (H_1, H_0, H_1), (N_1, N_0, N_1), (W_1, W_0, W_1), and (R_1, R_0, R_1). That is, the patterns for residence, employment, housing, industry, wages and land rents.

Our analysis below is organized around studying the centrality of these six patterns. To facilitate this, define the corresponding centrality ratios,

$$m \equiv \frac{M_0}{M_1}; \quad \ell \equiv \frac{L_0}{L_1}; \quad h \equiv \frac{H_0}{H_1}; \quad n \equiv \frac{N_0}{N_1}; \quad w \equiv \frac{W_0}{W_1}; \quad r \equiv \frac{R_0}{R_1}. \tag{20}$$

Under symmetry, given an aggregate constraint, e.g. $2x_1 + x_0 = 1$, a spatial pattern is uniquely determined by one more piece of information, such as the ratio x_0/x_1.

For this linear city, the iceberg commuting cost matrix is

$$\begin{pmatrix} \tau_{-1,-1} & \tau_{-1,0} & \tau_{-1,1} \\ \tau_{0,-1} & \tau_{0,0} & \tau_{0,1} \\ \tau_{1,-1} & \tau_{1,0} & \tau_{1,1} \end{pmatrix} = \begin{pmatrix} 1 & \tau & \tau^2 \\ \tau & 1 & \tau \\ \tau^2 & \tau & 1 \end{pmatrix}, \tag{21}$$

where $\tau > 1$. To ease notation, we define

$$a \equiv \frac{\alpha\beta}{1-\alpha} > 0, \tag{22}$$

$$b \equiv \frac{(1-\alpha)(1+\varepsilon)}{\alpha\beta\varepsilon} = \frac{1+\varepsilon}{a\varepsilon}, \tag{23}$$

$$\phi \equiv \tau^{-\varepsilon}. \tag{24}$$

The parameter a is important enough for our analysis to merit discussion. Recall that α and β are the labor share of output and the housing share of consumption. Thus, the denominator of a is the commercial land share and the numerator is the induced residential land share in production. It follows that a measures the relative intensity of the production sector's demand for commercial versus residential land. As a increases, firms are relatively less reliant on commercial land, and conversely. Recalling that the periphery is land rich and the center is land poor, it is not surprising that a plays an important role in determining whether employment or residence is more centralized.

We will sometimes refer to $\phi \in [0, 1]$ as the *spatial discount factor*, which decreases with the level of commuting costs ($\tau \uparrow$) and increases with the heterogeneity of the population ($\varepsilon \downarrow$). Hence, ϕ

may be large (resp., low) because either commuting costs are low (resp., high), or the population is very (resp., not very) heterogeneous, or both.[8]

Applying the symmetry assumption to the wage and land rent patterns $\mathbf{W}$ and $\mathbf{R}$ and using (6) and (21), we get the following equilibrium conditions:

$$\begin{aligned} s_{0,0} &= \kappa R_0^{-\beta\varepsilon} W_0^{\varepsilon}, \\ s_{0,1} = s_{0,-1} &= \kappa R_0^{-\beta\varepsilon} W_1^{\varepsilon} \tau^{-\varepsilon}, \\ s_{1,1} = s_{-1,-1} &= \kappa R_1^{-\beta\varepsilon} W_1^{\varepsilon}, \\ s_{1,-1} = s_{-1,1} &= \kappa R_1^{-\beta\varepsilon} W_1^{\varepsilon} \tau^{-2\varepsilon}, \\ s_{1,0} = s_{-1,0} &= \kappa R_1^{-\beta\varepsilon} W_0^{\varepsilon} \tau^{-\varepsilon}. \end{aligned} \tag{25}$$

Setting

$$\omega \equiv w^{\varepsilon} \quad \text{and} \quad \mathcal{R} \equiv r^{-\beta\varepsilon}, \tag{26}$$

the first-order condition (19) becomes:

$$\mathcal{R} = \omega^{a}. \tag{27}$$

Clearly, $\omega > 1$ means that jobs at the central area pay a higher wage than those at the peripheries, while $\mathcal{R} > 1$ means that land at the central area is cheaper than that at the peripheries. By implication of (27), under constant returns, $\omega > 1$ if and only if $\mathcal{R} > 1$, that is, wages are higher at the center than in the peripheries if and only if the land rent is lower at the center.

Before we turn attention to a characterization of equilibrium, we first discuss the economic forces at work in this model. To illustrate ideas, consider the case when wages and land rents are the same in all nine locations, and the resulting indirect utility is $V_{ij} = V$. In this case, using (4), a household's discrete choice problem is

$$\max_{ij} \left\{ \begin{array}{ccc} z_{-1,-1}V, & \frac{z_{-1,0}}{\tau}V, & \frac{z_{-1,1}}{\tau^2}V \\ \frac{z_{0,-1}}{\tau}V, & z_{0,0}V, & \frac{z_{0,1}}{\tau}V \\ \frac{z_{1,-1}}{\tau^2}V, & \frac{z_{1,0}}{\tau}V, & z_{1,1}V \end{array} \right\}.$$

This is the standard way of stating a discrete choice problem, except that we have arranged the nine choices in a matrix so that the row choice corresponds to a choice of residence and choice of column to a choice workplace.

In this case, because the distribution of shocks is identical for all nine location pairs, the average payoff for a household choosing a central residence is

$$E\left(\max\left\{\frac{z_{0,-1}}{\tau}V, z_{0,0}V, \frac{z_{0,1}}{\tau}V\right\}\right) = \Gamma\left(\frac{\varepsilon-1}{\varepsilon}\right)\left(1+\frac{2}{\tau^{\varepsilon}}\right)^{1/\varepsilon} V. \tag{28}$$

Similarly, the average payoff for a household choosing one of the peripheral locations as a residence is

$$E\left(\max\left\{z_{-1,-1}V, \frac{z_{-1,0}}{\tau}V, \frac{z_{-1,1}}{\tau^2}V\right\}\right) = \Gamma\left(\frac{\varepsilon-1}{\varepsilon}\right)\left(1+\frac{1}{\tau^{\varepsilon}}+\frac{1}{\tau^{2\varepsilon}}\right)^{1/\varepsilon} V. \tag{29}$$

[8] It is common in the trade and geography literature with CES monopolistic competition to use a spatial discount factor defined by $\phi \equiv \tau^{-\sigma}$ where σ is the elasticity of substitution of the CES utility function and τ the trade cost. This spatial discount factor does not differ much from ours. Indeed, we know from the relationship between the CES and the MNL that $\mu = \sigma - 1$ where $\mu > 0$ decreases with the heterogeneity of the population (Anderson *et al.*, 1992). Therefore, a lower σ may be interpreted as a more heterogeneous population, very much like a lower ε.

Because $\tau > 1$, it follows that the average payoff for a household choosing the peripheral location is lower than that of an average household choosing the central location. By symmetry, the corresponding statement is also true for the choice of work location. As a result, the average payoff for a household choosing a peripheral work location is lower than that of an average household choosing the central location for work. In this sense, the structure of the discrete choice problem, together with purely psychic commuting costs, creates what we will call an *average preference* for residence in the central location, and a similar average preference for work in the central location.

This is noteworthy for the following reasons. First, even in the absence of more familiar agglomeration effects operating through production, this model has *two* 'agglomeration' forces, the average preference for central work and the average preference for central residence. Second, these average preferences are not agglomeration forces in the conventional sense. They do not incentivize geographic concentration, rather they incentivize concentration in the central location. Third, the urban economics literature based on homogenous agents takes seriously the possibility of differential labor productivity across locations. However, the possibility of *preferences* over work locations is never considered. This is a feature of quantitative spatial models without analog in the urban economics literature. Fourth, on the basis of the existing literature, we expect that the preference for central residence will be capitalized into higher central land rents. Extrapolating from this intuition suggests that the preference for central work location will be capitalized into lower wages. We will see that this turns out to be the case.

Against the two centralizing forces of average preferences are set two centrifugal forces. There is twice as much land in the periphery as the center. Because land contributes to utility and productivity, the scarcity of central land incentivizes the movement of employment and residence to the periphery.

We can now guess at the form of a symmetric equilibrium under constant returns to scale. In equilibrium, the centrifugal force of land scarcity and the centralizing force of average preferences balance. Access to the center will be scarce and so land rent will be higher and wages lower in the center. Whether the center ends up relatively specialized in residence or work will depend on which of the two activities has the highest demand for land, and this activity will locate disproportionately in the land abundant periphery. We postpone a discussion of equilibrium under increasing returns to scale until after we provide a more formal characterization of such equilibria.

5 Existence of equilibria

We now turn to the characterization of symmetric equilibria in our three location city. We proceed in three main steps. In the first, we derive the demand for housing, the demand for commercial space, the supply of workers, and the supply of residents as functions of ω and $\mathcal{R}$. This done, we derive a system of two equations in wages and land rent that characterize equilibrium. In the third step, we solve this system of equations.

The next proposition shows that in equilibrium all variables can be expressed in terms of $\mathcal{R}$ and ω.

Proposition 1 *The equilibrium demand for housing and commercial space and the equilibrium*

supply of workers and residents are:

$$M_0 = \frac{\mathcal{R}(\omega+2\phi)}{\mathcal{R}(\omega+2\phi)+2(\phi\omega+1+\phi^2)}, \qquad M_1 = \frac{1-M_0}{2}, \tag{30}$$

$$L_0 = \frac{\omega(\mathcal{R}+2\phi)}{\omega(\mathcal{R}+2\phi)+2(\phi\mathcal{R}+1+\phi^2)}, \qquad L_1 = \frac{1-L_0}{2}, \tag{31}$$

$$H_0 = \frac{a\mathcal{R}(1+2\phi\omega^{-\frac{1+\varepsilon}{\varepsilon}})}{a\mathcal{R}(1+2\phi\omega^{-\frac{1+\varepsilon}{\varepsilon}})+\mathcal{R}+2\phi}, \qquad N_0 = 1-H_0, \tag{32}$$

$$H_1 = \frac{a(\phi\omega^{\frac{1+\varepsilon}{\varepsilon}}+1+\phi^2)}{a(\phi\omega^{\frac{1+\varepsilon}{\varepsilon}}+1+\phi^2)+\phi\mathcal{R}+1+\phi^2}, \qquad N_1 = 1-H_1. \tag{33}$$

Proof: See Appendix A.

The derivation of these functions involves algebraic manipulation of the equilibrium conditions. For example, each of M_i and L_j is derived from the expressions for the share of households choosing each workplace-residence pair ij. In one case, we sum over workplace-residence pairs with a common residence, and in the other over pairs with a common residence. The identical denominator for each of the four functions is simply the three-location version of the denominator on the right-hand side of equation (5). The expressions for H_i and N_j are more complicated because they must also satisfy the land market clearing condition (11). Notice that we write M_i, L_i, N_j, and H_i exclusively in terms of price ratios, r and w, rather than the prices themselves, R_i and W_j. This simplifies our effort to characterize equilibrium because we need to solve for only two equilibrium quantities instead of four.

The functions described in Proposition 1 permit the straightforward evaluation of comparative statics. For example, when the relative land price at the center decreases ($\mathcal{R}$ increases), central residents and their land consumption (M_0 and H_0) both increase, with opposite effects in the periphery. Similarly, a decrease in $\mathcal{R}$ implies an increase in central employment, L_0, and a decrease in N_0 as the production sector substitutes away from relatively expensive central land toward relatively inexpensive central labor and peripheral land. An increase in ω requires an increase in the ratio of the central to the peripheral wage and has an effect similar to a decrease in $\mathcal{R}$. As ω increases, the relative wage at the center increases. As this happens, L_0 decreases, N_0 increases and H_0 increases, i.e., the production sector substitutes land for labor and households consume more land. Effects in the periphery are opposite. These comparative statics do not depend on the values of the structural parameters α, β, γ, ε, and τ.

We now turn to finding the equilibrium values of ω and $\mathcal{R}$. We would like to derive a system of equations involving only ω and $\mathcal{R}$ that incorporates all of the equilibrium conditions given in definition 1. The following proposition describes such a system.

Proposition 2 *Assume $\gamma \neq \alpha/\varepsilon$. Then, a pair $(\mathcal{R}^*, \omega^*)$ is an interior equilibrium if and only if it solves the following two equations:*

$$\omega^{\frac{1+\varepsilon}{\varepsilon}} = f(\mathcal{R}) \equiv \frac{\phi\mathcal{R} - 2a\phi\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + (1+\phi^2)(1+a)}{(1+a)\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + 2\phi\mathcal{R}^{\frac{1}{\beta\varepsilon}} - a\phi}, \tag{34}$$

$$\omega^{\frac{1+\varepsilon}{\varepsilon}} = g(\mathcal{R};\gamma) \equiv \mathcal{R}^{\frac{b}{1-\gamma\varepsilon/\alpha}} \left(\frac{\mathcal{R}+2\phi}{\phi\mathcal{R}+1+\phi^2}\right)^{\frac{\gamma\varepsilon/\alpha}{1-\gamma\varepsilon/\alpha}\frac{1+\varepsilon}{\varepsilon}}. \tag{35}$$

Proof: See Appendix B.

Combining (34)-(35), we arrive at a single equation, entirely in terms of $\mathcal{R}$, that determines the equilibrium. Thus, studying the equilibrium behavior of our discrete linear city amounts to studying the solution(s) of the equation,

$$f(\mathcal{R}) = g(\mathcal{R}; \gamma). \tag{36}$$

In particular, we can show the existence of an equilibrium by showing that (36) has a positive solution $\mathcal{R}^*$. Likewise, we can determine the number of possible equilibria by determining the number of positive solutions of equation (36).

This argument requires two comments. First, equation (36) is not defined when $\gamma = \alpha/\varepsilon$, so that we will need a specific argument for this case. Second, we will see that $\gamma = \alpha/\varepsilon$ is a threshold value of γ, below which equilibrium is unique, and above which multiple equilibria may emerge. To ease exposition of equilibrium behavior around this threshold, define $\gamma_m \equiv \alpha/\varepsilon$.

We now turn to a characterization of equilibrium. To begin, we establish the properties of the functions f and g. Observe that the function f does not involve the parameter γ, and thus remains the same for all values of γ. The following lemma states the main properties of function the f that are important for the characterization of spatial equilibria.

Lemma 1 *The function $f(\mathcal{R})$ has a vertical asymptote at $\mathcal{R} = \mathcal{R}_0 \in (0, 1)$, decreases over $(\mathcal{R}_0, \mathcal{R}_1)$ and is equal to 0 at $\mathcal{R} = \mathcal{R}_1 > 1$, while $f(\mathcal{R}) < 0$ outside $[\mathcal{R}_0, \mathcal{R}_1]$.*

Proof: See Appendix C.

Unlike f, the function g varies with γ. We show in Lemma 2 that when $\gamma < \gamma_m$, g is an increasing function that converges to an increasing step function as $\gamma \nearrow \gamma_m$. When $\gamma > \gamma_m$, g is a decreasing function that converges to a step function as $\gamma \searrow \gamma_m$. In both limiting cases, the value $\mathcal{R}_L$ at which the step occurs solves

$$\mathcal{R}^{\frac{1}{a}} \frac{\mathcal{R} + 2\phi}{\phi\mathcal{R} + 1 + \phi^2} = 1. \tag{37}$$

Because the left-hand side of (37) increases with $\mathcal{R}$ and is equal to 0 (resp., ∞) when $\mathcal{R} = 0$ (resp., $\mathcal{R} \to \infty$), $\mathcal{R}_L$ is unique and $\mathcal{R}_L < 1$.

We note that the relationship between land rent and employment is surprisingly simple. Using (31) leads to

$$\ell = \omega \frac{\mathcal{R} + 2\phi}{\phi\mathcal{R} + 1 + \phi^2}.$$

Combining this with (35), this equation becomes

$$\ell = \left(\mathcal{R}^{\frac{1}{a}} \frac{\mathcal{R} + 2\phi}{\phi\mathcal{R} + 1 + \phi^2} \right)^{\frac{\gamma_m}{\gamma_m - \gamma}}. \tag{38}$$

If $\gamma < \gamma_m$, then ℓ increases over $(\mathcal{R}_0, \mathcal{R}_1)$ and $\ell(\mathcal{R}) > 1$ if and only if $\mathcal{R} > \mathcal{R}_L$. On the other hand, when $\gamma > \gamma_m$, the opposite holds: ℓ decreases over $(\mathcal{R}_0, \mathcal{R}_1)$ and $\ell(\mathcal{R}) > 1$ if and only if $\mathcal{R} < \mathcal{R}_L$. Unexpectedly (at least to us), we will see that ℓ may be smaller than 1 and need not rise with γ.

The following lemma provides a more formal statement of the relevant properties of g.

Lemma 2 *(i) If $\gamma \neq \gamma_m$, then $g(\mathcal{R};\gamma)$ is strictly positive and finite over $[\mathcal{R}_0, \mathcal{R}_1]$. (ii) If $\gamma < \gamma_m$, then g is increasing over $[\mathcal{R}_0, \mathcal{R}_1]$. (iii) If $\gamma > \gamma_m$, then g is decreasing over $[\mathcal{R}_0, \mathcal{R}_1]$. (iv) As γ converges to γ_m, we have:*

$$\lim_{\gamma \nearrow \gamma_m} g(\mathcal{R};\gamma) = \begin{cases} 0, & \mathcal{R} < \mathcal{R}_L; \\ \left(\frac{\mathcal{R}_L + 2\phi}{\phi \mathcal{R}_L + 1 + \phi^2}\right)^{-\frac{1+\varepsilon}{\varepsilon}} & \mathcal{R} = \mathcal{R}_L; \\ \infty, & \mathcal{R} > \mathcal{R}_L; \end{cases} \qquad \lim_{\gamma \searrow \gamma_m} g(\mathcal{R};\gamma) = \begin{cases} \infty, & \mathcal{R} < \mathcal{R}_L; \\ \left(\frac{\mathcal{R}_L + 2\phi}{\phi \mathcal{R}_L + 1 + \phi^2}\right)^{-\frac{1+\varepsilon}{\varepsilon}} & \mathcal{R} = \mathcal{R}_L; \\ 0, & \mathcal{R} > \mathcal{R}_L. \end{cases} \cdot \tag{39}$$

Proof: Part (i) follows from combining (35) with $0 < \mathcal{R}_0 < \mathcal{R}_1 < \infty$. Parts (ii) and (iii) are obtained by differentiating g with respect to $\mathcal{R}$. Part (iv) holds because $g(\mathcal{R};\gamma)$ may be rewritten as follows:

$$g(\mathcal{R};\gamma) = \left[\left(\mathcal{R}^{\frac{1}{a}} \frac{\mathcal{R} + 2\phi}{\phi \mathcal{R} + 1 + \phi^2}\right)^{\frac{\gamma_m}{\gamma_m - \gamma}} \cdot \left(\frac{\mathcal{R} + 2\phi}{\phi \mathcal{R} + 1 + \phi^2}\right)^{-1}\right]^{\frac{1+\varepsilon}{\varepsilon}}. \tag{40}$$

Q.E.D.

The four panels of figure 1 illustrate the functions f and g for various parameter values. In each panel, the horizontal axis is $\mathcal{R}$ and the vertical axis is an increasing transformation of ω. The behavior of f, the red line in each panel, is relatively simple. It is a decreasing, continuous function that has a positive asymptote at $\mathcal{R}_0 < 1$ and declines monotonically to 0 at $\mathcal{R}_1 > 1$. The behavior of g is more complicated. The two panels on the left each describe g for three different values of γ, with dark blue the smallest, light blue the largest, medium blue in between, and all three less than γ_m. In every case, g is a continuous, increasing function. The right column is the same as the left, but considers larger values of γ. Here, the light blue line traces g for the smallest value of γ, dark blue uses the largest value, medium blue is intermediate value, and all three are greater than γ_m.

This figure makes clear that, in general, f and g cross for a positive value $\mathcal{R}$, and when $\gamma > \gamma_m$, they may cross more than once. On this basis, we can surmise that an equilibrium exists in our model city throughout much or all of the parameter space. The following proposition confirms that this is the case.

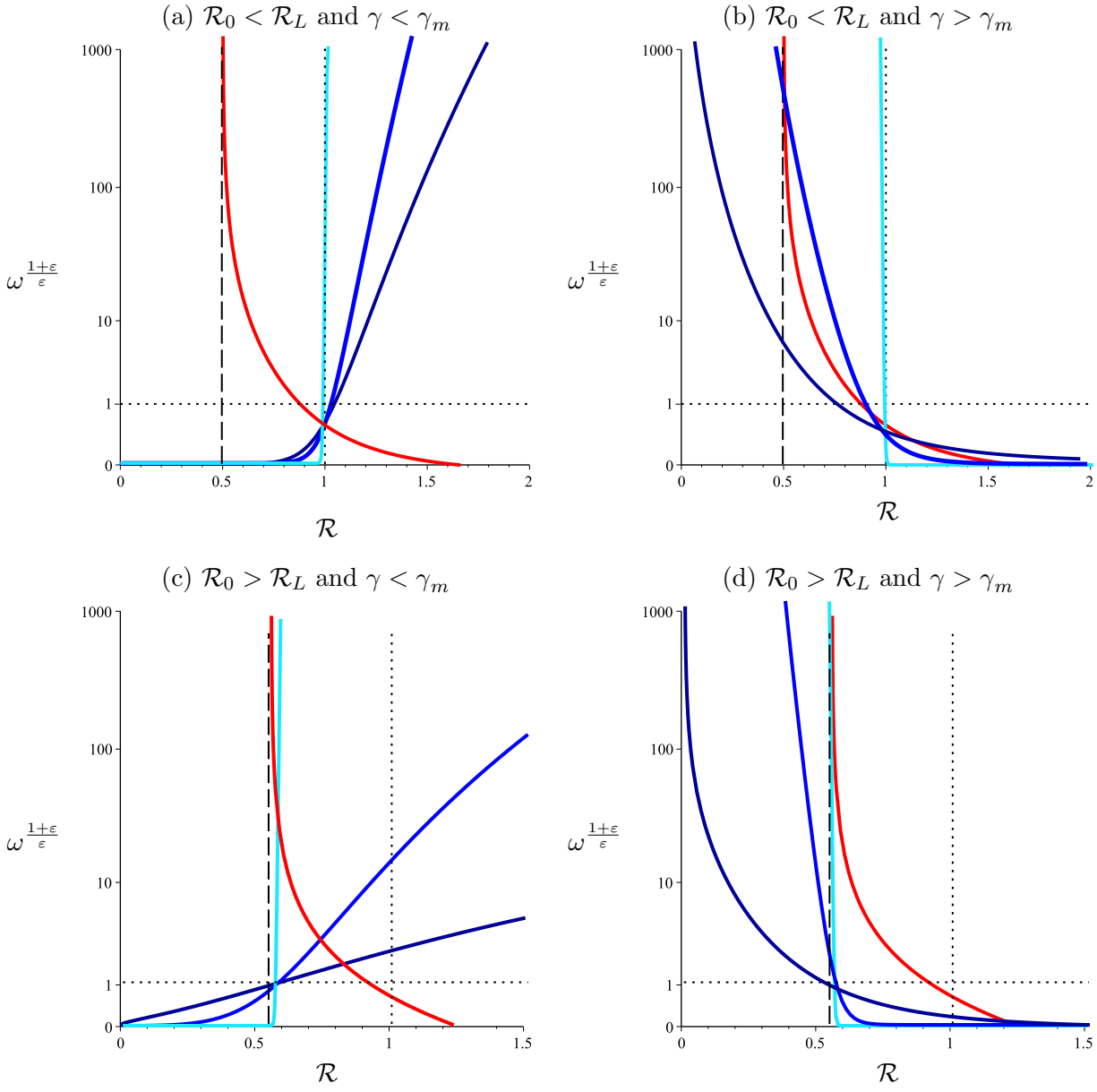
Proposition 3 *For all $\gamma \geq 0$, there exists an equilibrium.*

Proof: Using the properties of f and g given in Lemmas 1 and 2, the intermediate value theorem implies the existence of an equilibrium when $\gamma \neq \gamma_m$. The case where $\gamma = \gamma_m$ is considered in Propositions 6 and 7. Q.E.D.

6 Comparative statics

We now turn our attention to an investigation of how the equilibrium behavior of our city changes as fundamental characteristics of the economy change. More specifically, we focus on the

Figure 1: Graphical demonstration of equilibrium for a range of parameter values.



Notes: These figures illustrate equilibrium in several different cases. In all panels, f is given by the red line. The blue lines describe g. In the left two panels, darker colors of blue indicate smaller values of γ and in the right two panels darker colors of blue indicate larger values of γ.

implications of changes in returns to scale, γ, commuting costs, τ, and preference dispersion, ε.

6.1 Returns to scale

Lemma 1 shows that f has an asymptote at $\mathcal{R}_0 < 1$ and is zero at $\mathcal{R}_1 > 1$. Lemma 2 shows that as $\gamma \to \gamma_m$, g also approaches its asymptote at $\mathcal{R}_L < 1$. While Lemmas 1 and 2 guarantee that $\mathcal{R}_1 > \mathcal{R}_L$, they do not allow us to order $\mathcal{R}_L$ and $\mathcal{R}_0$. Unsurprisingly, the equilibrium configuration of our city depends sensitively on whether or not $\mathcal{R}_L > \mathcal{R}_0$. Therefore, as a precursor to our analysis of changes in γ, we determine necessary and sufficient conditions on fundamental parameters for $\mathcal{R}_L > \mathcal{R}_0$. More specifically, Lemma 3 tells us that the corresponding domains depend on commuting costs and the demands for commercial and residential land.

Lemma 3 *There exists a function $\overline{\phi}(\beta, \varepsilon) \in (0, 1)$ and scalar $\overline{a} > 0$ such that if $\phi < \overline{\phi}$ or $a < \overline{a}$ then $\mathcal{R}_0 < \mathcal{R}_L$. Conversely, if $\phi > \overline{\phi}$ and $a > \overline{a}$ then $\mathcal{R}_0 > \mathcal{R}_L$.*

Proof: See Appendix D.

Restating this lemma informally, we have $\mathcal{R}_0 < \mathcal{R}_L$ if either the spatial discount factor is low *or* the demand for commercial land is sufficiently large relative to the demand for residential land. Conversely, if the spatial discount factor is high *and* the demand for commercial land is low, then $\mathcal{R}_0 > \mathcal{R}_L$. We will see that the inequality $\mathcal{R}_0 \gtrless \mathcal{R}_L$ plays a key role in the determination of the city structure.

We now investigate changes in the equilibrium behavior of our city as returns to scale changes. To ease exposition, we introduce terminology to describe the three important domains of return to scale. These ranges will correspond to qualitatively different equilibrium behavior.

Definition 2 *Increasing returns to scale (IRS) are:*

i. weak $\Leftrightarrow 0 < \gamma < \gamma_m$,

ii. moderate $\Leftrightarrow \gamma_m < \gamma < \gamma_s$,

iii. strong $\Leftrightarrow \gamma > \gamma_s$, *where* $\gamma_s \in (\gamma_m, \frac{1+\varepsilon}{(1-\beta)\varepsilon} - \alpha]$.

Production involves constant returns to scale when $\gamma = 0$.

This definition requires two comments. First, population must be heterogeneous (ε finite) for weak returns to scale to occur. Hence, the behavior we observe in this domain of the parameter space cannot arise in models with the homogenous households that are standard in the urban economics literature. Second, these definitions indicate that the parameters α, β and γ are not sufficient to determine the regime to which the economy belongs. This also depends on ε. Because equilibrium behavior varies qualitatively with the type of returns to scale, it follows that *qualitative features of the equilibrium outcome depend on the degree of preference heterogeneity.*[9]

[9]This concurs with models of spatial competition models where firms locations depend on the degree of consumer heterogeneity (Anderson *et al.*, 1992, ch.7).

Figure 1 summarizes Lemmas 1-3 graphically. The top two panels describe cities where $\mathcal{R}_0 < \mathcal{R}_L$, that is, an economy where commuting costs are high or land is valuable in production. The bottom two panels describe an economy where $\mathcal{R}_0 > \mathcal{R}_L$, that is, a city where commuting costs are low and land has a low value in production. The two panels on the left describe cities where production involves constant or weak increasing returns to scale. The right two panels describe cities where increasing returns are moderate or strong. Each panel evaluates g for three different values of γ. In every panel, the light blue line describes g for a value of γ close to the weak/moderate threshold γ_m, and darker blue lines describe g functions for values of γ that are progressively further from this threshold, smaller in panels (a) and (c) and larger in panels (b) and (d). Taken together, these figures permit a fairly complete, if informal, description of the equilibrium behavior of our linear city with three locations.

Begin with the case of constant or weak increasing returns to scale illustrated in panels (a) and (c). In panel (a), $\mathcal{R}_0 < \mathcal{R}_L$ so that high commuting costs encourage households to work where they live or land hungry production faces pressure to disperse to the periphery (or both). We expect an equilibrium in such an economy to exhibit low levels of commuting and dispersed production. In fact, regardless of γ, f and g cross near $\mathcal{R} = 1$ and at a moderate value of ω. Since $\mathcal{R}_0 < \mathcal{R}_L$, as $\gamma \nearrow \gamma_m$ and g approaches its asymptote at $\mathcal{R}_L$, it does so when f is well away from its asymptote at $\mathcal{R}_0$. Hence, the equilibrium value of ω is relatively low. In this economy, the ratio of central and peripheral wages can never grow too large. Thus, panel (a) describes a city where employment and residence are not highly concentrated in the center or the periphery.

In contrast, panel (c) shows an economy where $\mathcal{R}_L < \mathcal{R}_0$. In this economy, low commuting costs allow households to separate work and residence locations in response to a small wage premium, and land is not very productive in production, so productivity is not sensitive to the relatively abundant land of the periphery. In such an economy, we expect that increasing returns to scale compounds the average preference for central employment to concentrate employment in the center, as households are able to cheaply disperse their residences to the land abundant periphery. In this city, as γ approaches γ_m, it does so near the asymptote of f. As a result, the value of ω at which the two curves intersects becomes large. This reflects that fact that peripheral productivity is falling almost to zero as production concentrates in the center and consumes progressively less peripheral land, even as central productivity rises with agglomeration economies. Panel (c) describes a city where commuting costs are low enough and the centrifugal forces on production weak enough for the average preference for central employment and increasing returns to scale complement each other to concentrate employment in the center.

Propositions 4, 5 and 6, presented below, formalize and refine this intuition about equilibrium with constant returns to scale or weak increasing returns to scale, and establish its generality.

Panels (b) and (d) of Figure 1 illustrate equilibria with moderate or strong increasing returns. As returns to scale increase beyond γ_m, behavior becomes more complicated. In panel (b) we again consider the city where $\mathcal{R}_0 < \mathcal{R}_L$ so that commuting is costly or land has a relatively high return in production. To begin, consider the case of moderate increasing returns described by the medium blue line. At this value of γ, g crosses f three times, first from below, then from above, and then again from below. At the first intersection point, we have $\mathcal{R}_1^* < 1$ and $\omega_1^* > 1$; at the second one, we see that $\mathcal{R}_2^*$ approaches $\mathcal{R}_L$ as γ decreases toward γ_m; at the third intersection point, we have $\mathcal{R}_3^* > 1$ and $\omega_3^* < 1$. and once near where it crossed when γ was close to but below γ_m. As equilibria are stable if and only if g crosses f from below are stable (see Appendix E), the intermediate equilibrium is unstable. Thus, when γ increases across the threshold from weak to

moderate increasing returns to scale, the equilibrium with decentralized production and low levels of commuting becomes unstable. In its place, two new equilibria arise with extreme values of ω. These extreme values of ω, in turn require that employment occur primarily in either the center or be divided between the two peripheral locations.

The light blue line in panel (b) describes g when γ is just above γ_m. As gamma approaches this threshold, for one of the two stable equilibria ω grows without bound (and occurs outside the frame of the figure) while ω approaches zero ever more closely in the other stable equilibrium. That is, just above the threshold, the stable equilibria approach corner patterns where all employment is either central or peripheral.

The dark blue line in panel (b) describes g when returns to scale are strong. As γ increases, the stable equilibrium with high ω converges toward the unstable equilibrium and when γ increases above the threshold γ_s between strong and moderate returns to scale, it disappears. The only remaining equilibrium involves a moderate value of ω, that is, an equilibrium where employment is evenly distributed across the three locations. As γ increases, households get wealthier, consume more of the numéraire good, and become less sensitive to the cost of commuting. As this occurs, they bid away differences in land rent and distribute themselves more uniformly across the three locations. Thus we arrive at the surprising conclusion that a low degree of increasing returns leads to agglomeration while a high degree fosters dispersion.

Summing up, in a city with $\mathcal{R}_0 < \mathcal{R}_L$, as γ increases, the city converges to an interior pattern in which employment does not become highly concentrated in the center or the periphery. When returns to scale increases beyond the weak/moderate threshold, this equilibrium becomes unstable and two new equilibria arise. These equilibria involve extreme concentration of employment in the center or in the periphery. As γ increases further, these two extreme equilibria become flatter, and ultimately, when γ crosses the moderate/strong threshold, we are left with a single equilibrium where employment is not highly concentrated in the center or periphery.

Panel (d) illustrates what occurs when returns to scale are moderate or strong in an economy where commuting costs are low and the productivity of land in production is relatively low ($\mathcal{R}_0 > \mathcal{R}_L$). The medium blue line illustrates the case when γ is moderate. Here we see that f crosses g only once and this crossing occurs when ω is close to zero. For such an economy, there is a unique equilibrium, and in this equilibrium most employment occurs in the peripheral locations. As γ decreases toward the threshold, a second equilibrium arises. This equilibrium involves a large value of ω, and hence, employment highly concentrated in the center. As γ increases so that returns to scale are strong, ω gradually increases, just as panel (b) where $\mathcal{R}_0 < \mathcal{R}_L$.

Recalling that just below the γ_m threshold, this city had employment concentrated in the center, we can describe equilibrium throughout the range of γ. As returns to scale increases from low levels, employment is more and more concentrated in the center. When returns to scale cross the weak/moderate threshold this highly centralized equilibrium persists, but a second equilibria with employment concentrated in the periphery arises. As returns to scales increases away from the weak/moderate threshold, the equilibrium with employment concentrated in the periphery persists, but the one with centralized employment does not. As γ increases further, the distribution of employment becomes more uniform as wealthier households arbitrage away small differences in land price.

Propositions 7, 8 and 9 confirm and refine this intuition about equilibrium with moderate and strong returns to scale, and its generality.

Before we turn to a more formal presentation of our results, we draw attention to three findings that we find most surprising. First, in some parts of the parameter space crossing the γ_m threshold involves the movement of nearly all employment from the center to the periphery, and in others from an interior equilibrium to one where nearly all employment is at the center or the periphery. Everywhere in the parameter space, equilibrium is always discontinuous at the γ_m threshold. Second, multiple equilibria may arise in the range of moderate γ, but in parts of the parameter space with low commuting costs or land that is relatively unproductive, equilibrium is unique for any level of returns to scale. Finally, everywhere in the parameter space, for returns to scale strong enough, the city has a unique equilibrium outcome, and this outcome no longer involves highly concentrated employment in either the center or the periphery.

We now turn to a formal statement of our results.

6.1.1 Constant returns to scale

Under constant returns to scale equation (35) becomes $g(\mathcal{R};0) = \mathcal{R}^b$ so that (36) may be rewritten as,

$$\omega^{\frac{1+\varepsilon}{\varepsilon}} = \mathcal{R}^b. \tag{41}$$

Since $g(\mathcal{R};0)$ increases from 0 to 1 as $\mathcal{R}$ increases over $[0,1]$, Lemma 1 implies that the two curves must cross exactly once. Furthermore, the intersection must occur strictly between $\mathcal{R}_0$ and 1, which implies $0 < \mathcal{R}^* < 1$. Since $g < 1$ over this interval, it must be that $\omega^* < 1$, while (41) implies $\omega^* > 0$.

Proposition 4 *Under constant returns to scale, there exists a unique equilibrium. Furthermore, this equilibrium is such that* $0 < \mathcal{R}^* < 1$ *and* $0 < \omega^* < 1$. *Furthermore, if*

$$\frac{\varepsilon}{1+\varepsilon} < \frac{1-\alpha}{\alpha\beta} \tag{42}$$

holds, then the equilibrium employment pattern is bell-shaped.

Proof: The first part follows from the above. For the second part, see Appendix F. Q.E.D.

The inequalities $0 < \mathcal{R}^* < 1$ and $0 < \omega^* < 1$ imply that all rents and wages are positive and finite. Furthermore, the equilibrium land rent is higher in the center while the equilibrium wage is lower. Since condition (42) is equivalent to $b > 1$, and a bell-shaped employment pattern is equivalent to $\ell^* > 1$, while (42), we have that $b > 1$ implies $\ell^* > 1$. It may seem surprising that (42) does not involve τ. We believe this reflects the fact that this condition is sufficient for $\ell^* > 1$, but not necessary.

We can refine our understanding of how the average preference for central work and residence affect equilibrium by looking carefully at the expressions for residential population and housing (30)-(33). Assume that wages and land rents are equal across locations so that $\omega = \mathcal{R} = 1$.

Substituting in the expressions for M_i and L_i we find that $M_0(1) = L_0(1) > M_1(1) = L_1(1)$. Therefore, central employment and residence is greater than in the periphery even though the different locations have the same relative pecuniary appeal. This reflects the average preference for central residence and workplace.

6.1.2 Weak increasing returns

The next two propositions describe the equilibrium when increasing returns to scale are weak.

Proposition 5 *Assume that $0 < \gamma < \gamma_m$. Then, there is a unique spatial equilibrium and $\mathcal{R}^* < 1$. Furthermore, if (42) holds, then the equilibrium employment pattern is bell-shaped and such that*

$$\frac{d\ell^*}{d\gamma} > 0 \qquad \frac{d\mathcal{R}^*}{d\gamma} < 0 < \frac{d\omega^*}{d\gamma}.$$

Proof: Under weak IRS, given Lemmas 1 and 2, f and g must intersect exactly once and because $f(1) < 1 < g(1;\gamma)$, the intersection must occur at $\mathcal{R}^* < 1$. This proves the first part of the proposition. See Appendix G for a proof of the second part. Q.E.D.

The substance of the proposition seems intuitive. *As scale economies increase, the central business district attracts more workers.* In addition, both the relative land price and relative wage increase as the land rent and wage at the center capitalize the agglomeration force resulting from increasing scale economies.

The comparative statics in Proposition 5 hold whenever $0 < \gamma < \gamma_m$. That is, whenever increasing returns to scale are weak. This somewhat surprising because, consistent with our earlier discussion of panels (a) and (c) of Figure 1, there are two distinct types of equilibrium behavior with weak increasing returns to scale, depending on whether $\mathcal{R}_0$ is larger or smaller than $\mathcal{R}_L$. The following proposition formalizes our earlier discussion and describes how the two different types of equilibrium diverge as γ approaches γ_m.

Proposition 6 *Assume γ is slightly below γ_m. Then,*

i. If $\mathcal{R}_0 < \mathcal{R}_L$ as $\gamma \nearrow \gamma_m$, the equilibrium wage ratio remains bounded and the limiting employment pattern remains interior.[10]

ii. if $\mathcal{R}_0 > \mathcal{R}_L$ as $\gamma \nearrow \gamma_m$, then $\omega^* \to \infty$ while the equilibrium employment pattern converges to $(0, 1, 0)$.

This proposition confirms that two distinct regimes are possible. When commuting costs are high, the full concentration of production in the central location requires prohibitively costly commuting. When commuting costs are low, almost full concentration of employment at the CBD occurs as long as the productivity of land in production is low enough to stop the migration of firms to the land abundant periphery. In this case, increasing returns become strong enough to

[10] This pattern is bell-shaped when (42) holds.

complement the preference for central employment and generate almost the full concentration of jobs at the center. This intuition seems straightforward, although the possibility of a corner solution when the support of the taste shocks is unbounded is surprising.

6.1.3 Moderate increasing returns

When $\gamma > \gamma_m$, the function f remains unchanged, but the function g changes from an increasing to a decreasing function. When both f and g are decreasing, we are no longer assured of the existence of a unique equilibrium. As shown in Appendix E, a solution $\mathcal{R}^*$ of (36) is stable if and only if f crosses g from above at $\mathcal{R}^*$.

The following proposition formalizes the intuition suggested by the figures confirms its generality.

Proposition 7 *Assume γ is slightly above γ_m. Then,*

i. $\mathcal{R}_0 < \mathcal{R}_L$ there exist two stable interior equilibria, $(\mathcal{R}_1^*, \omega_1^*)$ and $(\mathcal{R}_3^*, \omega_3^*)$, such that

$$\omega_1^* > 1 > \omega_3^* \qquad \text{and} \qquad \mathcal{R}_1^* < 1 < \mathcal{R}_3^*,$$

as well as an unstable interior equilibrium. Furthermore, as $\gamma \searrow \gamma_m$, the stable equilibrium employment patterns converge to, respectively, $(1/2, 0, 1/2)$ and $(0, 1, 0)$.

ii. $\mathcal{R}_0 > \mathcal{R}_L$ there exists a unique interior equilibrium $(\mathcal{R}^*, \omega^*)$ such that:

$$\omega^* < 1 \qquad \text{and} \qquad \mathcal{R}^* > 1.$$

Furthermore, as $\gamma \searrow \gamma_m$, the equilibrium employment pattern converges to $(1/2, 0, 1/2)$.

Proof: See Appendix H.

This proposition is the counterpart of Proposition 6 when γ converges toward γ_m from above rather than from below. Proposition 7 establishes the existence of multiple stable equilibria when either commuting costs are high or the productivity of land in production is relatively low. When γ exceeds γ_m, the interior equilibrium, which was stable under weak IRS, becomes unstable while two new stable equilibria emerge. Formally, for any γ slightly larger than γ_m, there exists an arbitrary small $\Delta > 0$ such that the equilibrium employment patterns are given respectively by $((1-\Delta)/2, \Delta, (1-\Delta)/2)$ and $(\Delta/2, 1-\Delta, \Delta/2)$. One of these equilibria involves *agglomeration in the central location* while the other leads to *partial agglomeration in the two peripheries*. Both involve different types of agglomeration.

As shown by Propositions 6 and 7, there is a bifurcation at $\gamma = \gamma_m$, which is driven by the discontinuity of function g with respect to γ (see Lemma 2). This result may come as a surprise because assuming heterogeneous agents is generally sufficient to smooth out payoffs (see, e.g., Anderson *et al.*, 1992). We have seen in Lemma 2 that $\mathcal{R}_L$ is the only admissible value of $\mathcal{R}$ for the function g to be defined at $\gamma = \gamma_m$. This makes (R_L, ω_L) a natural candidate equilibrium where $\omega_L = [f(R_L)]^{\frac{\varepsilon}{1+\varepsilon}}$. However, for this to hold, $\mathcal{R}_L$ must belong to the interval $(\mathcal{R}_0, \mathcal{R}_1)$. We know from the proof of Lemma 1 that $\mathcal{R}_1 > 1$. Since $\mathcal{R}_L < 1$, $\mathcal{R}_L < \mathcal{R}_1$ always holds. However,

the comparison of $\mathcal{R}_0$ and $\mathcal{R}_L$ is less straightforward because $\mathcal{R}_0$ is also smaller than 1 (see Lemma 1). If $\mathcal{R}_L > \mathcal{R}_0$, then $(\mathcal{R}_L, \omega_L)$ is the only interior equilibrium. Otherwise, there is no interior equilibrium. Also, there are two "corner" equilibria where the employment patterns are respectively given by $(1/2, 0, 1/2)$ and $(0, 1, 0)$.

Having studied the equilibria for γ slightly above γ_m, we now turn our attention to the values of γ that belong to the whole domain where moderate IRS prevail.

Proposition 8 *There is a unique value $\gamma_s > \gamma_m$ such that*

i. if $\mathcal{R}_0 < \mathcal{R}_L$, then three equilibria exist, with $\mathcal{R}_1^* < \mathcal{R}_2^* < \mathcal{R}_3^*$, for all $\gamma \in (\gamma_m, \gamma_s)$; the equilibria $\mathcal{R}_1^*$ and $\mathcal{R}_3^*$ are stable whereas $\mathcal{R}_2^*$ is unstable;

ii. if $\mathcal{R}_0 > \mathcal{R}_L$, then there exists a unique equilibrium.

Proof: TBD.

Hence, multiple stable equilibria to arise throughout the range on moderate increasing returns to scale when commuting costs are high or the demand for commercial land is high, or both.

As γ rises above γ_s, we enter a third domain that is described below.

6.1.4 Strong increasing returns

The above discussion shows that only a single equilibrium persists when γ exceeds γ_s. More specifically, we have:

Proposition 9 *If $\gamma > \frac{1+\varepsilon}{(1-\beta)\varepsilon} - \alpha \geq \gamma_s$, then there exists a unique spatial equilibrium.*

Proof: See Appendix J.

To our knowledge, this result is new to the literature. While it has long been understood that increasing returns to scale could lead to multiple equilibria, the idea that sufficiently high increasing returns leads, once again, to unique equilibrium is novel. With this said, the intuition behind this result seems straightforward. As γ increases, all else equal, wages must rise. Some of this extra income must be devoted to consumption. This increases the marginal utility of land, which increases a households' willingness to move in response to difference in land prices. As this occurs, employment follows workers and reduces the wage gap across locations.

6.2 Commuting cost and preference dispersion

We now discuss what our result become when the spatial discount factor converges to 1 or to 0. Recalling that $\phi = \tau^{-\varepsilon}$, it is straightforward that (i) $\phi \to 1$ as $\tau \to 1$ or $\varepsilon \to 0$ and that (ii) $\phi \to 0$ as $\tau \to \infty$ or $\varepsilon \to \infty$. When ϕ goes to either zero or one, the average preference for central work

and residence weakens. One can see this effect even more clearly in (28) and (29). These two expressions evaluate the average payoff from the choice of central workplace versus peripheral workplace or the corresponding payoff from choices of residential locations. These payoffs clearly converge toward each other when ϕ goes to either zero or one.

Setting $\phi = 0$ and $\phi = 1$ in (34) and (35), we obtain:

$$\begin{aligned} f(\mathcal{R})|_{\phi=0} &= \mathcal{R}^{-1-\frac{1}{\beta\varepsilon}}, \qquad f(\mathcal{R})|_{\phi=1} = \frac{\mathcal{R} - 2a\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + 2(1+a)}{(1+a)\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + 2\mathcal{R}^{\frac{1}{\beta\varepsilon}} - a}, \\ g(\mathcal{R};\gamma)|_{\phi=0} &= \mathcal{R}^{\left(\frac{1}{a}+\frac{\gamma}{\alpha\varepsilon}\right)\frac{\gamma\varepsilon/\alpha}{1-\gamma\varepsilon/\alpha}\frac{1+\varepsilon}{\varepsilon}}, \qquad g(\mathcal{R};\gamma)|_{\phi=1} = \mathcal{R}^{\frac{b}{1-\gamma\varepsilon/\alpha}}. \end{aligned}$$

Evaluating each of these functions at $\mathcal{R} = 1$ shows that $f(1)|_{\phi=0} = g(1;\gamma)|_{\phi=0} = 1$ and $f(1)|_{\phi=1} = g(1;\gamma)|_{\phi=1} = 1$. Hence, in both cases $\mathcal{R}^* = 1$ is an equilibrium. As $\varepsilon \to 0$, hence $\gamma_m \to \infty$, this equilibrium is unique because the only possible case is the weak increasing returns case. When $\varepsilon \to \infty$, plugging $\phi = 0$ in (41) yields

$$\frac{1}{\mathcal{R}} = \mathcal{R}^{1/a},$$

whose unique solution is also given by $\mathcal{R}^* = 1$. Thus, $\omega^* = (\mathcal{R}^*)^{1/a} = 1$, when $\varepsilon \to \infty$, which implies $\ell^*(1) = m^*(1) = h^*(1) = n^*(1) = 1$. Consequently, using (30) and (31) implies that, when the population is homogeneous, the equilibrium involves *backyard capitalism*, i.e., each individual works and resides at the same place. In other words, *the equilibrium involves working from home.* This equilibrium is such that $L_i^* = M_i^* = 1/3$ for $i = -1, 0, 1$ and $s_{ij}^* = 0$ for $i \neq j$ and $s_{ii}^* = 1/3$. In addition, land is shared between consumption and production according to the same proportion $H_i^*/N_i^* = a = \alpha\beta/(1-\alpha)$ across locations, which increases with the land share β in consumption and decreases with the land share $1-\alpha$ in production. What is more, land is shared between consumption and production according to the same proportion $H_i^*/N_i^* = a = \alpha\beta/(1-\alpha)$, which increases with the land share β in consumption and decreases with the land share $1-\alpha$ in production.

The same pattern emerges when the population is infinitely heterogeneous. Yet, the two situations are not completely identical. As $\varepsilon \to 0$, the payoffs associated with each of the nine location pair become identical. As $\varepsilon \to \infty$, the payoff (29) attached to off-diagonal pairs goes to zero. Even though each location hosts the same distribution of activities, the commuting patterns are opposite. In the former case, there is equal cross-commuting between any location pair whereas each location is autarchic in the latter.

Last, when $\tau \to 1$ or $\tau \to \infty$, the same result holds if the assumptions of Proposition 6 or 8 are satisfied.

The following proposition states the corresponding conditions in terms of model parameters for the different limits.

Proposition 10. *The equilibrium employment, residential, housing, and commercial patterns converge to a stable uniform pattern when one of the following condition holds:*

(i) $\tau \to 1$ *or* $\tau \to \infty$ *and* $\gamma < \gamma_m$ *or* $\gamma > \frac{1-\frac{1}{\varepsilon}}{1-\beta} - \alpha$;

(ii) $\varepsilon \to 0$;

(iii) $\varepsilon \to \infty$ *and* $\gamma > \frac{1}{1-\beta} - \alpha$.

The intuition behind two of the four cases is transparent. As $\varepsilon \to 0$, preference dispersion increases. As this happens, taste heterogeneity over pairwise choices becomes increasingly important relative to commuting costs and, in the limit, households ignore land price and wage differences, and the distribution of households across pairs is uniform. Likewise, as $\tau \to \infty$ the cost of commuting grows so high that it never makes sense to commute. Since the distribution of types is the same across locations, households must be uniformly distributed across locations. In this case, only autarchy is possible and all households work where they live. In both cases, the distribution of residence and employment is uniform, although city functions differently in the two cases. When $\varepsilon \to 0$ we have 'urban sprawl', where many people commute and there is no city center. When $\tau \to \infty$ we have a city of backyard home-workers.

As $\varepsilon \to \infty$, preference dispersion disappears and preferences become homogenous. Simultaneously, the range of weak returns to scale collapses toward $\{0\}$. When returns to scale are constant and agents are homogenous, then there is no longer a centralize or force in the model and the distribution of activity is uniform. The distribution of activity is also uniform when agents are homogenous and returns to scale are strong. This is less intuitive. When agents are homogenous, they will all change work and residence locations in response to a tiny change in wage or land rent. More intuitively, *once workers no longer care about where they live and work, high increasing returns make them rich enough for workers to avoid bearing the psychic commuting loss however small they are by residing and living at the same place.* Hence, in spite of increasing returns to scale, the only equilibrium involves a uniform distribution of productive activity across locations.

As $\tau \to 1$ something similar occurs. In the absence of increasing returns to scale, there is no reason for firms to concentrate, and so we arrive at a uniform distribution of activity. That this also occurs when increasing returns to scale are present is more puzzling. As shown by Panel (b) of Figure 2, m^* is larger than 1 because households earn a sufficiently high income to secure a large share of land in location 0 as a consequence of their average preference for the center (recall that ε is finite). As commuting costs fall sufficiently, the appeal of the center gets weak and the residential share at 0 decreases, while the employment share rises (see Figure 3). In the limit, the uniform pattern emerges.

We regard these comparative statics as among the more surprising and important that we find for two reasons. First, in the monocentric city model, decreases in commuting costs lead households to spread out. Our results contradict this intuition In our heterogenous households model, *comparative statics on commuting costs are not monotone.* Second, as we describe in our review of the literature, discrete heterogenous agent models, similar to ours, provide the basis for a rapidly growing quantitative literature. Often, such quantitative exercises evaluate the effects of counterfactual changes in commuting costs. To the extent that these counterfactual exercises are comparative statics of commuting costs, our results suggest the qualitative features of such counterfactual exercises may change sign in response to changes on other incidental parameters.

7 A simplified history of cities in the western world

We now compare the historical record to the evolution of our model as returns to scale increase. We begin with a summary of the history of urban form and returns to scale since 1500, and then compare this history to the behavior of our model.

We note that the discussion below omits any mention or analysis of other long run trends in the economy. Of these, decreases in transportation cost and in the land intensity of production seem of particular interest. We focus on returns to scale for three reasons. First, up until now, our understanding of the implications of changes in returns to scale has been restricted by the intractability of this problem: these results are our most novel. Second, changes in returns to scale alone seem to have a remarkable ability to explain qualitative features of the history of cities. Finally, as important as other changes have been, increases in productivity are surely more important. By focusing on changes returns to scale we keep attention on the central economic force behind the industrial revolution.

The existence of various types of agglomeration effects in modern cities is well established, even if debate continues about exactly how large they are. We do not have systematic evidence about time series change in the importance of agglomeration effects. Casual empiricism, however, strongly suggests that returns to scale are increasing. In particular, three pieces of evidence argue for increases in agglomeration economies since pre-industrial time.

First, the number and size of cities has increased steadily since around 1500, as has the share of urban population. During the period around 1500, de Vries (1984) reports 154 European cities with population above $10,000$, while Bairoch (1988) reports 89 cities of at least $20,000$. At this time, Europe (without Russia) had only between 10 and 12 cities of more than $100,000$ inhabitants. By 1800, the count of cities with a population of at least $10,000$ and $20,000$ increased to 364 and 194, respectively. Similarly, the share of the urban population was low and rose slowly from 10.7 in 1500 to 12.2 percent in 1750. The urbanization rate was still around 12 percent in 1800, but grew rapidly to 19 percent in 1850, 38 percent in 1900 and 48 percent in 1950 (Bairoch, 1988). Without a doubt, European cities have become progressively more attractive over time. While these increases are surely not entirely due to increases in returns to scale in production,[11] it is equally sure that stronger returns to scale are partly responsible.

Second, economic historians have documented many small changes in pre-industrial Europe that seem to have contributed to the productivity of cities. For example, Cantoni and Yuchtman (2014) observe the spread of Universities in 14th century Germany and argue that, by spreading knowledge of Roman law, these Universities contributed to larger and more productive market cities. Dittmar (2011) documents the spread of the printing press and finds that it contributed casually to the sizes of cities where it was introduced. Finally, de la Croix, Doepke and Mokyr (2018) argue that apprenticeship may have played a role similar to the one that Dittmar finds for the printing press. In all, the history of the pre-industrial period points to slow increases in the productivity of cities.

Finally, the nature of production has changed in a way that seems to argue for an increase in returns to scale. For the period prior to the industrial revolution, de Vries (1984) reports on the pervasiveness of proto-industrialization in manufacturing. Under this system, rural households

[11] See Nunn and Qian (2011), for example.

performed manufacturing, often of textiles, using materials provided by manufacturers. It is hard to imagine a system of industrial production that more strongly suggests constant or decreasing returns to scale in manufacturing.[12]

Summing up, although econometric estimates of the magnitude of agglomeration economies and increasing returns in the pre-industrial period are not available, we are confident that agglomeration economies are positive in the modern economy and the available evidence strongly suggests that they were small or zero in the pre-industrial period. Taken together these two observations indicate a trend upwards.

We now turn to a more ambitious undertaking, a stylized description of the evolution of urban form from the pre-industrial period to the present. de Vries (1984) describes cities in pre-industrial Europe as being organized as much for protection as production. Much of the population was employed in agriculture, either within city walls or without, manufacturing was at least as rural as urban, and the preponderance of urban residents were abjectly poor. While the absolute poverty of an average urban resident is clear, there is suggestive evidence from the US for the relative poverty of urban residents as well. Costa (1984) reports in a survey of US soldiers in the US civil war, early in the US industrial revolution, and finds that soldiers with urban backgrounds were shorter than those from rural backgrounds.

Clark (1951) is the first effort to provide systematic evidence about patterns of city density. In this landmark study, Clark collects historical census data for a collection of US and European cities. While the precise time period he considers varies from city to city, for the most part his data begins early in the industrial revolution and continues until the early 20th century. His findings are as dramatic as they are unequivocal. For the European and American cities in Clark's sample, center city density falls monotonically and the density gradient flattens over time.

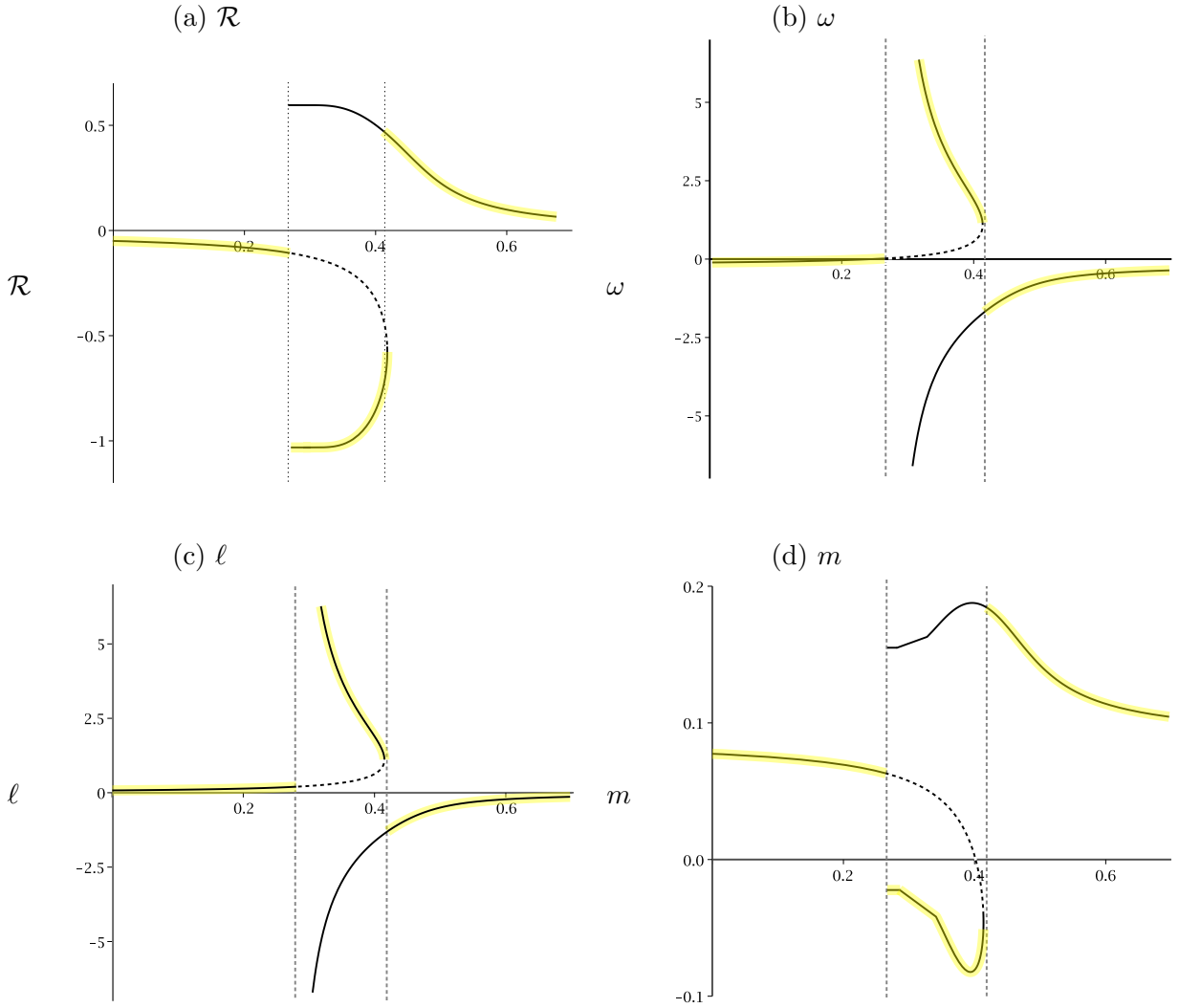
Turning to the late 20th and early 21st centuries, Baum-Snow (2007) documents the decline in US central city population from 1950 to 1990, a phenomena earlier described by Meyer, Kain and Wohl (1965).[13] Glaeser and Kahn (2004) document the changing spatial patterns of the late 20th century US. In particular, they find that employment follows decentralizing population and that peripheral urban residents of US cities tended to have much longer commutes than did central residents. Related to this, Garreau (1992) documents the rise of 'edge cities' in late 20th century US, while McMillen and MacDonald (JUE1998) provide econometric support for this idea using detailed data for Chicago. More recently, Couture and Handbury (2019) and Couture *et al.* (2020) document the a resurgence of central cities in the US.

Summing up, we arrive at the following stylized 500 year history of cities. During the pre-industrial period, population was dispersed and cities were small, poor and not obviously more productive than rural regions. In the early part of the Industrial Revolution, cities grew rapidly. Cities at this time had very dense centers and steep population density gradients. As the industrial revolution progressed, central city density decreased and the population density in peripheral suburbs increased. During the later part of the 20th century, we saw the rise of complicated poly-centric cities and longer commutes. In the 21st century there is suggestive

[12] We note the Cesaretti et al (2020) study finding that tax collection per capita increases with city size in England between 1450 and 1670. While this result seems relevant, it is unclear whether it reflects increases in the efficiency or effort of tax collectors, or increases in productivity of city residents, or simply the first nature advantages of larger pre-industrial cities.

[13] Angel-Garcia et al. (2015) document the corresponding phenomena in European cities.

Figure 2: Evolution of cities and returns to scale.



Notes: These figures illustrate how an equilibrium city evolves as γ increases. The solid dark lines indicate stable equilibrium outcomes, the dashed black lines indicate unstable equilibrium outcomes, and the highlighted portions indicate the particular equilibrium paths that we compare with history in the text. In all figures, the x-axis is γ and the y-axis is (a) $\ln(\mathcal{R})$, (b) $\ln(\omega)$, (c) $\ln(\ell)$, and (d) $\ln(m)$. In all four panels, the dashed vertical lines indicate the threshold values of γ, γ_m and γ_s. The parameter values on which these figures are based are $(\varepsilon, \tau, \alpha, \beta) = (3, 1.3, 0.8, 0.25)$.

evidence for a slow resurgence of central cities.

Let us now compare this stylized history of cities to the behavior of our model. Figure 2 illustrates all possible equilibrium outcomes for our model as γ increases in a city satisfying $\mathcal{R}_0 < \mathcal{R}_L$. That is a city where either commuting costs are sufficiently high, the population is not too homogeneous, or land is sufficiently valuable in production (in the precise sense of Lemma 3). There are four panels in the figure. Panel (a) describes the evolution of $\mathcal{R}$; (b) describes ω; (c) describes ℓ, the ratio of central to peripheral employment; and (d) describes m, the ratio of

central to peripheral residence. In all four panels, we apply a logarithmic transformation to the y-axis to increase legibility. This compresses large values and assigns values between zero and one to negative values.

Our object in figure 2 is to demonstrate qualitative features of the way that equilibria change as γ increases. With that said, we rely on parameter values that are consistent with the available evidence. In every panel, the figures are based on a city where the fundamental parameters are $(\beta, \alpha, \tau, \varepsilon) = (0.25, 0.8, 1.3, 3)$, and γ ranges from zero to 0.7. A housing share of consumption, β, equal to 0.25 is conventional and in line with modern empirical evidence on the issue (Davis and Ortalo-Mangé, 2011). A labor share of 0.8 is large relative to the observed labor share in a modern economy, however, our model omits capital, so we are effectively consolidating the labor and capital share to drive down the land share of production to 0.2. This is a compromise value. It is likely too high for a modern economy and too low for the more agricultural preindustrial world. Choosing τ is difficult since it reflects a purely psychic cost that cannot be observed directly. By choosing $\tau = 1.3$ we require that commuting from periphery to center results in about a 1/3 decline in welfare, all else equal. This is probably too large, but results in more legible figures than does a smaller value of τ. Similarly, $\varepsilon = 3$ is smaller than the values 5 or 6 suggested by Heblich, Redding and Sturm (2020),[14] but contributes to the legibility of our figures. Setting $\varepsilon = 5$ does not change the figures qualitatively, but expands variation along the y-axis.

The range of γ described by the x-axes of the four panels of figure 2 is 0 to 0.7. This range is large compared to modern estimates of agglomeration economies (around 0.05; see, e.g., Rosenthal and Strange, 2004). However, if we allow for any sort of compounding over the approximately 200 years since the industrial revolution began, considering values of γ up to 0.7 seems modest. The increase in returns to scale that we consider in figure 2 is large only in comparison to modern estimates of cross-sectional agglomeration effects. It is not large in comparison to any defensible measure of the increase in productivity between the 16th and 21st centuries. With this said, we repeat that our goal is to illustrate that the model can replicate qualitative features of history, not to approximate observed magnitudes.

In panel (a) we see that when γ is close to zero, $\mathcal{R}$ is less than one, and decreases gradually as γ increases. Recalling the definition of $\mathcal{R}$, this means that for low levels of γ, central land rent is modestly higher than peripheral, and this ratio increases with γ. When γ reaches the threshold value separating weak from moderate returns to scale, γ_m, two new equilibria with extreme values of $\mathcal{R}$ emerge. In these equilibria, either central land rent is much higher than peripheral (the negative value of $\ln(\mathcal{R})$) or conversely, and the unique equilibrium that prevailed under weak increasing returns becomes unstable. As γ increases further, central and peripheral land rents converge in both equilibria. Ultimately, when returns to scale reach the strong increasing returns threshold, γ_s, the more centralized equilibria (i.e., $\ln(\mathcal{R}) < 0$) converges with the unstable equilibria and disappears. For cities following this more centralized equilibrium path, this leads to a second discontinuity in which $\mathcal{R}$ increases and central land rents decrease discontinuously to below peripheral land rents. As γ increases still further, central land rents gradually increase as the central city recovers from the second discontinuity.

[14] We have seen that $\mu = \sigma - 1$ where μ is the heterogeneity parameter of the MNL and σ the elasticity of substitution of the CES (Anderson *et al.*, 1992). It is well known that the logarithm of a Gumbel-distributed variable is a Fréchet-distributed variable, so that $\mu = \varepsilon$. Therefore,we have $\varepsilon = \sigma - 1$. Since several estimations of the elasticity of substitution of the CES suggest values of σ that vary from 6 to 7 (**references**), it is reasonable to assume that ε varies from 5 to 6.

In panel (b) we show the corresponding evolution of ω. As γ increases from 0 to γ_m, $\ln(\omega)$ increases from slightly below to slightly above 0, so that ω increases from slightly below (as required by Proposition 4) to slightly above one. Thus, at sufficiently low levels of returns to scale, the gap between central and peripheral workers is small, though peripheral workers are better off for γ near zero. As γ increases past γ_m, the weak increasing returns to scale equilibrium becomes unstable and two extreme equilibria emerge. In one, central wages are much larger than peripheral wages. In this equilibria central land rent is also high, so that the positive equilibrium values of $\ln(\omega)$ correspond to the equilibria with negative values of $\ln(\mathcal{R})$. Conversely, along the second equilibrium path, peripheral wages are much higher than those in the center. As γ continues to increase, central and peripheral wages converge in both stable equilibria. When γ reaches the strong increasing returns to scale threshold, the more centralized equilibrium disappears and a city on this path experiences a discontinuous decrease in ω. After this drop, peripheral wages are higher than central wages. As γ increases still further, the central wages gradually recover.

Panel (c) shows the evolution of the logarithm of ℓ, the ratio of central to peripheral employment. Unsurprisingly, the equilibrium behavior of this ratio strongly resembles that of the ω (a monotone transformation of the wage ratio). There is one noteworthy divergence. Both ω and ℓ are close to one and increase slowly as γ increases from zero. However, ℓ is always strictly greater than one in this range, so that there is always more employment in the center than the periphery, while ω can be below one for γ sufficiently close to zero.

Panel (d) shows the evolution of the logarithm of m, the ratio of central to peripheral residence. For low levels of γ, residence is concentrated in the center. As γ increases, this ratio gradually decreases as central employment gradually increases.

At the moderate returns to scale threshold, the weak returns to scale equilibrium becomes unstable and two new equilibria emerge in which residence is either highly concentrated in the center or in the periphery. The equilibrium with residence concentrated in the periphery is also the equilibrium where central land rent, wages and employment are very high relative to the periphery. That is, in one of the two equilibria, something very like a monocentric 'mill town' emerges for values of γ just above γ_m. Employment is highly concentrated in the center, central rents and wages are high, and most residents live in the periphery.[15] As γ continues to increases, for both equilibria, residence becomes progressively more concentrated, as employment begins to disperse. Eventually, the path of residence begins to follow the path of employment. When γ reaches the strong returns to scale threshold, γ_s, cities with decentralized residence experience a discrete increase central residence. People move back into the city. As γ increases still further residence becomes less concentrated in the center as city's residence profile, like all other quantities in the figure, flattens out.[16]

We now consider the correspondence between the behavior of our model, and our stylized history of returns to scale and urban form. For the purpose of this exercise, let the pre-industrial period be a period of time during which returns to scale increase from zero to the threshold of moderate

[15] We note the similarity of the rapid centralization of employment and migration of residence to the periphery to the changes that Heblich, Redding and Sturm (2020) observe in 18th century London. In their model, this effect arises as a consequence of changes in transportation costs, while we derive it as a consequence of increases in returns to scale.

[16] Note that the behavior of m has proved more resistant to analysis than have $\mathcal{R}$, ω and ℓ, and in particular, none of our propositions address the behavior of m. Given this, we have less confidence in the generality of the behavior we see in panel (d) than for the other panels.

returns to scale, γ_m; let the period from the industrial revolution until the mid-20th century be a period of time during which returns to scale increases from the moderate returns to scale threshold to the point in the moderate returns to scale domain where central residence reaches its minimum; and finally, let the post-industrial period beginning in the late-20th century correspond to a period of time when returns to scale increase beyond the point where central residential population reaches its minimum.

With this correspondence in place, we can compare our stylized history to the behavior of our model. We consider the more centralized of the two possible equilibria in the moderate increasing returns to scale range of γ. During the preindustrial period, cities were small and grew slowly. Employment and residence both concentrated in city centers, but the differences between center and periphery were modest. Employment in the center was at most only slightly more productive than in the periphery.

At the beginning of the Industrial Revolution, city centers grew rapidly. They became sites of highly concentrated employment, high wages and rents, even as residence was pushed out to the periphery, just as Clarke (1951) describes. No sooner did employment displace residence to the periphery, than employment began to follow people to be closer to their peripheral residences. Over time, the concentration of employment in central cities declines. With the arrival of the post-industrial period we see a continued decentralization of employment, although it still remains relatively concentrated in the center, and gradual increase in central residence.

There does not seem to be a feature in our history that corresponds to the discontinuity that occurs at the boundary of moderate and strong increasing returns to scale, and so our ability to match history to our model ends, as does our stylized history, with a resurgence of central cities. It is natural to wonder whether some new innovation, perhaps related to information and communications technology, will bring us up to this threshold.

Summing up, while both our model and our description of history are stylized, the discussion above shows that many of the important features of the history of cities can be rationalized by our model and a trend upward in the strength of returns to scale.

8 Conclusion

Understanding how people arrange themselves when they are free to choose work and residence locations, when commuting is costly, and when increasing returns to scale affect production is one of the defining problems of urban economics. We address this problem by combining the discrete choice models employed by the recent quantitative spatial models literature, with the stylized geographies of the urban economics literature. This has, at last, permitted a complete description of equilibria, throughout the parameter space.

Equilibrium behavior is surprising and interesting for a number of reasons. First, while the presence of multiple equilibria as returns to scale increase is not surprising, the fact that in much of the parameter space multiple equilibria do not arise is less expected. Second, comparative statics as returns to scale increase contradict the conventional wisdom: increasing returns to scale in production can cause dispersion as well as agglomeration. Third, comparative statics on commuting costs also contradict conventional wisdom. As in the monocentric model, reductions

in commuting costs can lead to dispersed economic activity, but they can also lead to greater concentration.

By combining the urban economics and the QSM toolbox, we have also learned about how the two frameworks are different. Introducing heterogeneous agents with psychic commuting costs is not merely mathematical a trick for solving a difficult choice problem, it is fundamental change from the older urban economics literature. Introducing agents with heterogenous preferences over work-residence location pairs and psychic commute costs results in an average preference for central work and residence. These average preferences are a *preference* for the central location in the assumed geography, and lead to centralized employment and residence even in the absence of increasing returns to scale. Neither preference is present in the older urban economics literature.

Two of our results should contribute to the refinement of efforts at quantitative spatial modelling of cities. First, for all parameter values, equilibrium is always discontinuous at the moderate increasing returns to scale threshold and, in our simple model, these discontinuities can be catastrophically large. This seems worrying for empirical application working within plausible measurement error of this threshold and suggests the importance of at least rudimentary numerical explorations of the part of the parameter space where multiple equilibria can occur. If a quantitative model is evaluated within reasonable measurement error of a catastrophic discontinuity, then the resulting uncertainty should probably be reflected in the conclusions drawn from such a model. Secondly, one of the most common applications of quantitative spatial models is to the evaluation of observed or counterfactual changes in commuting costs. Our results demonstrate that the qualitative effect of such changes is ambiguous and depends on both the level of commuting costs and on the values of other fundamental parameters. This suggests that robustness testing of the effects of changes in commuting costs are important. Our results suggest that these comparative statics can change qualitatively as other parameters change. Our simple model does not provide a basis for thinking about how sensitive richer empirical models may be to this problem, but it does suggest that the possibility should at least be considered.

Finally, we show that the qualitative behavior of our model as returns to scale increase (along one of the possible equilibrium paths), can reproduce many of the qualitative features observed over the last 500 years of urban history in the Western world. Urbanization and increasing productivity are surely two of the most important economic phenomena in history and they appear to have been closely linked. That an economic model of geography relating the two can reproduce basic features that we observe in the history of urbanization is striking, but it should probably not be surprising.

References

Ahlfeldt, Gabriel M, Stephen J Redding, Daniel M Sturm, and Nikolaus Wolf. 2015. The economics of density: Evidence from the berlin wall. *Econometrica* 83(6): 2127–2189.

Allen, Treb and Costas Arkolakis. 2014. Trade and the topography of the spatial economy. *The Quarterly Journal of Economics* 129(3): 1085–1140.

Allen, Treb, Costas Arkolakis, and Xiangliang Li. 2015. Optimal city structure. *Yale University, mimeograph* .

Alonso, William. 1964. *Location and land use*. Harvard University Press.

Anas, Alex. 1983. Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological* 17(1): 13–23.

Anas, Alex. 1990. Taste heterogeneity and urban spatial structure: The logit model and monocentric theory reconciled. *Journal of Urban Economics* 28(3): 318–335.

Anderson, Simon P, Andre De Palma, and Jacques-Francois Thisse. 1992. *Discrete choice theory of product differentiation*. MIT press.

Armington, Paul S. 1969. A theory of demand for products distinguished by place of production. *IMF Staff Papers* 16(1): 159.

Arzaghi, Mohammad and J Vernon Henderson. 2008. Networking off madison avenue. *The Review of Economic Studies* 75(4): 1011–1038.

Bairoch, Paul. 1988. *Cities and economic development: from the dawn of history to the present*. University of Chicago Press.

Balboni, Clare Alexandra. 2019. *In Harm's Way? Infrastructure Investments and the Persistence of Coastal Cities*. Ph.D. thesis, London School of Economics.

Baum-Snow, Nathaniel. 2007. Did highways cause suburbanization? *The Quarterly Journal of Economics* 122(2): 775–805.

Beckmann, Martin J. 1976. Spatial equilibrium in the dispersed city. In *Environment, Regional Science and Interregional Modeling*. Springer, 132–141.

Cantoni, Davide and Noam Yuchtman. 2014. Medieval universities, legal institutions, and the commercial revolution. *The Quarterly Journal of Economics* 129(2): 823–887.

Cesaretti, Rudolf, José Lobo, Luis MA Bettencourt, and Michael E Smith. 2020. Increasing returns to scale in the towns of early tudor england. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53(3): 147–165.

Clark, Colin. 1951. Urban population densities. *Journal of the Royal Statistical Society* 114(4): 490–496.

Costa, Dora L. 2015. Health and the economy in the united states from 1750 to the present. *Journal of Economic Literature* 53(3): 503–70.

Couture, Victor, Cecile Gaubert, Jessie Handbury, and Erik Hurst. 2019. Income growth and the distributional effects of urban spatial sorting. Technical report, National Bureau of Economic Research.

Couture, Victor and Jessie Handbury. 2017. Urban revival in america, 2000 to 2010. Technical report, National Bureau of Economic Research.

Davis, Morris A and François Ortalo-Magné. 2011. Household expenditures, wages, rents. *Review of Economic Dynamics* 14(2): 248–261.

De la Croix, David, Matthias Doepke, and Joel Mokyr. 2018. Clans, guilds, and markets: Apprenticeship institutions and growth in the preindustrial economy. *The Quarterly Journal of Economics* 133(1): 1–70.

De Palma, André, Victor Ginsburgh, Yorgo Y Papageorgiou, and J-F Thisse. 1985. The principle of minimum differentiation holds under sufficient heterogeneity. *Econometrica: Journal of the Econometric Society* : 767–781.

De Vries, Jan. 1984. *European Urbanization, 1500-1800*. Routledge.

Dingel, Jonathan I and Felix Tintelnot. 2020. Spatial economics for granular settings. Technical report, National Bureau of Economic Research.

Dittmar, Jeremiah E. 2011. Information technology and economic change: the impact of the printing press. *The Quarterly Journal of Economics* 126(3): 1133–1172.

Eaton, Jonathan and Samuel Kortum. 2002. Technology, geography, and trade. *Econometrica* 70(5): 1741–1779.

Feldman, Mark and Christian Gilles. 1985. An expository note on individual risk without aggregate uncertainty. *Journal of Economic Theory* 35(1): 26–32.

Fujita, Masahisa. 1989. *Urban economic theory: Land use and city size*. Cambridge university press.

Fujita, Masahisa and Hideaki Ogawa. 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics* 12(2): 161–196.

Garcia-López, Miquel-Àngel, Adelheid Holl, and Elisabet Viladecans-Marsal. 2015. Suburbanization and highways in spain when the romans and the bourbons still shape its cities. *Journal of Urban Economics* 85: 52–67.

Garreau, Joel. 1992. *Edge city: Life on the new frontier*. Anchor.

Glaeser, Edward L and Matthew E Kahn. 2004. Sprawl and urban growth. In *Handbook of Regional and Urban Economics*, volume 4. Elsevier, 2481–2527.

Heblich, Stephan, Stephen J Redding, and Daniel M Sturm. 2020. The making of the modern metropolis: evidence from london. *The Quarterly Journal of Economics* 135(4): 2059–2133.

Herzog, Ian. 2020. The city-wide effects of tolling downtown drivers: Evidence from london's congestion charge. Technical report, University of Toronto.

Judd, Kenneth L. 1985. The law of large numbers with a continuum of iid random variables. *Journal of Economic theory* 35(1): 19–25.

Lucas, Robert E and Esteban Rossi-Hansberg. 2002. On the internal structure of cities. *Econometrica* 70(4): 1445–1476.

Lucas Jr, Robert E. 2001. Externalities and cities. *Review of Economic Dynamics* 4(2): 245–274.

Luce, R Duncan. 1959. *Individual choice behavior.* John Wiley.

McFadden, Daniel. 1973. Conditional logit analysis of qualitative choice behavior .

McMillen, Daniel P and John F McDonald. 1998. Suburban subcenters and employment density in metropolitan chicago. *Journal of Urban Economics* 43(2): 157–180.

Meyer, John Robert, John F Kain, and Martin Wohl. 1966. *The Urban Transportation Problem*. Harvard University Press.

Mills, Edwin S. 1967. An aggregative model of resource allocation in a metropolitan area. *The American Economic Review* 57(2): 197–210.

Monte, Ferdinando, Stephen J Redding, and Esteban Rossi-Hansberg. 2018. Commuting, migration, and local employment elasticities. *American Economic Review* 108(12): 3855–90.

Muth, Richard F. 1969. Cities and housing university of chicago press. *Chicago, IL* .

Nunn, Nathan and Nancy Qian. 2011. The potato's contribution to population and urbanization: evidence from a historical experiment. *The Quarterly Journal of Economics* 126(2): 593–650.

Ogawa, Hideaki and Masahisa Fujita. 1980. Equilibrium land use patterns in a nonmonocentric city. *Journal of regional science* 20(4): 455–475.

Redding, Stephen J. 2020. Trade and geography. Technical Report w27821, National Bureau of Economic Research Working Paper Series.

Redding, Stephen J and Esteban Rossi-Hansberg. 2017. Quantitative spatial economics. *Annual Review of Economics* 9: 21–58.

Rosenthal, Stuart S and William C Strange. 2004. Evidence on the nature and sources of agglomeration economies. In *Handbook of Regional and Urban Economics*, volume 4. Elsevier, 2119–2171.

Severen, Christopher. 2018. Commuting, labor, and housing market effects of mass transportation: Welfare and identification. Technical report, Federal Reserve Bank of Philadelphia.

Tsivanidis, Nick. 2019. The aggregate and distributional effects of urban transit infrastructure: Evidence from Bogotá's Transmilenio. *Unpublished manuscript* .

Uhlig, Harald. 1996. A law of large numbers for large economies. *Economic Theory* 8(1): 41–50.

Appendix

A. Proof of Proposition 1

We first evaluate κ for our three location city. Using (7), symmetry and (26) we have that

$$\kappa = \left[R_1^{\beta\varepsilon} W_1^{-\varepsilon} \left(2(1+\phi^2) + 2\phi\omega + 2\phi\mathcal{R} + \omega\mathcal{R}\right)\right]^{-1}. \tag{A.1}$$

Using the expression for utility maximizing choice shares (25, the definition of residential population (9), along with symmetry and (26), we have:

$$M_0 = s_{00} + 2s_{01} = \kappa R_1^{\beta\varepsilon} W_1^{-\varepsilon}(\mathcal{R}(2\phi + \omega)).$$

Substituting for κ from (A.1), we have M_0 as a function of $\mathcal{R}$ and ω. Under symmetry, the residential pattern satisfies $2M_1 + M_0 = 1$. Substituting from (30) gives M_1 as a function of $\mathcal{R}$ and ω.

Using the expression for utility maximizing choice shares (25), the definition of employment (10), along with symmetry and (26), we obtain:

$$L_0 = s_{00} + 2s_{10} = \kappa R_1^{\beta\varepsilon} W_1^{-\varepsilon}(\omega(2\phi + \mathcal{R})). \tag{A.2}$$

Substituting for κ from (A.1), we have L_0 as a function of $\mathcal{R}$ and ω. Under symmetry, the employment pattern satisfies $2L_1 + L_0 = 1$. Substituting from (A.2) gives L_1.

To evaluate expressions for residential and commercial land, evaluate (12) for our three location model to get

$$R_0 H_0 = \beta\kappa \left(W_0^{1+\varepsilon} R_0^{-\beta\varepsilon} + 2\phi W_1^{1+\varepsilon} R_0^{-\beta\varepsilon}\right),$$

$$R_1 H_1 = \beta\kappa \left[\phi W_0^{1+\varepsilon} R_1^{-\beta\varepsilon} + (1+\phi^2) W_1^{1+\varepsilon} R_1^{-\beta\varepsilon}\right].$$

Substituting from (A.1) and using (26) we have that

$$H_0 = \beta \frac{W_0}{R_0} \frac{\mathcal{R}\omega + 2\phi\mathcal{R}\omega^{-\frac{1}{\varepsilon}}}{\omega\mathcal{R} + 2\phi\mathcal{R} + 2\phi\omega + 2(1+\phi^2)},$$

$$H_1 = \beta \frac{W_1}{R_1} \frac{\phi\omega^{\frac{1+\varepsilon}{\varepsilon}} + (1+\phi^2)}{\omega\mathcal{R} + 2\phi\mathcal{R} + 2\phi\omega + 2(1+\phi^2)}.$$

Using the Cobb-Douglas property (18) and the land market clearing conditions (11), we get

$$1 - N_0 = a \frac{N_0}{L_0} \frac{\mathcal{R}\omega + 2\phi\mathcal{R}\omega^{-\frac{1}{\varepsilon}}}{\omega\mathcal{R} + 2\phi\mathcal{R} + 2\phi\omega + 2(1+\phi^2)},$$

$$1 - N_1 = a \frac{N_1}{L_1} \frac{\phi\omega^{\frac{1+\varepsilon}{\varepsilon}} + (1+\phi^2)}{\omega\mathcal{R} + 2\phi\mathcal{R} + 2\phi\omega + 2(1+\phi^2)}.$$

Solving for N_0 and N_1 and using the expressions L_0 and L_1 from (31), we arrive at the commercial land use pattern,

$$N_0 = \frac{\mathcal{R} + 2\phi}{(1+a)\mathcal{R} + 2\phi + 2a\phi\mathcal{R}\omega^{-\frac{1+\varepsilon}{\varepsilon}}},$$

$$N_1 = \frac{\phi\mathcal{R} + (1+\phi^2)}{\phi\mathcal{R} + a\phi\omega^{\frac{1+\varepsilon}{\varepsilon}} + (1+a)(1+\phi^2)}.$$

Substituting these expressions into the land market clearing conditions, $H_i + N_i = 1$, we find the housing pattern (32) and (33). Q.E.D.

B. Proof of Proposition 2

The land rent equilibrium condition (13) at $i = 0, 1$ leads to

$$R_0 = \left[\frac{\beta\kappa}{H_0}\left(W_0^{1+\varepsilon} + 2\phi W_1^{1+\varepsilon}\right)\right]^{\frac{1}{1+\beta\varepsilon}}, \tag{B.1}$$

$$R_1 = \left[\frac{\beta\kappa}{H_1}\left(\phi W_0^{1+\varepsilon} + (1+\phi^2) W_1^{1+\varepsilon}\right)\right]^{\frac{1}{1+\beta\varepsilon}}. \tag{B.2}$$

Multiplying by R_j both sides of the land market balance condition (11), we get:

$$R_0 H_0 + R_0 N_0 = R_0 H_0 + \frac{1-\alpha}{\alpha} W_0 L_0 = R_0, \tag{B.3}$$

$$R_1 H_1 + R_1 N_1 = R_1 H_1 + \frac{1-\alpha}{\alpha} W_1 L_1 = R_1. \tag{B.4}$$

Dividing (B.3) over (B.4), we obtain:

$$r = \frac{R_0}{R_1} = \frac{R_0 H_0 + \frac{1-\alpha}{\alpha} W_0 L_0}{R_1 H_1 + \frac{1-\alpha}{\alpha} W_1 L_1}. \tag{B.5}$$

It then follows from (B.1) – (B.2) that

$$R_0 H_0 = \beta\kappa\left(W_0^{1+\varepsilon} R_0^{-\beta\varepsilon} + 2\phi W_1^{1+\varepsilon} R_0^{-\beta\varepsilon}\right), \tag{B.6}$$

$$R_1 H_1 = \beta\kappa\left[\phi W_0^{1+\varepsilon} R_1^{-\beta\varepsilon} + (1+\phi^2) W_1^{1+\varepsilon} R_1^{-\beta\varepsilon}\right]. \tag{B.7}$$

Using (25), the labor market balance conditions at $i = 0, 1$ are given by

$$L_0 = s_{00} + 2s_{10} = \kappa W_0^{\varepsilon}\left(R_0^{-\beta\varepsilon} + 2\phi R_1^{-\beta\varepsilon}\right),$$

$$L_1 = s_{11} + s_{01} + s_{-11} = \kappa W_1^{\varepsilon}\left[(1+\phi^2) R_1^{-\beta\varepsilon} + \phi R_0^{-\beta\varepsilon}\right],$$

so that

$$W_0 L_0 = \kappa\left(W_0^{1+\varepsilon} R_0^{-\beta\varepsilon} + 2\phi W_0^{1+\varepsilon} R_1^{-\beta\varepsilon}\right), \tag{B.8}$$

$$W_1 L_1 = \kappa \left[\phi W_1^{1+\varepsilon} R_0^{-\beta\varepsilon} + (1+\phi^2) W_1^{1+\varepsilon} R_1^{-\beta\varepsilon} \right]. \tag{B.9}$$

Plugging (B.6) – (B.7) and (B.8) – (B.9) into (B.5), we get after simplifications:

$$r = \frac{(1+a)w^{1+\varepsilon} r^{-\beta\varepsilon} + 2a\phi r^{-\beta\varepsilon} + 2\phi w^{1+\varepsilon}}{a\phi w^{1+\varepsilon} + \phi r^{-\beta\varepsilon} + (1+\phi^2)(1+a)}. \tag{B.10}$$

Combining (15) with (16)-(17), we get:

$$w = \ell^{\gamma} \left(\frac{n}{\ell} \right)^{1-\alpha} \qquad \text{and} \qquad r = \ell^{\gamma} \left(\frac{n}{\ell} \right)^{-\alpha}. \tag{B.11}$$

Dividing (B.8) by (B.9) yields:

$$\ell = w^{\varepsilon} \frac{r^{-\beta\varepsilon} + 2\phi}{\phi r^{-\beta\varepsilon} + (1+\phi^2)}. \tag{B.12}$$

Using (B.11) and (B.12), we get:

$$w^{\alpha} r^{1-\alpha} = \ell^{\gamma} = \left(w^{\varepsilon} \frac{r^{-\beta\varepsilon} + 2\phi}{\phi r^{-\beta\varepsilon} + (1+\phi^2)} \right)^{\gamma},$$

or, equivalently,

$$w = r^{-\frac{1-\alpha}{\alpha}} \left(w^{\varepsilon} \frac{r^{-\beta\varepsilon} + 2\phi}{\phi r^{-\beta\varepsilon} + (1+\phi^2)} \right)^{\frac{\gamma}{\alpha}}. \tag{B.13}$$

The conditions (B.10) and (B.13) are the equilibrium conditions which pin down the equilibrium price ratios (r, w). They can be reformulated in terms of $(\mathcal{R}, \omega)$:

$$\mathcal{R} = \left[\frac{(1+a)\mathcal{R}\omega^{\frac{1+\varepsilon}{\varepsilon}} + 2a\phi\mathcal{R} + 2\phi\omega^{\frac{1+\varepsilon}{\varepsilon}}}{\phi\mathcal{R} + a\phi\omega^{\frac{1+\varepsilon}{\varepsilon}} + (1+\phi^2)(1+a)} \right]^{-\beta\varepsilon}, \tag{B.14}$$

$$\omega = \mathcal{R}^{\frac{1}{a}} \left(\omega \frac{\mathcal{R} + 2\phi}{\phi\mathcal{R} + (1+\phi^2)} \right)^{\frac{\gamma\varepsilon}{\alpha}}. \tag{B.15}$$

Solving (B.14) and (B.15) for $\omega^{\frac{1+\varepsilon}{\varepsilon}}$ yields (34) and (35). Q.E.D.

C. Proof of Lemma 1

As seen from (34), the denominator of $f(\mathcal{R})$ is an increasing function of $\mathcal{R}$, which is unbounded from above and negative at $\mathcal{R} = 0$. Therefore, the equation

$$D(\mathcal{R}) \equiv (1+a)\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + 2\phi\mathcal{R}^{\frac{1}{\beta\varepsilon}} - a\phi = 0$$

has a unique positive root $\mathcal{R}_0$. Clearly, $D(\mathcal{R}) > 0$ if and only if $\mathcal{R} > \mathcal{R}_0$. Since $D(1) > 0$, it must be that $\mathcal{R}_0 < 1$.

Next, it is readily verified using (34) that the numerator of $f(\mathcal{R})$ is a concave function which is positive at zero. It increases in the vicinity of 0 and then starts decreasing and goes to $-\infty$ as $\mathcal{R} \to \infty$. Hence, the equation

$$N(\mathcal{R}) \equiv \phi\mathcal{R} - 2a\phi\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + (1+\phi^2)(1+a) = 0$$

has a unique positive root $\mathcal{R}_1$. Clearly, $N(\mathcal{R}) < 0$ if and only if $\mathcal{R} > \mathcal{R}_1$. Since $N(1) > 0$, it must be that $\mathcal{R}_1 > 1$.

The numerator $N(\mathcal{R})$ and the denominator $D(\mathcal{R})$ of $f(\mathcal{R})$ have the same sign if and only if $\mathcal{R}_0 < \mathcal{R} < \mathcal{R}_1$. This interval is not empty because $\mathcal{R}_0 < 1 < \mathcal{R}_1$. Since $\mathcal{R}_1$ (resp., $\mathcal{R}_0$) is the root of the numerator (resp., the denominator) of $f(\mathcal{R})$, it follows that $f(\mathcal{R})$ decreases from ∞ to 0 over $(\mathcal{R}_0, \mathcal{R}_1)$. Q.E.D.

D. Proof of Lemma 3

It follows from the proof of Lemma 1 that $\mathcal{R}_0$ is the unique solution of

$$D(\mathcal{R}) \equiv (1+a)\mathcal{R}^{\frac{1+\beta\varepsilon}{\beta\varepsilon}} + 2\phi\mathcal{R}^{\frac{1}{\beta\varepsilon}} - a\phi = 0. \tag{D.1}$$

The expressions (37) and (D.1) imply that $\mathcal{R}_L$and $\mathcal{R}_0$ are functions of a. We next show that $\mathcal{R}_0$ and $\mathcal{R}_L$ vary with a as follows,

$$\lim_{a\to 0}\mathcal{R}_L = 1, \quad \frac{d\mathcal{R}_L}{da} < 0, \quad \lim_{a\to\infty}\mathcal{R}_L = 1-\phi,$$

$$\lim_{a\to 0}\mathcal{R}_0 = 0, \quad \frac{d\mathcal{R}_0}{da} > 0, \quad \lim_{a\to\infty}\mathcal{R}_0 = \phi^{\frac{\beta\varepsilon}{1+\beta\varepsilon}}.$$

We can show that $\mathcal{R}_0$ (resp., $\mathcal{R}_L$) increases (resp., decreases) in a by applying the implicit function theorem to $D(\mathcal{R}) = 0$ (resp., (37)). Observe further that, when $a \to \infty$ (resp., $a \to 0$), dividing $D(\mathcal{R}) = 0$ by a and taking the limit yields $\mathcal{R}_0 = \phi^{\beta\varepsilon/(1+\beta\varepsilon)}$ (resp., $\mathcal{R}_0 = 1$). Last, when $a \to \infty$ (resp., $a \to 0$), taking (D.1) at the power a and the limit yields $\mathcal{R}_L = 1-\phi$ (resp., $\mathcal{R}_L = 1$).

To determine where $\mathcal{R}_0$ and $\mathcal{R}_L$ intersect, we compare $\lim_{a\to\infty}\mathcal{R}_L$ and $\lim_{a\to\infty}\mathcal{R}_0$ by considering the equation

$$\phi^{\beta\varepsilon/(1+\beta\varepsilon)} + \phi = 1. \tag{D.2}$$

Differentiating the left-hand side of (D.2) with respect to ϕ shows that it increases from 0 to 2 when ϕ increases from 0 to 1. The intermediate value theorem then implies that, for any given β and ε, the equation (D.2) has a unique solution $\bar{\phi}(\beta,\varepsilon) \in (0,1)$, which increases with both β and ε.

The inequality $\mathcal{R}_0 \le \mathcal{R}_L$ holds if $\phi^{\beta\varepsilon/(1+\beta\varepsilon)} \le 1-\phi$, which amounts to $\phi \le \bar{\phi}$. If $\bar{\phi} < \phi \le 1$, then there exists a unique value $\bar{a} > 0$ that solves the condition $\mathcal{R}_L(a) = \mathcal{R}_0(a)$. Consequently, if $a < \bar{a}$, then $\mathcal{R}_0 \le \mathcal{R}_L$. If $a \ge \bar{a}$, then $\mathcal{R}_0 > \mathcal{R}_L$. Summing up, $\mathcal{R}_0 \le \mathcal{R}_L$ if $\phi \le \bar{\phi}$ or $a \le \bar{a}$, and $\mathcal{R}_0 > \mathcal{R}_L$ when both conditions fail. Q.E.D.

E. Lemma (Stability)

A solution $\mathcal{R}^*$ of (36) is stable if and only if f crosses g from above at $\mathcal{R}^*$. Consider a differentiable function $h : \mathbb{R}_+ \to \mathbb{R}_+$. Say that $x^* \in \mathbb{R}_+$ is a fixed point of h if x^* solves the equation $x = h(x)$. Say that x^* is stable if $h'(x^*) < 1$, and unstable otherwise.

Restate the equilibrium condition $f(\mathcal{R}) = g(\mathcal{R};\gamma)$ as follows:

$$\mathcal{R} = h(\mathcal{R};\gamma) \equiv f^{-1}(g(\mathcal{R};\gamma)), \tag{E.1}$$

where $f^{-1} : \mathbb{R}_+ \to (\mathcal{R}_0, \mathcal{R}_1]$ is the inverse of f. It follows from Lemma 1 that f^{-1} is strictly decreasing, $f^{-1}(0) = \mathcal{R}_1$, and $f^{-1}(\infty) = \mathcal{R}_0$. We know from Lemma 2 that $g(\mathcal{R};\gamma)$ increases with $\mathcal{R}$ from 0 to ∞ for $\gamma < \gamma_m$, and decreases with $\mathcal{R}$ from ∞ to 0 for $\gamma > \gamma_m$. Hence,

$$h(0) = \begin{cases} \mathcal{R}_1, & \gamma < \gamma_m, \\ \mathcal{R}_0, & \gamma > \gamma_m, \end{cases} \qquad \text{and} \qquad h(\infty) = \begin{cases} \mathcal{R}_0, & \gamma < \gamma_m, \\ \mathcal{R}_1, & \gamma > \gamma_m, \end{cases}$$

and

$$h'(\cdot) \gtrless 0 \quad \Longleftrightarrow \quad \gamma \gtrless \gamma_m.$$

As $\gamma \nearrow \gamma_m$ (resp., $\gamma \searrow \gamma_m$), $h(\cdot)$ converges to a piecewise constant function taking on one of two values, $\mathcal{R}_0$ and$\mathcal{R}_1$, and displaying a downward jump (resp., an upward jump) at $\mathcal{R} = \mathcal{R}_L$.

For any solution $\mathcal{R}^*$ of (E.1), we have:

$$h'(\mathcal{R}^*) = \frac{g'(\mathcal{R}^*;\gamma)}{f'(\mathcal{R}^*)}.$$

For $\gamma < \gamma_m$, we have $h'(\mathcal{R}^*) < 0 < 1$, hence, $\mathcal{R}^*$ is stable. For $\gamma > \gamma_m$, we have: $h'_\gamma(\mathcal{R}^*) < 1$ if and only if $|g'(\mathcal{R}^*;\gamma)| < |f'(\mathcal{R}^*)|$. In other words, $\mathcal{R}^*$ is stable if and only if f intersects g from above.[17] Q.E.D.

F. Proof of Proposition 4

Assume $\gamma = 0$. Since $0 < \mathcal{R}^* < 1$, we may assume throughout that $\mathcal{R} \in (0,1)$. Note that the equilibrium employment pattern is bell-shaped if and only if $\ell^* > 1$, while (42) is equivalent to $b > 1$.

Some tedious calculations show that the equilibrium condition $f(\mathcal{R}) = g(\mathcal{R};0)$ may be rewritten as follows:

$$\frac{\left(\frac{a}{1+a}\mathcal{R}^{-b} + \frac{1}{1+a}\mathcal{R}^{-1}\right)^{-1} + 2\phi}{\phi\left(\frac{a}{1+a}\mathcal{R}^{b} + \frac{1}{1+a}\mathcal{R}\right) + 1 + \phi^2}\left(\frac{a}{1+a}\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + \frac{1}{1+a}\mathcal{R}^{b+\frac{1}{\beta\varepsilon}}\right) = 1. \tag{F.1}$$

[17] A symmetric statement of this result is possible in which we define $h = g^{-1}(f)$, and arrive at the opposite condition for stability. Conceptually, the choice of how to conduct the inversion of $f = g$ is like a choice of the direction in which to let time run, and there is no formal basis for this choice. We resolve this ambiguity so that the unique equilibria that occur in the case when $\gamma < \gamma_m$ must be stable.

Since $1/x$ is convex, for every $\mathcal{R} < 1$ Jensen's inequality implies

$$\left(\frac{a}{1+a}\mathcal{R}^{-b} + \frac{1}{1+a}\mathcal{R}^{-1}\right)^{-1} < \frac{a}{1+a}\mathcal{R}^{b} + \frac{1}{1+a}\mathcal{R} < \mathcal{R}. \tag{F.2}$$

Plugging (F.2) in (F.1) leads to

$$1 < \frac{\frac{a}{1+a}\mathcal{R}^{b} + \frac{1}{1+a}\mathcal{R} + 2\phi}{\phi\left(\frac{a}{1+a}\mathcal{R}^{b} + \frac{1}{1+a}\mathcal{R}\right) + 1 + \phi^2}\left(\frac{a}{1+a}\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + \frac{1}{1+a}\mathcal{R}^{b+\frac{1}{\beta\varepsilon}}\right).$$

Using $b > 1$ yields

$$\frac{a}{1+a}\mathcal{R}^{b} + \frac{1}{1+a}\mathcal{R} < \frac{a}{1+a}\mathcal{R} + \frac{1}{1+a}\mathcal{R} = \mathcal{R}.$$

Since the function $\frac{x+2\phi}{\phi x+1+\phi^2}$ is increasing for all $x \geq 0$, we obtain

$$1 < \frac{\mathcal{R} + 2\phi}{\phi\mathcal{R} + 1 + \phi^2}\left(\frac{a}{1+a}\mathcal{R}^{1+\frac{1}{\beta\varepsilon}} + \frac{1}{1+a}\mathcal{R}^{b+\frac{1}{\beta\varepsilon}}\right). \tag{F.3}$$

As (42) implies

$$\frac{1}{a} < 1 + \frac{1}{\beta\varepsilon} < b + \frac{1}{\beta\varepsilon},$$

while $\mathcal{R}^* < 1$, we have

$$\frac{a}{1+a}\left(\mathcal{R}^*\right)^{1+\frac{1}{\beta\varepsilon}} + \frac{1}{1+a}\left(\mathcal{R}^*\right)^{b+\frac{1}{\beta\varepsilon}} < \left(\mathcal{R}^*\right)^{\frac{1}{a}}.$$

Replacing the bracketed term in (F.3), we obtain the inequality:

$$1 < \left(\mathcal{R}^*\right)^{\frac{1}{a}}\frac{\mathcal{R}^* + 2\phi}{\phi\mathcal{R}^* + 1 + \phi^2},$$

which is equivalent to $\mathcal{R}^* > \mathcal{R}_L$, hence $\ell^* > 1$ (see (38)). Q.E.D.

G. Proof of Proposition 5

Assume $0 < \gamma < \alpha/\varepsilon$. Since $\mathcal{R}^* > \mathcal{R}_L$, we have

$$f(\mathcal{R}_L) > f(\mathcal{R}^*) = g(\mathcal{R}^*;\gamma) > g(\mathcal{R}_L) \tag{G.1}$$

because f is decreasing by Lemma 1 and g is increasing in $\mathcal{R}$ by Lemma 2. As shown by (40), $g(\mathcal{R}_L)$ is independent of γ. Combining this with (G.1), we obtain $f(\mathcal{R}_L) - g(\mathcal{R}_L;\gamma) > 0$. Since $f(\mathcal{R}^*) - g(\mathcal{R}^*;\gamma) = 0$ while $f - g$ is decreasing by Lemmas 1 and 2, we have $\mathcal{R}_L < \mathcal{R}^*$ for all $\gamma < \alpha/\varepsilon$, which amounts to $\ell^* > 1$.

We now study the impact of γ on (i) $\mathcal{R}^*$, (ii) $\omega^*$and (iii) ℓ^*.

(i) Since $\partial g(\mathcal{R};\gamma)/\partial\gamma > 0$, applying the implicit function theorem to (36) leads to

$$\frac{d\mathcal{R}^*}{d\gamma} = \left.\frac{\partial g(\mathcal{R};\gamma)/\partial\gamma}{\partial f'(\mathcal{R})/\partial\mathcal{R} - \partial g(\mathcal{R};\gamma)/\partial\mathcal{R}}\right|_{\mathcal{R}=\mathcal{R}^*} < 0,$$

where the numerator is positive because $\mathcal{R}^* > \mathcal{R}_L$ while the denominator is negative because $f(\mathcal{R})$ is decreasing and $g(\mathcal{R};\gamma)$ is increasing in $\mathcal{R}$.

(ii) Differentiating (34) with respect to γ, we obtain:

$$\frac{1+\varepsilon}{\varepsilon}\omega^{\frac{1}{\varepsilon}}\frac{d\omega^*}{d\gamma} = \frac{df}{d\mathcal{R}}\frac{d\mathcal{R}^*}{d\gamma} > 0.$$

(iii) Observe that, combining (38) with (40), the equilibrium condition (36) can be restated as

$$\ell^{\frac{1+\varepsilon}{\varepsilon}} = \left(\frac{\mathcal{R}+2\phi}{\phi\mathcal{R}+1+\phi^2}\right)^{\frac{1+\varepsilon}{\varepsilon}} f(\mathcal{R}). \tag{G.2}$$

Since $f(\mathcal{R})$ can be decomposed as

$$f(\mathcal{R}) = \mathcal{R}^{-\frac{1}{\beta\varepsilon}} \cdot \frac{\phi\mathcal{R}+1+\phi^2}{\mathcal{R}+2\phi} \cdot \frac{1+a\frac{1+\phi^2-2\phi\mathcal{R}^{1+\frac{1}{\beta\varepsilon}}}{\phi\mathcal{R}+1+\phi^2}}{1+a\frac{\mathcal{R}-\phi\mathcal{R}^{-\frac{1}{\beta\varepsilon}}}{\mathcal{R}+2\phi}}, \tag{G.3}$$

plugging (G.3) in (G.2) leads to

$$\ell^{\frac{1+\varepsilon}{\varepsilon}} = \left(\frac{\mathcal{R}^{-\frac{1-\beta}{\beta}}+2\phi\mathcal{R}^{-\frac{1}{\beta}}}{\phi\mathcal{R}+1+\phi^2}\right)^{\frac{1}{\varepsilon}} \cdot \frac{1+a\frac{1+\phi^2-2\phi\mathcal{R}^{1+\frac{1}{\beta\varepsilon}}}{\phi\mathcal{R}+1+\phi^2}}{1+a\frac{\mathcal{R}-\phi\mathcal{R}^{-\frac{1}{\beta\varepsilon}}}{\mathcal{R}+2\phi}}.$$

The first term in the RHS clearly decreases in $\mathcal{R}$. Since the numerator (resp., denominator) of the second term is decreasing (resp., increasing), the second term also decreases in $\mathcal{R}$. Hence, the RHS is decreasing in γ. Combining this with $d\mathcal{R}^*/d\gamma < 0$, we obtain $d\ell^*/d\gamma > 0$. Q.E.D.

H. Proof of Proposition 7

(i) Consider first the case when commuting costs are high. It then follows from (37) and Lemma 3 that $\mathcal{R}_0 < \mathcal{R}_L < 1 < \mathcal{R}_1$. Therefore, for $\Delta > 0$ sufficiently small, we have:

$$\mathcal{R}_0 + \Delta < \mathcal{R}_L - \Delta < \mathcal{R}_L + \Delta < 1 < \mathcal{R}_1.$$

If γ is sufficiently close to α/ε (but still such that $\gamma > \alpha/\varepsilon$ holds), Lemma 2 implies the following inequalities:

$$\begin{gathered} g(\mathcal{R}_0+\Delta;\gamma) < f(\mathcal{R}_0+\Delta), \\ g(\mathcal{R}_L-\Delta;\gamma) > f(\mathcal{R}_L-\Delta), \\ g(\mathcal{R}_L+\Delta;\gamma) < f(\mathcal{R}_L+\Delta), \\ g(\mathcal{R}_1;\gamma) > f(\mathcal{R}_1) = 0, \end{gathered}$$

where the last inequality holds because (35) implies that, for $\gamma > \alpha/\varepsilon$, $g(\mathcal{R};\gamma) > 0$ for all $\mathcal{R} > 0$ while $f(\mathcal{R}_1) = 0$ for any γ by definition of $\mathcal{R}_1$. Therefore, by continuity of f and g, the equation (36) has at least *three* distinct solutions, which we denote as follows:

$$\mathcal{R}_1^\star > \mathcal{R}_2^\star > \mathcal{R}_3^\star.$$

Furthermore, the properties of function g imply the following:

$$\begin{aligned}
\lim_{\gamma\varepsilon\searrow\alpha} \mathcal{R}_1^\star &= \mathcal{R}_1,\\
\lim_{\gamma\varepsilon\searrow\alpha} \mathcal{R}_2^\star &= \mathcal{R}_L,\\
\lim_{\gamma\varepsilon\searrow\alpha} \mathcal{R}_3^\star &= \mathcal{R}_0.
\end{aligned}$$

The solution $\mathcal{R}_2^\star$ matches the equilibrium of Proposition 6. As for the other two solutions, $\mathcal{R}^\star$ and $\mathcal{R}_3^\star$, when γ is close enough to α/ε, we have $\mathcal{R}^\star > 1 > \mathcal{R}_3^\star$.

As $\gamma \searrow \alpha/\varepsilon$, it follows from Lemma 1 that $f(\mathcal{R}^\star)$ and $f(\mathcal{R}_3^\star)$ converge, respectively, to 0 and ∞, which implies:

$$\lim_{\gamma\varepsilon\searrow\alpha} \omega_1^\star = 0 \qquad \text{and} \qquad \lim_{\gamma\varepsilon\searrow\alpha} \omega_3^\star = \infty.$$

Hence, $\omega_1^\star < 1 < \omega_3^\star$ when $\gamma\varepsilon$ is close enough to α. It then follows from (35) that

$$\lim_{\gamma\varepsilon\searrow\alpha} \ell_1^\star = 0 \qquad \text{and} \qquad \lim_{\gamma\varepsilon\searrow\alpha} \ell_3^\star = \infty.$$

(ii) Consider now the case of high commuting costs. Then, we know from the proof of Proposition 6 that there exists a value $\bar{a} \in (0,1)$ such that

$$\mathcal{R}_L \leq \mathcal{R}_0 < 1 < \mathcal{R}_1 \tag{H.1}$$

is satisfied for $a \geq \bar{a}$, and $\mathcal{R}_0 < \mathcal{R}_L < 1 < \mathcal{R}_1$ holds otherwise. Under (H.1), there is a small $\Delta > 0$ such that the following inequalities hold:

$$\begin{aligned}
g(\mathcal{R}_1 - \Delta;\gamma) &< f(\mathcal{R}_1 - \Delta),\\
g(\mathcal{R}_1;\gamma) &> f(\mathcal{R}_1) = 0.
\end{aligned}$$

while $\mathcal{R}^\star > 1$ when γ slightly exceeds α/ε.

Furthermore,

$$\lim_{\gamma\varepsilon\searrow\alpha} (\omega_1^*)^{\frac{\varepsilon}{1+\varepsilon}} = f(\mathcal{R}_1) = 0.$$

Since $\lim_{\gamma\varepsilon\searrow\alpha} \omega_1^\star = 0$, $\omega_1^\star < 1$ when $\gamma\varepsilon$ is sufficiently close to α. Last, using (38), we have:

$$\lim_{\gamma\varepsilon\searrow\alpha} \ell_1^\star = 0.$$

Q.E.D.

I. Proof of Proposition 8

We study here the behavior of the equilibrium outcome within the domain of moderate IRS. To this end, we define the following relationship between $(\mathcal{R}_0, \mathcal{R}_1)$ and $\mathbb{R}$:

$$x \equiv \log\left(\mathcal{R}^{\frac{1}{a}} \frac{\mathcal{R}+2\phi}{\phi\mathcal{R}+1+\phi^2}\right),$$

where $x > 0$ (resp., $x < 0$) if $\mathcal{R} > \mathcal{R}_L$ (resp., $\mathcal{R} < \mathcal{R}_L$). Set $x_0 \equiv x(\mathcal{R}_0)$ and $x_1 \equiv x(\mathcal{R}_1)$ with $x_1 > 0$ and $x_0 > 0$ if and only if $\mathcal{R}_0 > \mathcal{R}_L$. Since the function $x(\mathcal{R})$ is increasing, it has an inverse, denoted $\rho(x)$, which increases with x.

It follows from (40) that the function $g(\mathcal{R};\gamma)$ is given by

$$g(\mathcal{R};\gamma) = \mathcal{R}^b \left(\mathcal{R}^{\frac{1}{a}} \frac{\mathcal{R}+2\phi}{\phi\mathcal{R}+1+\phi^2}\right)^{\frac{\gamma}{\gamma_m-\gamma}\frac{1+\varepsilon}{\varepsilon}}.$$

which amounts to

$$\mathcal{R}^{-b} g(\mathcal{R};\gamma) = \left(\mathcal{R}^{\frac{1}{a}} \frac{\mathcal{R}+2\phi}{\phi\mathcal{R}+1+\phi^2}\right)^{\frac{\gamma}{\gamma_m-\gamma}\frac{1+\varepsilon}{\varepsilon}}.$$

Therefore, the equilibrium condition (36) may be rewritten as follows:

$$\mathcal{R}^{-b} f(\mathcal{R}) = \left(\mathcal{R}^{\frac{1}{a}} \frac{\mathcal{R}+2\phi}{\phi\mathcal{R}+1+\phi^2}\right)^{\frac{\gamma}{\gamma_m-\gamma}\frac{1+\varepsilon}{\varepsilon}}.$$

Taking the logarithm of this expression and using $\mathcal{R} = \rho(x)$ yields

$$\varphi(x) \equiv \log\Big((\rho(x))^{-b} f(\rho(x))\Big) = \nu \log x,$$

where

$$\nu \equiv -\frac{\gamma}{\gamma-\gamma_m}\frac{1+\varepsilon}{\varepsilon} < 0.$$

Since $\rho(x)$ increases in x and $f(\rho)$ decreases in ρ, $f(\rho(x))$ decreases in x. Hence, $\varphi(x)$ decreases over (x_0, x_1) and has two vertical asymptotes at $x = x_0$ and $x = x_1$. As a result, $\varphi(x)$ is convex in the vicinity of x_0 and concave in the vicinity of x_1.

8.1 J. Proof of Proposition 9

Using (G.3) and (40), the equilibrium condition (36) becomes after simplifications:

$$\frac{1}{\phi\mathcal{R}+1+\phi^2}\left(\frac{1+a\frac{1+\phi^2-2\phi\mathcal{R}^{1+\frac{1}{\beta\varepsilon}}}{\phi\mathcal{R}+1+\phi^2}}{1+a\frac{\mathcal{R}-\phi\mathcal{R}^{-\frac{1}{\beta\varepsilon}}}{\mathcal{R}+2\phi}}\right)^{\lambda} = \mathcal{R}^{\mu}\frac{\mathcal{R}}{\mathcal{R}+2\phi}, \tag{J.1}$$

where λ and μ are defined by

$$\lambda \equiv \frac{\gamma\varepsilon - \alpha}{\gamma + \alpha} > 0 \qquad \text{and} \qquad \mu \equiv \frac{\gamma\varepsilon - \alpha - (1-\alpha)(1+\varepsilon)}{\beta\varepsilon(\gamma+\alpha)}.$$

The first term of the left-hand side of (H.1) decreases in $\mathcal{R}$; the second term also decreases because the numerator decreases while the denominator increases in $\mathcal{R}$. Therefore, the left-hand side of (H.1) is an increasing function of $\mathcal{R}$. Furthermore, the right-hand side of (H.1) increases from 0 to ∞ in $\mathcal{R}$ when $\mu > 0$. It is readily verified that $\gamma > 0$ if and only if

$$\gamma > \frac{1+\varepsilon}{(1-\beta)\varepsilon} - \alpha.$$

Hence, (J.1) has a unique solution $\mathcal{R}^*$. Q.E.D.

Note that

$$\frac{d}{d\gamma}\left(\frac{\gamma\varepsilon - \alpha - (1-\alpha)(1+\varepsilon)}{\beta\varepsilon(\gamma+\alpha)}\right) = \frac{\varepsilon+1}{\beta\varepsilon\,(\alpha+\gamma)^2} > 0$$

Thus, if $\mathcal{R}^* > 1$, a higher γ shift the RHS of (J.1) upward, which reduces $\mathcal{R}^*$. Hence, ω^* and ℓ^* also decrease with γ.