The optimal distribution of population across cities[a]

David Albouy[b] Kristian Behrens[c] Frédéric Robert-Nicoud[d] Nathan Seegert[e]

ABSTRACT

We develop an urban model that incorporates: (1) heterogeneous sites; (2) fiscal and urban externalities; and (3) an endogenous number of cities, i.e., the extensive margin of urban development. Within- and across-city decreasing returns to scale cause agents to perceive their city as being too large in the socially optimal allocation. As a consequence, in equilibrium the largest cities on the most amenable sites are undersized, whereas the smaller cities on less amenable sites are oversized. We propose a test for optimal city size with heterogeneous sites extending the Henry George Theorem.

Keywords: Heterogeneous sites; optimal and equilibrium city sizes; fiscal wedges; local governments; Henry George Theorem.

[a] This paper merges and supersedes Albouy and Seegert (2011) and Behrens and Robert-Nicoud (2015b). We are grateful to the editor Gilles Duranton, two anonymous referees, and Vernon Henderson for their detailed and extremely valuable comments. We are also grateful to Costas Arkolakis, Richard Arnott, Spencer Banzhaf, Klaus Desmet, Morris Davis, Pablo Fajgelbaum, David Pines, Esteban Rossi-Hansberg, Bernard Salanié, William Strange, Jacques Thisse, Tony Venables, Wouter Vermeulen, Dave Wildasin, and numerous conference and seminar audiences for discussions and feedback. Noah Smith provided preliminary research assistance. Albouy would like to thank the Lincoln Institute for Land Policy for generous assistance on this project. Behrens and Robert-Nicoud gratefully acknowledge financial support from the CRC Program of the Social Sciences and Humanities Research Council (SSHRC) of Canada for the funding of the Canada Research Chair in Regional Impacts of Globalization. This project was funded by the Russian Academic Excellence Project '5-100'. The Michigan Center for Local, State, and Urban Policy (CLOSUP) provided research funding. Any remaining errors are ours.

[b] University of Illinois at Urbana-Champaign, USA; and NBER, USA. Email: albouy@illinois.edu.

[c] Université du Québec à Montréal, Canada; National Research University Higher School of Economics, Russian Federation; and CEPR, UK. E-mail:behrens.kristian@uqam.ca.

[d] Corresponding author. Geneva School of Economics and Management, University of Geneva, Switzerland; SERC, UK; and CEPR, UK. E-mail: frederic.robert-nicoud@unige.ch. Institute of Economics and Econometrics, Uni Mail, Bd Pont d'Arve 40, CH 1211

[e] University of Utah, USA. E-mail: nathan.seegert@business.utah.edu.

Most of the world's population is now urban, and future population growth is expected to all be in cities. Indeed, cities not only define civilization, but are the engines of modern growth. Positive urban externalities – from greater sharing, better matching, and faster learning – provide agglomeration economies that attract firms and workers to cities.

Despite their economic importance, cities are often perceived as being too large. Many economists argue that free migration causes cities to become inefficiently overpopulated. Urban migrants increase competition for land and other services. They also create negative externalities such as congestion, crime, pollution, and disease. Urban economics textbooks present the view that cities are oversized as fact. Policy circles and the larger public accept this view as it reinforces ancient negative stereotypes of cities. Ultimately, this view fuels policies that limit urban growth and subsidizes rural areas.

We revisit the question of whether cities are always overpopulated. We also consider whether there are too few cities because of similar forces. These question require us to consider the incentives that determine city populations, and what these imply for land and congestion costs. Our tractable yet rich framework It builds on the seminal work of Henderson (1974b), and uses the functional forms of Behrens and Robert-Nicoud (2015a) City populations, or "sizes," result from the *fundamental trade-off in urban economics* (Fujita, 1989), which causes the average return to city size to be single-peaked hills. Productivity benefits rise with size, but are eventually dominated by congestion costs.

Moreover, cities are finite in number and built on sites that vary in exogenous productivity, or "natural advantage." The efficient city scale, the population that maximizes the average return of an isolated city, increases with this productivity. Additionally, sites may remain un-populated, making the extensive margin to new city development nontrivial .

To allocate populations across cities, we compare three different scenarios: (1) the social optimum, where a central planner allocates individuals to places; (2) local governments, in which city governments control their size directly; and (3) free-mobility equilibria, in which individuals choose in which city or urban area to live. Solving for the central optimum, we provide a simple theory-based test of population optimality (or suboptimality) building on the Henry George Theorem.

Importantly, local governments and central planners have different objectives. Local governments maximize the welfare of residents in a single city, setting population to the locally efficient scale. This equates each city's marginal return with its average return. They ignore non-residents and what happens if other sites are developed.

A benevolent central planner maximizes the welfare of those in the entire urban system. Thus, she equalizes the marginal returns to city size across all sites. Furthermore, she considers how limiting the size of cities will put require putting people on supermarginal sites, with lower productivity (or a rural area with diminishing returns). This is the extensive margin of development. Thus the planner must balance diminishing returns at the city level from congestion with those at the system level from lower site quality.

In the third case, With free migration, individuals move to the place offering the highest average return. In equilibrium, the average returns will be equal. This is the case where oversized cities are most likely. That is, in the standard case where land is given for free to newcomers, and taken without compensation from leavers — much as in a standard highway congestion problem.

When sites are homogeneous, the centralized optimum, local-government, and (most efficient) free-migration population allocations all coincide. All cities have the same marginal and average returns at their identical efficient scales (Henderson and Becker, 2000). The entire urban population is divided equally across sites at an identical optimum city size.

Everything changes When cities are heterogeneous. The notion of a single optimal city size no longer applies, and local governments will not enforce the centralized optimum. When marginal returns are equal, average returns will be to the right of their peak. Thus, local governments will want sizes that are too small. Individuals shed from the optimum must then live on lower quality sites, creating too many cities. The model thus provides a natural explanation for the ubiquitous NIMBY-ism of local governments. It also explains why many New Yorkers may feel that their city is too large, even though it might be too small from the point of view of the American urban system as a whole.

Heterogeneity also creates issues for free migration. If sites are homogeneous, then all cities can be optimally sized in a stable equilibrium. They could also all be equally too large due to a coordination failure (Harris and Todaro, 1970; Tolley, 1974; Arnott, 1979; Upton, 1981; Abdel-Rahman, 1988; Fenge and Meier, 2002; O'Sullivan, 2007).[1] However, heterogeneity causes average returns to differ from each other when marginal returns are equal. This means free migration will undo an optimal allocation. Absent any externalities across cities, the best sites will be over-populated.

[1] The argument goes as follows. The cost migrants pay to enter a city is equal to the average cost, rather than the marginal cost, associated with urban life. With homogeneous sites, and analogous to the efficient scale of a firm, the efficient population scale of all cities is the size at which the marginal and the average benefits coincide. Migrants ignore externalities (agglomeration economies and congestion costs) they impose on others and thus respond to the average benefit. They will enter a city until the average benefit of migration equals the outside option (the average benefit in another city or the countryside). The resulting equilibrium population level is only stable when benefits are falling with city size, thus implying that cities are too large.

Yet, as we point out, a natural fiscal (or pecuniary) externality arises from the need for migrants to rent (or purchase) land in their destination city. This creates a difference between the benefit produced and the benefit consumed within a city. Higher land rents in the most productive cities devalue migrants' gains from agglomeration economies, driving a wedge between private and social returns. This wedge distorts the location choices of mobile individuals at both the intensive and extensive margins of urban development. When land is sufficiently important in the economy, free migration can cause there to be too many cities, and for the largest cities to be too small.

The main issue we address in this article—the (sub)optimality of population allocations across possible urban sites—is policy relevant for many reasons. First, local governments often block or try to attract new residentss. Typically, they limit populations by restricting development, often in the form of restrictive land-use regulations or urban containment seen in cities such as Portland, Oregon, and in the UK's Green Belt (Cheshire and Sheppard, 2002; Glaeser et al., 2005; Hilber and Robert-Nicoud, 2013).[2] We find that achieving the locally efficient scale of each city individually is suboptimal nationally. This result is not specific to our theoretical design.

Second, our finding that freely-mobile agents do not have incentives compatible with the optimum is also quite general. Real-world features, such as proper land markets — common in the OECD and beyond — can result in the largest, most advantaged, cities being too small even without local limitations. By contrast, in the absence of a *fiscal externality*, competitive land developers who own the land typically achieve the first-best allocation by pricing the agglomeration externalities and congestion costs correctly.[3] This result contrasts sharply with the view famously echoed by Molotch's (1976) urban growth machine theory, which asserts that the economic interests of landowners are in overdeveloping local land.

Our paper contributes to several academic debates. Importantly, we contribute to the quickly expanding work on mis-allocated urban populations. Formal reasoning on optimal systems of regions was pioneered by Buchanan and Goetz (1972) and Flatters et al. (1974), developed extensively by Henderson (1974b) and Henderson and Becker (2000), and given comprehensive treatments by Kanemoto (1980), Fujita (1989), and Abdel-Rahman and Anas (2004). Our work expand theirs by

[2] See, for example, the current debate in California, where "lawmakers are considering extraordinary legislation to, in effect, crack down on communities that have, in their view, systematically delayed or derailed housing construction proposals, often at the behest of local neighborhood groups." (Source: New York Times, July 17, 2017: "The Cost of a Hot Economy in California: A Severe Housing Crisis" available online at `https://www.nytimes.com/2017/07/17/us/california-housing-crisis.html`).

[3] What we model as a fiscal externality potentially incorporates any number of externalities, including pecuniary externalities from federal taxes and redistributions, to technological externalities, such as environmental ones (e.g., Helpman and Pines, 1980; Albouy, 2012; Borck and Tabuchi, forthcoming). What matters is that these externalities occur across cities and become more positive with city size.

adding an extensive margin of heterogeneous sites, with variable returns to city scale. Thus, we can characterize an optimal number of cities, whose optimal size of on the entire system it is in. Our results complement the growing literature on various distortions to the population distribution (Henderson, 1974b, 1986; Au and Henderson, 2006; Albouy, 2009; Desmet and Rossi-Hansberg, 2013; Redding, 2016; Eeckhout and Guner, 2017; Borck and Tabuchi, forthcoming; Hsieh and Moretti, forthcoming). The work of Henderson and Becker (2000) in particularly close to ours. In an extension of their framework, they feature one city with a productivity advantage over all others, and show that this city is at once the largest and undersized in equilibrium. We generalize their model with a wide range of advantages among many potential sites, as well as adding fiscal externalities.

Our work also clarifes the often implicit assumptions of land ownership in urban models. It demonstrates that land-rent payments may cause a fiscal externality that distorts the distribution of population. In doing so, we extend the Henry George Theorem (Stiglitz, 1977; Vickrey, 1977; Arnott, 1979) to the case where cities cannot be replicated perfectly. We develop a single statistic for aggregate land values that characterizes a necessary condition for optimality of an urban system.

The rest of the paper is structured as follows. Section 1 introduces the model. Section 2 characterizes the central and local optimal population distributions and highlights the presence of local externalities. This produces our first result: with heterogeneous cities, populations preferred by local governments differ from those of central governments. Section 3 introduces fiscal externalities with a generalized model of land rents. It produces our second result: with heterogeneous sites, the most productive cities are undersized if fiscal externalities are greater than local externalities. Section 4 compares the size and number of cities under the three solution concepts for various parameter values, and presents the extended Henry George Theorem. Section 5 concludes.

1. An urban system with heterogeneous sites

The economy comprises a mass $\overline{N}$ of homogeneous agents to be allocated in a spatial economy consisting of a continuum of various potential city sites and a rural area. The mass of agents is endogenously split into an urban population N and a rural population R, such that $R+N=\overline{N}$. There is a single, homogeneous consumption good in the economy and utility is linear in the consumption of that good. The consumption good is produced in cities as well as the rural area. It is traded at no cost and thus a natural candidate for the numéraire.[4]

[4] Assuming that urban and rural locations produce the same good and that trade involves no cost simplifies the analysis considerably. Relaxing these assumptions would not change the essence of our qualitative results, but it would

We focus on heterogeneity in natural advantages of cities and lump all sources of heterogeneity (including geography) into a single parameter for each city. Specifically, the only exogenous difference among cities is in their natural production amenities or local productivity, which we denote $a \in \mathbb{R}_+$.[5] The production amenities are distributed with cumulative density function $G(\cdot)$ over $\mathbb{R}_+$, which is twice continuously differentiable and unbounded from above.[6] We refer to a city with productivity a as "city a" for short. We model rural site heterogeneity implicitly by assuming that total agricultural output, $A = F(R)$, is an increasing and strictly concave function of the rural population, R. We further assume that the Inada condition $\lim_{R\to 0} F'(R) = +\infty$ is satisfied, which ensures that there is always an interior urban-rural split of population.

Cities result from a conflict of agglomeration economies and urban costs. The latter eventually dominate the former, generate the hill-shaped graph of utility with respect to population. The single interior peak is the local efficient scale. We use additively separable benefits and costs to obtain clear-cut results, although the insights we develop are much more general.

Urban dwellers benefit from agglomeration economies that translate into higher wages. The gross output (and wage) of a worker in city a with population $n(a)$ is given by

$$w(a, n(a)) \equiv an(a)^{\varepsilon}, \tag{1}$$

where $a > 0$ is city productivity and $n(a)^{\varepsilon}$ are agglomeration economies external to the representative firm.[7]

The elasticity of agglomeration economies with respect to city size is given by $\varepsilon > 0$. We are agnostic about the precise microeconomic foundations of these agglomeration economies and simply assert their existence. Duranton and Puga (2004) survey a wide class of models that deliver the reduced form in equation (1). The ample empirical evidence documenting the existence of agglomeration economies is surveyed by Combes and Gobillon (2015).

The urban costs urban dwellers bear consiste of commuting costs and land rents.[8] Let ρ define the share of per capita land values, γn^{γ} that accrues to agents who reside outside the city. The

do so at the cost of blurring some of the central results of the paper.

[5] For simplicity, we focus on a single dimension of site heterogeneity assuming that land offers heterogeneous production amenities. For a model with both production and quality of life amenities, see Seegert (2011).

[6] Including an upper bound for a does not change the qualitative results yet makes all proofs more cumbersome. See Albouy, Behrens, Robert-Nicoud and Seegert (2016) for an analysis without an upper bound for a.

[7] The representative firm uses labor only, is perfectly competitive, produces $an(a)^{\varepsilon}$ units of output per worker under constant returns to scale, and its output is freely traded. Perfect competition in the labor market of that city implies that equilibrium wages are equal to per capita gross output as in (1).

[8] In our model, housing production only land, so that housing costs equal land rents.

literature often considers the polar cases where either $\rho = 0$ or $\rho = 1$. With $\rho = 0$, land rents are redestributed evenly *within* the city. With $\rho = 1$, land rents are redistributed evenly across the urban system, or given to absentee landlords.[9] We let $\rho \in [0,1]$ For simplicity, we assume aggregate land rents (ALR) in a city are proportional to aggregate commuting costs (ACC) in the city. We use $\gamma \equiv \text{ALR}/\text{ACC}$ as that factor of proportionality. The Alonso (1964)–Muth (1969)–Mills (1967) monocentric model is a classic way to deliver such urban costs (Fujita, 1989; Duranton and Puga, 2015). Assume the cost of commuting from distance x from the CBD takes the form νx^{γ}. Then, the land rent that leaves agents indifferent between locations within a linear city on the interval $[-n/2, n/2]$ is given by $L(x) = \nu\left[(n/2)^{\gamma} - |x|^{\gamma}\right]$. This implies the following Aggregate Land Rent, Aggregate Commuting Costs, and Aggregate Urban Costs:

$$\text{ALR} = 2\nu\left(\frac{n}{2}\right)^{1+\gamma}\frac{\gamma}{1+\gamma}, \quad \text{ACC} = 2\nu\left(\frac{n}{2}\right)^{1+\gamma}\frac{1}{1+\gamma}, \quad \text{and} \quad \text{AUC} = 2\nu\left(\frac{n}{2}\right)^{1+\gamma}\frac{\rho\gamma}{1+\gamma},$$

with $\rho\text{ALR} + \text{ACC} = \text{AUC}$. It then immediately follows that $\gamma \equiv \text{ALR}/\text{ACC}$. We set $\nu \equiv (1+\gamma)2^{\gamma}$ by choice of units so that Aggregate Urban Costs simplify to $\text{AUC} = (1+\rho\gamma)n^{1+\gamma}$ (the expressions for ALR and ACC simplify in the same way).

The consumption of the numéraire good of an urban dweller is equal to her wage minus her urban costs, net of redistributed land rents.[10] As a result, urban dwellers derive an average utility from living in city a of size $n(a)$ equal to:

$$v(a, n(a)) = an(a)^{\varepsilon} - (1+\rho\gamma)n(a)^{\gamma}. \tag{2}$$

We impose $\gamma > \varepsilon$ to ensure that urban costs dominate agglomeration economies beyond some city size. This assumption ensures that multiple cities exist and it is consistent with empirical evidence (see, e.g., Combes, Duranton and Gobillon, 2018).[11]

Rural dwellers work in a competitive agricultural sector and earn a wage income $w_R(R) =$

[9] See Pines and Sadka (1986) for an early analysis of the consequences of relaxing the absentee landlord hypothesis in a closed city setting. If land rents are redistributed nationally there are no general equilibrium effects as all individuals treat their land rental income as exogenous to their own location decision by atomicity, and because there are no wealth effects on location.

[10] We require commuting costs not to be fully paid in time—i.e., they are not strictly iceberg costs—to ensure that city size is not independent of the local amenity as discussed by Duranton and Puga (2004). Allowing for a different functional form such that γ is increasing in n would reinforce of our qualitative results.

[11] Our functional form for utility differs from that used in some previous works. It is, however, not more specific. An alternative utility function that is sometimes used (see, e.g., Eeckhout 2004) is one where utility is monotonically decreasing with population, e.g., $u(a, n(a)) = an(a)^{\varepsilon-\gamma}$. Such a utility function is decreasing in population and the individually efficient city size is zero. In this context, it always pays off to create a new city and make existing cities smaller. With such preferences the extensive margin of the urban system is almost impossible to analyze.

$F'(R) > 0$, which is decreasing with the rural population by $F''(\cdot) < 0$. The rural area does not suffer from congestion, and therefore the benefit an agent receives in the rural area is given by

$$u_R(R) = F'(R). \tag{3}$$

A fixed factor, agricultural land, collects the Ricardian rents that are generated in the rural area. The latter are assumed to be redistributed evenly across the entire population.

2. Local and Central Allocations: The Role of Heterogeneity

We now characterize the systems of cities when local governments and a central government allocate population. The population allocations are denoted by n_l and n_* for the local and central allocations, respectively. To derive sharp results that highlight the important role of site heterogeneity, we focus on the case where land rents are fully redistributed locally, i.e., $\rho = 0$. This case is the most frequent assumption in the optimal city-size literature. Comparing the outcomes with local governments and a central government produces our first result: with heterogeneous cities, local governments prgoduce too many cities and underpopulate the most productive cities relative to the social optimum.

2.1. Local Optimum

Let $u(a,n(a)) \equiv an(a)^{\varepsilon} - n(a)^{\gamma}$ denote the utility in (2) when $\rho = 0$. The objective of each local government is to maximize the utility each individual receives in its city (and since agents are homogeneous, this amounts to maximizing the average utility in its city):

$$\max_{n(a)} \; u(a,n(a)) = an(a)^{\varepsilon} - n(a)^{\gamma}. \tag{4}$$

To achieve this aim, the local government determines the population in its city, $n_l(a)$, by setting different policies.[12] Local governments have the power to exclude people from their cities either directly—by using urban growth bounds and other controls—or indirectly—by using various land use regulations that impose a regulatory tax on newcomers (Glaeser, Gyourko and Saks, 2005). For simplicity, we assume that local governments set $n_l(a)$ directly. The first-order condition that

[12] Here, our analysis differs to the analysis in Henderson and Becker (2000) in two ways. First, we do not assume that local government can set subsidies to attract urban dwellers. Second, we assume that all city governments are active but we assume that individual agents are passive.

characterizes the choice of the local government can be written as

$$\frac{\partial u(a,n(a))}{\partial n(a)} = \varepsilon a n(a)^{\varepsilon-1} - \gamma n(a)^{\gamma-1} = 0. \tag{5}$$

The second-order condition holds by $\varepsilon < \gamma$. The first-order condition implies that the population that maximizes individual utility within a city is

$$n_l(a) = \left(\frac{a\varepsilon}{\gamma}\right)^{1/(\gamma-\varepsilon)}, \tag{6}$$

which is an increasing function of local productivity a. Substituting the locally efficient population scale (6) into utility yields

$$u(a,n_l(a)) = \left(1-\frac{\varepsilon}{\gamma}\right) a n_l(a)^{\varepsilon} = \left(\frac{\gamma}{\varepsilon}-1\right) n_l(a)^{\gamma}. \tag{7}$$

We define the *intensive margin condition* by rewriting equation (7) as a function of utility, wage, and a *local externality*

$$\theta_l \equiv 1-\frac{\varepsilon}{\gamma} \in (0,1)$$

as follows:

$$u(a,n_l(a)) = \theta_l a n_l(a)^{\varepsilon} = \theta_l w(a,n_l(a)). \tag{8}$$

Then, for any pair of inhabited cities a, $b \in \mathbb{R}_+$, equations (7) and (8) yield:

$$u(a,n_l(a)) - \theta_l w(a,n_l(a)) = u(b,n_l(b)) - \theta_l w(b,n_l(b)). \tag{9}$$

The local externality θ_l is the difference between the local optimum condition and the condition that would set utilities equal across all cities (which arises in the limiting case when $\theta_l = 0$). In the local optimum, the local externality causes cities with more amenities to provide higher utility; if $a > b$, then $u(a,n_l(a)) > u(b,n_l(b))$ by $w(a,n_l(a)) > w(b,n_l(b))$. Individuals excluded from high amenity cities are forced to inhabit lower quality cities or live in the rural area.

The *extensive margin condition* for the local optimum ensures the marginal city, a_l, provides residents with the same utility as the rural sector:

$$u(a,n_l(a)) \geq F'(R_l), \quad \forall n_l(a) > 0, \quad \text{and} \quad u(a_l,n_l(a_l)) = F'(R_l), \tag{10}$$

where the equality comes from the Inada condition in the rural sector. All sites with a productivity lower than a_l host no population: $n(a_l) = 0$ for all $a < a_l$.

2.2. Central Optimum

The objective of the central government is to maximize aggregate utility in the spatial economy, which is equivalent to maximizing aggregate consumption of the numéraire good.[13] To achieve this aim, the central government determines the population in each city (the intensive margin), the number of cities inhabited (the extensive margin, given by the amenity level a_* in the least productive city inhabited), and the degree of urbanization, given by the urban population $\overline{N} - R$. We may write this problem as follows:

$$\max_{n(a),R,a_*} F(R) + \int_{a_*}^{\infty} n(a)u(a,n(a))dG(a) \quad \text{subject to} \quad R + \int_{a_*}^{\infty} n(a)dG(a) = \overline{N}. \tag{11}$$

The second equation above is the population adding-up constraint. Note the key difference with (4): the central government optimizes over both the intensive and extensive margins of urban development, i.e., $n(a)$ and a_*, respectively; whereas local governments only consider intensive margins.

Conditions (12)–(17) below fully characterize the socially optimal allocation, which we label using subscripts '$*$'. The population in the rural area is determined by the first-order condition with respect to R, which ensures that adding an extra worker to the economy increases total surplus by at least the marginal benefit in the rural area:

$$\mu_* \geq F'(R_*), \quad R_* \geq 0, \tag{12}$$

with complementary slackness, where μ_* denotes the Lagrange multiplier associated with the population adding-up constraint in (11) evaluated at the optimal allocation. Recall that $F(\cdot)$ satisfies the Inada conditions, hence the rual area is inhabited at the optimal allocation and $F'(R_*) = \mu_*$ holds for a unique $R_* > 0$.

The first-order conditions for the optimal city sizes, $n_*(a)$, state that the marginal benefit of residing in any city must be equal across all inhabited sites as in Flatters, Henderson and Mieszkowski

[13] All agents have a constant and identical marginal utility of income. Hence utility is transferable, and uniform transfers do not affect location (Mirrlees, 1982). This also implies that land rents do not enter the central government's problem, irrespective of the value of ρ, because they are transfers between agents.

(1974) or Arnott (1979):

$$a(1+\varepsilon)n_*(a)^{\varepsilon} - (1+\gamma)n_*(a)^{\gamma} \leq \mu_*, \quad n_*(a) \geq 0, \tag{13}$$

with complementary slackness. For any pair of inhabited sites, a and b, this intensive margin condition can be written as a function of utility and an *urban externality*,

$$\theta_u \equiv 1 - \frac{1+\varepsilon}{1+\gamma} \in (0,1) \tag{14}$$

as follows:

$$u(a,n_*(a)) - \theta_u w(a,n_*(a)) = u(b,n_*(b)) - \theta_u w(b,n_*(b)). \tag{15}$$

The urban externality is the difference between the central optimum condition and the condition that would set utilities equal across all cities (which arises in the limiting case $\theta_u = 0$). In the central optimum, the urban externality causes cities with more amenities to produce higher utilities; if $a > b$, then $u(a,n_*(a)) > u(b,n_*(b))$ by $w(a,n_*(a)) > w(b,n_*(b))$.[14] As we show below, the urban externality captures the standard urban forces that usually lead to oversized cities when agents are freely mobile.

The first-order condition for the optimal extensive margin of urban development, a_*, states that the least productive site to develop a city satisfies $\mu_* = u(a_*,n_*(a_*)) = a_* n_*(a_*)^{\varepsilon} - n_*(a_*)^{\gamma}$. Combining this expression with (13) evaluated at $a = a_*$ then yields $\varepsilon a_* n_*(a_*)^{\varepsilon} - \gamma n_*(a_*)^{\gamma} = 0$. This result establishes that the least productive site to be populated will be developed at its locally efficient scale:

$$n_*(a_*) = \left(\frac{a\varepsilon}{\gamma}\right)^{\frac{1}{\gamma-\varepsilon}} = n_l(a_*). \tag{16}$$

We thus have

$$\mu_* = u(a_*,n_l(a_*)) = \left(\frac{\gamma}{\varepsilon} - 1\right) n_l(a_*)^{\gamma}, \tag{17}$$

where the second equality comes from (7). This expression defines the common marginal benefit, μ_*, across all inhabited cities.

Note that there exists a unique allocation satisfying equations (12), (13), (16), and (17). To see this result, remember first that there are decreasing returns to labor in the rural sector by the assumption $F''(\cdot) < 0$. Furthermore, $F(\cdot)$ satisfies the Inada conditions so that $R_* > 0$ holds. Second, the urban system features decreasing returns to urban population $N_* \equiv \int_{a_*}^{\infty} n(a)dG(a)$. Indeed, for

[14] The central government can reallocate utility using city-specific lump-sum transfers such that all agents receive the same utility ex post.

all $a > a_*$, we may rewrite (13) as

$$an_*(a)^{\varepsilon} - n_*(a)^{\gamma} = \mu_* + [\gamma n_*(a)^{\gamma} - \varepsilon n_*(a)^{\varepsilon}] > \mu_*,$$

where the inequality holds by the second-order condition for $n_*(a)$, all $a > a_*$, which implies that all inhabited cities feature decreasing returns to city size at their optimal size (by $\varepsilon < \gamma$). Plugging this inequality into the expression for total urban output net of urban costs, $\int_{a_*}^{\infty} \left[an(a)^{1+\varepsilon} - n(a)^{1+\gamma}\right] \mathrm{d}G(a)$, yields

$$\frac{\int_{a_*}^{\infty} \left[an(a)^{1+\varepsilon} - n(a)^{1+\gamma}\right] \mathrm{d}G(a)}{N_*} > \mu_*,$$

namely, the average social benefit is greater than the social marginal benefit of urbanization, i.e., there are decreasing returns to urbanization in the efficient allocation. Combined with $\mu_* = F'(R_*)$ and $R_* > 0$, this result implies that there is a unique pair R_* and N_* that both belong to the interval $(0, \bar{N})$ and also satisfy the population adding-up constraint. Then, given N_*, the solutions to (13), (16), and (17) are unique.

2.3. Local-Urban Externalities: Are Cities Too Large? Part 1

Comparing the central and local optimum allocations produces our first result on the importance of modeling heterogeneous cities.[15]

Proposition 1 *If cities are endowed with heterogeneous productivity, then* ***(i)*** *more cities are inhabited in urban systems with local governments than at the central optimum* ($a_l < a_*$)*, and* ***(ii)*** *all cities inhabited at the central optimum have larger populations at that allocation than at the allocation with local governments:* $n_*(a) > n_l(a), \forall a > a_*$.

Proof. Let $smb(n(a)) \equiv (1+\varepsilon)an(a)^{\varepsilon} - (1+\gamma)n(a)^{\gamma}$ denote the social marginal benefit of populating site a, as given by the left-hand side of (13). smb is $\cap$-shaped in $n(a)$: it is increasing in $n(a)$ until $n(a)^{\gamma-\varepsilon} = a\frac{\varepsilon}{\gamma}\frac{1+\varepsilon}{1+\gamma}$ and decreasing in $n(a)$ beyond that threshold.

By the second-order condition of the optimization problem (11), $smb(n_l(a)) > \mu_*$ if and only if $n_l(a) < n_*(a)$. Such cities are undersized. Conversely, $smb(n_l(a)) < \mu_*$ if and only if $n_l(a) > n_*(a)$. Such cities are oversized. If $n_l(a) > 0$ for some $a < a_*$, then there are too many cities in equilibrium:

[15] Henderson and Becker (2000) obtain a similar result in a different environment: in theirs, only one local government is active at a time but it can attract freely Mobile urban dwellers by designing subsidies; in their Proposition 6, the city size of the active local government is too small.

sites with amenities below a_* would not be developed by the central government. Hence, cities at such sites are oversized in a trivial sense—by virtue of existing—because they have zero population at the optimal allocation.

Using equation (17), which states that the marginal city is at its locally optimal size, we have

$$\frac{smb(n_l(a))}{\mu_*} = \frac{(1+\varepsilon)an_l(a)^{\varepsilon} - (1+\gamma)n_l(a)^{\gamma}}{n_l(a_*)^{\gamma}\left(\frac{\gamma}{\varepsilon}-1\right)} = \frac{n_l(a)^{\gamma}\left(\frac{\gamma}{\varepsilon}-1\right)}{n_l(a_*)^{\gamma}\left(\frac{\gamma}{\varepsilon}-1\right)} = \left(\frac{a}{a_*}\right)^{\gamma/(\gamma-\varepsilon)} > 1 \qquad (18)$$

for all $a > a_*$. This result establishes that all cities that are inhabited in the central optimum are larger at that allocation than they are in the local government allocation. This establishes part (**ii**) of Proposition 1. It then follows by the population adding-up constraint that $a_l < a_*$, as more cities have to be developed to accommodate the urban population that is in excess supply when local governments control city sizes, which establishes part (**i**) of the result. Therefore, all cities with $a < a_*$ are oversized by virtue of existing, but all other cities are too small.

Finally, note that, when cities are homogeneous, $a = a_* = a_l$, all cities are at their optimal size, irrespective of whether local governments or the central government choose their sizes. □

Proposition 1 highlights the importance of modeling heterogeneity of urban productivity and the extensive margin of urban development. Agents in a city with a population greater than its efficient scale consider their city oversized. But, from the perspective of the central government, it is efficient to crowd that city in order to limit the number of cities that have to be inhabited. Additional, supra-marginal, cities have to be built on sites with poorer productivity. These different incentives lead local governments to enact over-restrictive land use regulations and urban containment. To limit this form of NIMBY-ism, policies that restrict development should not be designed only by local authorities that do not internalize consequences on the urban system.[16]

Let us stress that the qualitative result summarized in Proposition 1 is quite general and holds in generic environments that nest our own. Consider any model with variable returns to city size, first increasing and then decreasing in city size. Let $u(n,a)$ denote the per capita utility level, with $u_a > 0$, $u_{an} > 0$, and $u_{nn} < 0$ for any a and n. Obviously, (2) satisfies these properties, as does the canonical Henderson (1974b) model. The centrally optimal allocation satisfies $n_* + n_* u_n(n_*, a) = \mu_*$ and $u_n(n_*) < 0$ for all a such that $n_*(a) > 0$. The local optimum allocation satisfies $u_n(n_l, a) = 0$ for all a such that $n_l(a) > 0$. It then naturally follows from $u_{nn} < 0$ that $n_l(a) < n_*(a)$ for all a's that are inhabited in both allocations. These cities are the largest ones. Thus, we can conclude that the

[16] The implication that local and central governments have different incentives is consistent with Vermeulen (2017).

largest cities are too small in the local optimum allocation also in this generic environment.

2.4. Implementing the optimal allocation through developers

As shown by Henderson (1974b) and Henderson and Becker (2000), the optimal allocation may be implemented as an equilibrium outcome with perfectly competitive land developers when sites are homogeneous. Remarkably, this result extends to our setting with heterogeneous sites.[17] Assume that there is a continuum of atomistic developers who each own one site, so that they have no individual impact on the aggregate variables of the economy. Local land developers attract urban dwellers by offering (possibly negative) subsidies, s, such that the net utility is weakly larger than the economy-wide utility level, denoted by u_d (where subscript d stands for developers). They collect aggregate land rents, $\gamma n(a)^{\gamma+1}$, in the site they develop. Agents in the city receive utility $an(a)^{\varepsilon} - n(a)^{\gamma} - \gamma n(a)^{\gamma} + s$, i.e., the wage net of commuting costs and per capita land rents plus the subsidy. A developer who owns site a thus solves the following maximization problem:

$$\max_{n(a),s} \ \pi(n(a), s; u_d) = \gamma n(a)^{\gamma+1} - n(a)s \quad \text{subject to} \quad u_d \leq an(a)^{\varepsilon} - (1+\gamma)n(a)^{\gamma} + s,$$

where the second term in π is the cost of the subsidy to attract $n(a)$ people to the site. Local land developers choose s to set agents at their reservation utility, $s = u_d - an(a)^{\varepsilon} + (1+\gamma)n(a)^{\gamma}$, treating u_d as a parameter. Plugging this expression for s into the profit and optimizing with respect to $n(a)$, the first-order condition is given by

$$a(1+\varepsilon)n(a)^{\varepsilon} - (1+\gamma)n(a)^{\gamma} - u_d = 0 \tag{19}$$

if $n(a)$ is positive (the term on the left-hand side is negative otherwise). Observe that (19) is isomorphic to (13), with u_d instead of μ_*. Then, the solution $n_d(a)$ satisfies $n_d(a) = n_*(a)$ if $u_d = \mu_*$. We claim that this result holds at the equilibrium with competitive land developers. To establish this result, note first that $u_d > \mu_*$ cannot possibly arise: the allocation with developers cannot Pareto dominate the optimal allocation by definition. Second, we can rule out $u_d < \mu_*$ using a proof by contradiction. Assume hence that $u_d < \mu_*$ holds. Then, cities are more populated and fewer than at the optimal allocation. This implies that land developers who own empty land with an a slightly below a_* can attract workers by offering them a utility higher than u_d and yet make pure profits. To see this result, use (19) to solve for u_d and substitute into the definition for land developer profit to

[17] We provide the general version with heterogeneous sites and fiscal externalities in Appendix B.

see that the profit is positive for $n_d(a) > n_*(a)$, all $a \geq a_*$. Thus, even the marginal land developer (the one owning land in site a_*) makes positive profits, which violates the competitive assumption. Thus $u_d = \mu_*$, and the allocation with developers coincides with the central optimum allocation.

The equilibrium profits of land developers are strictly positive for all $a > a_*$, are increasing in a, and equal to zero for $a = a_*$ only. Developers who own superior sites make strictly positive profits since better sites are in limited supply and command Ricardian rents. When land is homogeneous, the equilibrium profits of land developers are zero. Here, equilibrium rents remain positive because sites are (vertically) differentiated goods.[18]

In sum, the allocation with competitive land developers is more efficient than the one with local governments. Competition for mobile agents prod competitive land developers to take the external effects of their choices into account. By maximizing the value of land, they have an incentive to expand the city and to maximize the *total* utility of its residents, whereas local government only seek to maximize their average utility.

3. Equilibrium Allocation With Free Migration

This section characterizes the system of cities in an equilibrium where agents maximize utility by choosing where to live (what Henderson and Becker (2000) refer to as 'self-organization'). We make two departures with respect to the previous section. First, we assume that local and central governments are passive.

Second, we emphasize the role of land ownership. Assumptions on land ownership are an important determinant of how mobile agents make their location decisions. In the previous section, we have adopted the most frequent assumption in the optimal city-size literature by setting $\rho = 0$, namely, that land rents are evenly shared among city residents. This assumption implies that, upon moving to a city, the agent (freely) gets a claim to local land rent. As migrants are rarely given land for free in the location they move to or expropriated from the land they own in the city they leave, setting a value of ρ close to unity seems more realistic for modeling migration (the case of $\rho = 1$

[18] It is useful to make the analogy between our problem and that of imperfect competition between firms. In our model, developers differ by productivity (i.e., the quality of the site they own), but varieties (i.e., sites) are viewed as perfect substitutes by mobile workers (consumers). The developer with the lowest cost (i.e., the best site) cannot capture the whole market because her production cost is convex in city size (because of increasing urban costs). This limits the size of any city. Since all agents eventually have to end up in some location (they can opt out of the urban system, but there are decreasing returns in the rural area), this implies that some of the developers who own worse sites also end up developing them.

is standard in a Roback 1982 equilibrium). A lower value may be justified if property rights for land are weak or if land rents are collected through local property taxes and redistributed locally.[19] Generally, a high value of ρ acts as an entry barrier that decreases incentives to move to large cities, which have more expensive land.[20]

We now demonstrate that, if some portion of land rents in a city accrues to agents that reside outside the city ($\rho > 0$), then there exists a *fiscal externality* between the benefits agents receive and the average benefits produced in a city. When the fiscal externality is sufficiently large, we find that the equilibrium urban system provides too many cities and that the largest cities are undersized.

3.1. Private Equilibrium Allocation

Each agent's utility (2) is a function of the population in the city in which she resides. Agents move freely across their options to maximize their utility, and the resulting utility in any inhabited location is equalized at the spatial equilibrium. Therefore, for all inhabited cities, the equilibrium population is characterized by the condition $v(a,n_p(a)) = \mu_p$, some $\mu_p \geq 0$, where subscript p indicates the private equilibrium allocation.

It is well known that the common utility level μ_p can take different values, thus leading to a continuum of equilibria (Henderson, 1974b). These multiple equilibria can be Pareto-ranked, and we focus on the most efficient, which we call the constrained-efficient equilibrium.[21]

We start by characterizing the extensive margin of urban development at the Pareto efficient spatial equilibrium. Denote by a_p the least productive populated site in equilibrium, namely, $n_p(a) > 0$ if and only if $a > a_p$. Focusing on the constrained efficient equilibrium, this site achieves the locally efficient size as perceived by private agents. From (2), we thus have

$$n_p(a_p) = \left(\frac{a_p\varepsilon}{\gamma}\frac{1}{1+\rho\gamma}\right)^{\frac{1}{\gamma-\varepsilon}}. \tag{20}$$

Note that this perceived efficient size differs from the actual size $n_l(a_p)$ if $\rho > 0$ because the land ownership distorts the perception of urban costs. The extensive margin equilibrium condition en-

[19] Land rights may be weak in developing countries where migrants squat on land (Jimenez, 1984).

[20] Seegert (2011) and Parkhomenko (2018) investigate urban models with housing regulations. Parkhomenko (2018) finds that productive cities endogenously vote for more housing regulations. Seegert (2011) and Salant and Seegert (n.d.) show that housing regulations can be efficient if they act as fees for future residents.

[21] In general, the coordination problem is an important problem in urban economics. For a recent discussion and solution, see Henderson and Venables (2009) and Seegert (2013).

sures the marginal city, a_p, provides residents with the same utility as all other cities and the rural sector:

$$\mu_p = F'(R_p) = v(a_p, n_p(a_p)) = \left(\frac{\gamma}{\varepsilon} - 1\right) n_p(a_p)^{\gamma}, \tag{21}$$

where the last equality comes from our assumption that the smallest city is of constrained efficient size. This condition resembles the central optimum extensive margin condition in equation (17) and differs only in the benefit agents receive in the marginal city, which is distorted by a fiscal externality arising from land ownership that reduces its size (this terminology will become clear in Section 4.2).

We turn next to characterizing the intensive margin of urban development at the Pareto efficient spatial equilibrium. Recall that $u\,(a, n(a))$ denotes the utility level that would prevail if land rents were fully redistributed locally, namely, if $\rho = 0$. The intensive margin condition for a spatial equilibrium can be written as a function of u and a *fiscal externality*,

$$\theta_f \equiv 1 - \frac{1}{1+\rho\gamma} = \frac{\rho\gamma}{1+\rho\gamma} \in [0,\ 1), \tag{22}$$

for any pair of inhabited cities a and b:[22]

$$u(a, n_p(a)) - \theta_f w(a, n_p(a)) = u(b, n_p(b)) - \theta_f w(b, n_p(b)). \tag{23}$$

The fiscal externality is a pecuniary externality capturing land rents (prices), in contrast to the urban and local externalities which are technological externalities capturing changes in agglomeration.

To sum up, land rents, some of which accrue to absentee landlords, create a wedge between the private utility agents receive and the average benefit produced in a city. Pines and Sadka (1986) emphasize the consequences of this assumption for comparative static results of closed cities. Our analysis complements theirs by studying its consequences in a system of open cities. Unlike the effect of federal taxes (Albouy, 2009), this effect has not been widely recognized in the literature so far. In the private free-mobility equilibrium, the fiscal externality causes cities with better amenities

[22] To see this result, note that

$$\begin{aligned}
v(a, n_p(a)) &= a n_p(a)^{\varepsilon} - (1+\rho\gamma) n_p(a)^{\gamma} = (1+\rho\gamma)\left[\frac{a}{1+\rho\gamma} n_p(a)^{\varepsilon} - n_p(a)^{\gamma}\right] \\
&= (1+\rho\gamma)\left[a\left(1 - \frac{\rho\gamma}{1+\rho\gamma}\right) n_p(a)^{\varepsilon} - n_p(a)^{\gamma}\right] = (1+\rho\gamma)\left[u(a, n_p(a)) - \left(1 - \frac{1}{1+\rho\gamma}\right) a n_p(a)^{\varepsilon}\right] \\
&= (1+\rho\gamma)\left[u(a, n_p(a)) - \theta_f w(a, n_p(a))\right] = \mu_p
\end{aligned}$$

Since μ_p is common to all sites, (23) then follows.

to be smaller than they would be without the fiscal externality. The complementarity between the fiscal externality and heterogeneous production amenities causes agents to undervalue the production amenity, the more so the higher the latter. The reason is that sites with high amenities command larger rents, but those rents leave the city and thus reduce consumption benefits there. Consequently, places with high a become disproportionately more expensive places to live in when rents are not rebated to their residents.

3.2. Fiscal-Urban Externalities: Are Cities Too Large? Part 2

We now compare the free-mobility equilibrium with the central optimum to produce our second result that cities can be too small in equilibrium.

Proposition 2 *Consider the socially optimal urban system and the Pareto efficient urban system with free migration. If cities are heterogeneous and the fiscal externality is larger than the urban externality in the sense $\theta_f > \theta_u$. Then **(i)** more cities are inhabited in urban systems with free migration than at the central optimum and **(ii)** the largest cities are undersized in equilibrium.*

Proof. Let $smb(a,n(a)) \equiv (1+\varepsilon)an(a)^{\varepsilon} - (1+\gamma)n(a)^{\gamma}$ denote the marginal social benefit given by equation (13). Evaluating the social marginal benefit in each city at $n_p(a_p)$, using equations (17) (for μ_*) and (20) (for $n_p(a_p)$), and using the definitions of θ_u and θ_f in (14) and (22), respectively, we obtain

$$\frac{smb(a_p,n(a_p))}{\mu_*} = \left(\frac{a_p}{a_*}\right)^{\gamma/(\gamma-\varepsilon)} \frac{\frac{\gamma}{\varepsilon}(1-\theta_u)(1-\theta_f)^{\varepsilon/(\gamma-\varepsilon)} - (1-\theta_f)^{\gamma/(\gamma-\varepsilon)}}{\frac{\gamma}{\varepsilon}(1-\theta_u)-1}. \tag{24}$$

The first ratio on the right-hand side of (24) is greater than unity if and only if a_p exceeds a_*. The second ratio is $\cap$-shaped with respect to $1-\theta_f$. It has a maximum at $1-\theta_f = 1-\theta_u$, is equal to zero at $1-\theta_f = 0$ and $1-\theta_f = \frac{\gamma}{\varepsilon}(1-\theta_u)$, and is equal to one for $1-\theta_f = 1$ and for some value $B \in (0, 1-\theta_u)$. Figure 1 provides an illustration of this function; there, $f(\cdot)$ is the second ratio in the right-hand side of (24).

Insert Figure 1 about here.

If $\theta_f \geq \theta_u$, then the social marginal benefit is increasing in a and large cities are too small. Two configurations need to be considered.

(a) If $B < 1-\theta_f \leq 1-\theta_u$, then $smb(a_p) > \mu_*$ and the social marginal benefit is increasing in a. Hence all cities with $a \geq a_*$ are too small in equilibrium. It follows by the population adding-up constraint that there are too many of them, i.e., $a_* > a_p$ must hold. Cities with productivity between a_p and a_* are thus too big in a trivial sense—because they should not exist—whereas all cities with a productivity above a_* are too small.

(b) If $1-\theta_f < B$, then city sizes are on the increasing range of the social marginal benefit. Thus cities with $a \geq a_*$ are too small, and, to satisfy the adding-up constraint, too many cities form, namely, there exist $a_p \in (0, a_*)$ such that $n_p(a) > n_*(a) = 0$ for all $a \in [a_p, a_*)$.

This completes the proof. □

Proposition 2 shows that cities are not always too large. More precisely, it states that large cities are undersized in the plausible case where $\theta_f > \theta_u$, which arises, e.g., when land rents accrue to agents outside of the city ($\rho = 1$) and cities are heterogeneous.[23] As a consequence, having to pay rent to enter a city reduces the incentives to move to larger cities, which may cause them to be undersized.

Proposition 2 also extends the optimal city size analysis to the extensive margin—will the system provide too many or too few cities? The intensive margin considers agglomeration benefits versus urban costs within each city. The extensive margin mediates the decreasing returns to scale due to congestion within each city with the decreasing returns to scale from inhabiting sites with inferior amenities. Proposition 2 shows that, when the fiscal externality is larger than the urban externality, then the system is distorted toward producing too many cities. Consequently, welfare could improve if agents in the equilibrium allocation abandoned the worst sites and moved to the best sites, despite the total increases in urban costs. Such an overall welfare-improving re-location does not occur in equilibrium: since moving to a high-amenity city entails paying much more in land rent.

4. Characterizing Systems of Cities

This section considers how the city populations vary with (1) heterogeneous productivity, (2) the urban externality, and (3) the fiscal externality. These comparisons contrast with most of the literature that considers cities that are homogeneous or differ by quality yet are infinitely replicable.[24]

[23] The condition $\theta_f > \theta_u$ in Proposition 2 holds if and only if $\rho > (\gamma-\varepsilon)/[\gamma(1+\varepsilon)]$. Hence, it always holds if $\rho \to 1$ since $\gamma > \varepsilon > 0$.

[24] An urban system with replicable heterogeneous cities still displays constant returns in the aggregate when the number of cities is large (see, e.g., Henderson, 1988, p. 176).

We then extend the Henry George Theorem and derive a simple sufficient statistic that can be used to test for the optimality of an urban system with heterogeneous sites.

4.1. The Role of Heterogeneous Sites

This section compares how city populations and utility for the central optimum and free-mobility equilibrium change as production amenities vary. In the central-optimum case, the elasticity of population with respect to productivity is

$$\frac{\mathrm{d}\ln n_*(a)}{\mathrm{d}\ln a} = \frac{1-\theta_u}{\varepsilon} \frac{1}{\left[\frac{n_*(a)}{n_l(a)}\right]^{\gamma-\varepsilon} - (1-\theta_u)} > 0, \tag{25}$$

where the inequality comes from the fact that, for all $a > a_*$, the optimal city size $n_*(a)$ is larger than the efficient scale, $n_l(a)$. In the free mobility case, the percentage change in population with respect to a change in productivity is

$$\frac{\mathrm{d}\ln n_p(a)}{\mathrm{d}\ln a} = \frac{1-\theta_f}{\varepsilon} \frac{1}{\left[\frac{n_p(a)}{n_l(a)}\right]^{\gamma-\varepsilon} - (1-\theta_f)} > 0, \tag{26}$$

where the inequality comes from the fact that for, all $a > a_p$, the free mobility equilibrium size $n_p(a)$ is larger than the efficient scale $n_l(a)$. The derivations of equations (25) and (26) are provided in Appendix A.1.

The differences between these two expressions are the city populations $n_*(a)$ and $n_p(a)$ and, crucially, the urban and fiscal externalities. If the fiscal externality equals the urban externality, then the intensive urban margin will be undistorted, and both systems will exhibit the same elasticity of population with respect to productivity.

To see how utility changes with production amenities in the different allocations, Figure 2 graphs the per-person utility with respect to population for three cities, $a_2 > a_1 > a_*$.[25] For each city, the utility is concave and reaches its peak where it intersects the marginal benefit (dashed line). How utility changes with production amenities differs between the local optimum, central optimum, and free-mobility equilibrium, denoted $u(a, n_l(a))$, $u(a, n_*(a))$, and $u(a, n_p(a))$, respectively.

Insert Figure 2 about here.

[25] For simplicity, we draw that figure for $\theta_f = 0$.

The local-optimum utility has the steepest slope with respect to productivity because the local optimum maximizes average utility in each city. The free-mobility equilibrium utility has the flattest (zero) slope because agents move across these cities until utility is equalized. The central-optimum utility has an intermediate slope because it balances the decreasing returns within each city with the decreasing returns across cities. The local optimum undercrowds cities with better amenities relative to the central optimum because it exclusively focuses on within-city decreasing returns to city size. The free-mobility equilibrium overcrowds cities with better amenities relative to the central optimum because it also focuses on between-city decreasing returns.

4.2. The Role of the Fiscal Externality

This section considers the fiscal externalities arising from needing to pay an entry fee into the city. For simplicity, it is easiest to rebate land rents lump sum, with some transfer T. This does not affect the equilibrium.

For comparison, combine the urban and local externalities into a single urban-to-local externality parameter: $\theta_{ul} \equiv 1-(1-\theta_u)/(1-\theta_l) \in (-\infty, 1)$. It is positive if and only if the urban externality is larger than the local externality (and negative otherwise).

Insert Table 1 about here.

Table 1 characterizes the system of cities under local governments and free mobility, and for different relative sizes of the fiscal and urban externalities. Moving across columns, the fiscal externality gets larger moving from left to right, with the rightmost column reporting the cases in Propositions 1 and 2. The last panel of Table 1 extends Proposition 2, considering parameter spaces featuring too many or too few cities. As the fiscal externality becomes non-negative, the system of cities moves from hosting too few cities to hosting too many of them. When the fiscal externality is larger than the urban-to-local externality, the largest cities become undersized, while the smallest cities become oversized, simply by virtue of existing. Given that migrants usually have to pay more to enter into a larger city in most countries, the two columns on the right of the table present the most realistic scenarios.

With local governments, the transition from too few to too many occurs for a relatively smaller value of the fiscal externality as compared to the free-mobility equilibrium. Furthermore, the largest cities are undersized unless there is a hugely negative fiscal externality: e.g., land is provided for free and additional subsidies — e.g. bread and circuses (Ades and Glaeser, 1995) — are provided. In either case, it seems quite possible that the largest cities will be undersized, especially if local governments have their way.

4.3. A Generalized Henry George Theorem

We now generalize the Henry George Theorem (HGT) to the case with heterogeneous cities. To our knowledge, heterogeneous sites have not been considered in that context until now.[26] The basic HGT states that urban land values equal the agglomeration benefits in cities at the optimal allocation with homogeneous land. Said differently, the agglomeration benefits in a city are capitalized in land rents such that a single confiscatory land tax could finance agglomeration benefits.

In order to operationalize the HGT to design a test of optimal city size, we characterize how the ratio of urban land values to agglomeration externalities is affected by heterogeneous land and the fiscal externality. The ratio of urban land values to agglomeration benefits can be written as a function of wage dispersion in the urban system,

$$\frac{\Delta w}{\overline{w}} \equiv \frac{\overline{w} - w(a_{\min}, n(a_{\min}))}{\overline{w}},$$

where $\overline{w}$ is the average urban wage and $w(a_{\min}, n(a_{\min}))$ is the lowest urban wage at the least productive inhabited site.

Proposition 3 (Generalized Henry George Theorem) *(**i**) In the central optimum, the ratio of urban land values to agglomeration externalities is weakly greater than one and equal to*

$$\eta_* \equiv \frac{\displaystyle\int_{a_*}^{\infty} \gamma n(a)^{\gamma+1} \mathrm{d}G(a)}{\displaystyle\int_{a_*}^{\infty} \varepsilon a n(a)^{\varepsilon+1} \mathrm{d}G(a)} = 1 + \frac{\theta_u}{\varepsilon} \frac{\overline{w} - w(a_*, n_*(a_*))}{\overline{w}} \geq 1. \tag{27}$$

*(**ii**) In the free-mobility equilibrium with a fiscal externality θ_f, the ratio of urban land values to*

[26] Arnott (2004, p.1072) states that “The Henry George Theorem is derived on the assumption that land is homogeneous, but in reality locations differ. ... How do these Ricardian differences in land affect the theorem qualitatively, and how important are they quantitatively? To my knowledge, this question has not been investigated in the literature.”

agglomeration externalities is equal to

$$\eta_p \equiv \frac{\displaystyle\int_{a_p}^{\infty} \gamma n(a)^{\gamma+1} \mathrm{d}G(a)}{\displaystyle\int_{a_p}^{\infty} \varepsilon a n(a)^{\varepsilon+1} \mathrm{d}G(a)} = (1-\theta_f)\left[1+\frac{(1+\gamma)\theta_u}{\varepsilon}\frac{\overline{w}-w(a_p,n_p(a))}{\overline{w}}\right]. \tag{28}$$

(iii) *If* $\eta_* > \eta_p$ *then the equilibrium system of cities allocates too little population to the most productive sites.*

Proof. See Appendix A.2. □

When cities are homogeneous, the average and lowest urban wage are identical and $\eta_* = 1$ holds, replicating the basic Henry George Theorem. When cities are heterogeneous, the average wage is higher than the lowest wage, and land values exceed urban agglomeration benefits. In this case, land rents capture the agglomeration benefit and heterogeneity in the production amenity. Diminishing returns at both the intensive and extensive margin contribute to higher land values. As a result, when cities are heterogeneous, a confiscatory land tax is more than sufficient to subsidize the agglomeration benefits of cities.

In principle, one could try to calculate how high land values in the central optimum would be relative to the free-mobility equilibrium, η_*/η_p. This would provide an indicator of whether large cities are too small. It is possible that a system of free migration could be optimal, in which case the ratio will be one. If cities are too small, the ratio will be less than one.

With local governments, the ratio of land values to agglomeration economies is $\eta_l = 1-\theta_f$. The condition that land values are too low with local governments, $1-\theta_f \leq \eta_*$ should hold for all but unrealistically negative values of θ_f.

5. Conclusion

https://v2.overleaf.com/project/5b5da072863c17647cd982f4 We provide a comprehensive, yet concise, analysis of optimal city sizes in a framework allowing for heterogeneous sites, fiscal externalities (including land ownership rules), and an endogenously determined number of cities. Within that framework, we show that large cities may be underpopulated. We highlight the role that agglomeration, urban costs, and the distribution of land rents play in generating that outcome.

Optimal city sizes balance decreasing returns to scale within *and* across cities. This trade-off produces several policy insights. First, local governments distort the trade-off between within- and across-city decreasing returns. Second, because local governments focus solely on within-city returns, all agents perceive their city as being too large in the central optimum. All cities but the smallest are larger than their locally efficient scale. By contrast, in an equilibrium with free mobility, the system of cities focuses solely on across-city decreasing returns. This outcome suggests that urban policies should be coordinated by a federal government that balances the within-city cost concerns of with the costs of inhabiting inferior sites.

We further show that a fiscal externality arises when any new-comers must pay, on net, for the land they inhabit. This externality distorts the distribution of population. The often-made assumption that land rents accrue to residents solely within each city implicitly assumes that migrants to a city are given the average land value as a welcome gift. This is an unrealistic assumption for developed countries with strong property rights. Relaxing this assumption causes the most productive cities to become underpopulated.

Last, we extend the Henry George Theorem to cases with heterogeneous cities. This provides an indicator for whether the most productive cities in a system of cities are oversized. When cities are heterogeneous, a confiscatory land tax is more than sufficient to finance agglomeration externalities. Furthermore, the wage dispersion across cities can be used to test whether the ratio of urban land values to urban agglomeration is optimal.

Building a city systems model that allows for fiscal externalities and heterogeneous urban amenities is notoriously difficult (Henderson, 1988). We solve the problem by taking a continuous approach and using parsimonious functional forms. We also assume away network effects by assuming that inter-city trade costs and city sizes do not interact: a low city-specific productivity parameter in our model may be interpreted as a city that is isolated from all other cities in general, but its vector of bilateral distances to other cities does not matter. We also assume that all agents are homogeneous. The canonical Henderson (1974a) model of systems of cities has homogeneous sites but often multiple sectors, and in equilibrium cities specialize in different sectors with those specializing in sectors with stronger agglomeration economies being more populated. Extending our model to allow for complex geographies, as do Allen and Arkolakis (2014), or for agents with heterogeneous talents, as do Behrens et al. (2014), or for differences in quality-of-life, as does Albouy (2008), or with multiple sectors featuring different agglomeration, are all promising venues for further research.

References

Abdel-Rahman, Hesham M., "Product differentiation, monopolistic competition and city size," *Regional Science and Urban Economics*, 1988, *18* (1), 69–86.

Abdel-Rahman, Hesham M. and Alex Anas, "Theories of systems of cities," in J. Vernon Henderson and Jacques-François Thisse, eds., *Handbook of Regional and Urban Economics*, Vol. 4, North-Holland: Elsevier B.V., 2004, pp. 2293–2339.

Ades, Alberto F. and Edward L. Glaeser, "Trade and circuses: explaining urban giants," *The Quarterly Journal of Economics*, 1995, *110* (1), 195–227.

Albouy, David, "Are big cities bad places to live? Estimating quality of life across metropolitan areas," Technical Report, National Bureau of Economic Research, No. w14472 2008.

Albouy, David, "The unequal geographic burden of federal taxation," *Journal of Political Economy*, 2009, *117* (4), 635–667.

Albouy, David, "Evaluating the efficiency and equity of federal fiscal equalization," *Journal of Public Economics*, 2012, *96* (9-10), 824–839.

Albouy, David and Nathan Seegert, "Optimal city size and the private-social wedge," in "46th Annual AREUEA Conference Paper," 2011.

Albouy, David, Kristian Behrens, Frédéric Robert-Nicoud, and Nathan Seegert, "The optimal distribution of population across cities," *National Bureau of Economic Research*, 2016, *No. 22823.*

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

Alonso, William, *Location and land use: Toward a general theory of land rent*, Vol. 204, Harvard University Press Cambridge, MA, 1964.

Arnott, Richard, "Optimal city size in a spatial economy," *Journal of Urban Economics*, 1979, *6* (1), 65–89.

Arnott, Richard, "Does the Henry George Theorem provide a practical guide to optimal city size?," *American Journal of Economics and Sociology*, 2004, *63* (5), 1057–1090.

Au, Chun-Chung and J. Vernon Henderson, "Are Chinese cities too small?," *The Review of Economic Studies*, 2006, *73* (3), 549–576.

Behrens, Kristian and Frédéric Robert-Nicoud, "Agglomeration theory with heterogeneous agents," in Gilles Duranton, Vernon J. Henderson, and William C. Strange, eds., *Handbook of Regional and Urban Economics*, Vol. 5, North-Holland: Elsevier B.V., 2015, pp. 171–244.

Behrens, Kristian and Frédéric Robert-Nicoud, "Are cities too small? Equilibrium and optimal urban systems with heterogenoues land," Mimeographed 2015.

Behrens, Kristian, Gilles Duranton, and Frédéric Robert-Nicoud, "Productive cities: Sorting, selection, and agglomeration," *Journal of Political Economy*, 2014, *122* (3), 507–553.

Borck, Rainald and Takatoshi Tabuchi, "Pollution and city size: Can cities be too small?," *Journal of Economic Geography*, forthcoming.

Buchanan, James M. and Charles J. Goetz, "Efficiency limits of fiscal mobility: An assessment of the Tiebout model," *Journal of Public Economics*, 1972, *1* (1), 25–43.

Cheshire, Paul and Stephen Sheppard, "The welfare economics of land use planning," *Journal of Urban Economics*, 2002, *52* (2), 242–269.

Combes, Pierre-Philippe and Laurent Gobillon, "The empirics of agglomeration economies," in Gilles Duranton, Vernon J. Henderson, and William C. Strange, eds., *Handbook of Regional and Urban Economics*, Vol. 5, North-Holland: Elsevier B.V., 2015, pp. 447–348.

Combes, Pierre-Philippe, Gilles Duranton, and Laurent Gobillon, "The costs of agglomeration: House and land prices in French cities," *CEPR Discussion Paper No. DP9240 (updated)*, 2018.

Desmet, Klaus and Esteban Rossi-Hansberg, "Urban accounting and welfare," *The American Economic Review*, 2013, *103* (6), 2296–2327.

Duranton, Gilles and Diego Puga, "Micro-foundations of urban agglomeration economies," in J. Vernon Henderson and Jacques-François Thisse, eds., *Handbook of Regional and Urban Economics*, Vol. 4, North-Holland: Elsevier B.V., 2004, pp. 2063–2117.

Duranton, Gilles and Diego Puga, "Urban land use," in Gilles Duranton, Vernon J. Henderson, and William C. Strange, eds., *Handbook of Regional and Urban Economics*, Vol. 5, North-Holland: Elsevier B.V., 2015, pp. 476–560.

Eeckhout, Jan, "Gibrat's law for (all) cities," *American Economic Review*, 2004, *94* (5), 1429–1450.

Eeckhout, Jan and Nezih Guner, "Optimal spatial taxation: Are big cities too small?," *mimeographed*, 2017.

Fenge, Robert and Volker Meier, "Why cities should not be subsidized," *Journal of Urban Economics*, 2002, *52* (3), 433–447.

Flatters, Frank, J. Vernon Henderson, and Peter Mieszkowski, "Public goods, efficiency, and regional fiscal equalization," *Journal of Public Economics*, 1974, *3* (2), 99–112.

Fujita, Masahisa, *Urban Economic Theory: Land Use and City Size*, Cambridge Univ Pr, 1989.

Glaeser, Edward L., Joseph Gyourko, and Raven Saks, "Why is Manhattan so expensive? Regulation and the rise in housing prices," *The Journal of Law and Economics*, 2005, *48* (2), 331–369.

Harris, John R and Michael P Todaro, "Migration, unemployment and development: A two-sector analysis," *The American Economic Review*, 1970, *60* (1), 126–142.

Helpman, Elhanan and David Pines, "Optimal public investment and dispersion policy in a system of open cities," *The American Economic Review*, 1980, *70* (3), 507–514.

Henderson, J Vernon, "Optimum city size: the external diseconomy question," *Journal of Political Economy*, 1974, *82* (2, Part 1), 373–388.

Henderson, J. Vernon, "The sizes and types of cities," *The American Economic Review*, 1974, *64* (4), 640–656.

Henderson, J. Vernon, "Efficiency of resource usage and city size," *Journal of Urban Economics*, 1986, *19* (1), 47–70.

Henderson, J. Vernon, *Urban Development: Theory, Fact, and Illusion*, Oxford Unviersity Press, 1988.

Henderson, J. Vernon and Anthony J. Venables, "The dynamics of city formation," *Review of Economic Dynamics*, 2009, *12* (2), 233–254.

Henderson, J Vernon and Randy Becker, "Political economy of city sizes and formation," *Journal of Urban Economics*, 2000, *48* (3), 243–484.

Hilber, Christian A. and Frédéric Robert-Nicoud, "On the origins of land use regulations: Theory and evidence from US metro areas," *Journal of Urban Economics*, 2013, *75*, 29–43.

Hsieh, Chang-Tai and Enrico Moretti, "Why do cities matter? Local growth and aggregate growth," *American Economic Journal: Macroeconomics*, forthcoming.

Jimenez, Emmanuel, "Tenure security and urban squatting," *The Review of Economics and Statistics*, 1984, *66* (4), 556–567.

Kanemoto, Yoshitsugu, *Theories of Urban Externalities*, North-Holland, Amsterdam, 1980.

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

Mirrlees, James A., "Migration and optimal income taxes," *Journal of Public Economics*, 1982, *18* (3), 319–341.

Molotch, Harvey, "The city as a growth machine: Toward a political economy of place," *American Journal of Sociology*, 1976, *82* (2), 309–332.

Muth, Richard F., *Cities and Housing*, Chicago, 1969.

O'Sullivan, Arthur, *Urban Economics*, McGraw-Hill/Irwin, 2007.

Parkhomenko, Andrii, "Housing supply regulation: Local causes and aggregate implications," Technical Report, University of Southern California 2018.

Pines, David and Efraim Sadka, “Comparative statics analysis of a fully closed city,” *Journal of Urban Economics*, 1986, *20* (1), 1–20.

Redding, Stephen J., “Goods trade, factor mobility and welfare,” *Journal of International Economics*, 2016, *101*, 148–167.

Roback, Jennifer, “Wages, rents, and the quality of life,” *The Journal of Political Economy*, 1982, *90* (6), 1257–1278.

Salant, Stephen and Nathan Seegert, “Should Congestion Tolls be Set by the Government or by the Private Sector? The Knight-Pigou Debate Revisited,” *Economica*, *85*, 428–448.

Seegert, Nathan, “Barriers to migration in a system of cities,” Technical Report, Social Science Research Network, No. 2557399 2011.

Seegert, Nathan, “Rushing to opportunities: a model of entrepreneurship and growth,” Technical Report, Social Science Research Network, No. 2857105 2013.

Stiglitz, Joseph E., “The theory of local public goods,” in “The Economics of Public Services,” Springer, 1977, pp. 274–333.

Tolley, George S., “The welfare economics of city bigness,” *Journal of Urban Economics*, 1974, *1* (3), 324–345.

Upton, Charles, “An equilibrium model of city size,” *Journal of Urban Economics*, 1981, *10* (1), 15–36.

Vermeulen, Wouter, “Agglomeration externalities and urban growth controls,” *Journal of Economic Geography*, 2017, pp. 59–94.

Vickrey, William, “The city as a firm,” in “The Economics of Public Services,” Springer, 1977, pp. 334–343.

Appendix A. Proofs

Appendix A.1. City Size is Increases with a

We begin by showing that city size increases with a in the central optimum allocation. More precisely, we establish

$$\frac{\mathrm{d}\ln(n_*(a))}{\mathrm{d}\ln(a)} = \frac{1-\theta_u}{\varepsilon} \frac{1}{\left[\frac{n_*(a)}{n_l(a)}\right]^{\gamma-\varepsilon} - (1-\theta_u)} > 0,$$

from Section 4. To see that city size is increasing in a, totally differentiate the first-order condition for all inhabited sites,

$$a(1+\varepsilon)n_*(a)^{\varepsilon} - (1+\gamma)n_*(a)^{\gamma} = \mu_*,$$

given in equation (13), and recall that μ_* is constant across all populated sites, to obtain

$$(1+\varepsilon)n_*(a)^{\varepsilon} + \frac{\mathrm{d}n_*}{\mathrm{d}a}\frac{a}{n_*(a)}\left[\varepsilon(1+\varepsilon)n_*(a)^{\varepsilon} - \frac{\gamma(1+\gamma)}{a}n_*(a)^{\gamma}\right] = 0,$$

which in turn yields

$$\frac{\mathrm{d}n_*}{\mathrm{d}a}\frac{a}{n_*(a)} = \frac{1}{\gamma\left(\frac{1}{a}\frac{1+\gamma}{1+\varepsilon}\right)n_*(a)^{\gamma-\varepsilon} - \varepsilon}. \tag{A-1}$$

Since $\theta_u = 1-(1+\varepsilon)/(1+\gamma)$, equation (A-1) can be rewritten as

$$\begin{aligned}\frac{\mathrm{d}n_*}{\mathrm{d}a}\frac{a}{n_*(a)} = \frac{\mathrm{d}\ln(n_*(a))}{\mathrm{d}\ln(a)} &= \frac{(1-\theta_u)}{\varepsilon\frac{1}{a}\frac{\gamma}{\varepsilon}n_*(a)^{\gamma-\varepsilon} - \varepsilon(1-\theta_u)} \\ &= \frac{1-\theta_u}{\varepsilon}\frac{1}{\left[\frac{n_*(a)}{n_l(a)}\right]^{\gamma-\varepsilon} - (1-\theta_u)} > 0,\end{aligned}$$

where we have used the definition of the locally efficient size $n_l(a)^{\gamma-\varepsilon} = a\varepsilon/\gamma$. The inequality follows from $\theta_u \in (0,1)$ and $n_*(a) > n_l(a)$ from Proposition 1.

We next establish that city size is increasing in a in the private equilibrium allocation. More precisely, we establish

$$\frac{\mathrm{d}\ln(n_p(a))}{\mathrm{d}\ln(a)} = \frac{1-\theta_f}{\varepsilon} \frac{1}{\left[\frac{n_p(a)}{n_l(a)}\right]^{\gamma-\varepsilon} - (1-\theta_f)} > 0,$$

from Section 4. To do so, fully differentiate the free mobility condition $v(a) = an_p(a)^{\varepsilon} - (1+$

$\rho\gamma)n_p(a)^\gamma = \bar{v}$, where the latter is constant with respect to a. This yields:

$$\begin{aligned}\frac{\mathrm{d}n_p}{\mathrm{d}a}\frac{a}{n_p(a)} = \frac{\mathrm{d}\ln(n_p(a))}{\mathrm{d}\ln(a)} &= \frac{1}{\varepsilon}\frac{1}{\frac{\gamma}{a\varepsilon}n_p(a)^{\gamma-\varepsilon}(1+\rho\gamma)-1} \\ &= \frac{1}{\varepsilon(1+\rho\gamma)}\frac{1}{\frac{\gamma}{a\varepsilon}n_p(a)^{\gamma-\varepsilon}-\frac{1}{(1+\rho\gamma)}} \\ &= \frac{1-\theta_f}{\varepsilon}\frac{1}{\left[\frac{n_p(a)}{n_l(a)}\right]^{\gamma-\varepsilon}-(1-\theta_f)} > 0,\end{aligned}$$

since $1+\rho\gamma = \frac{1}{1-\theta_f}$. The inequality follows from $\theta_u \in (0,1)$ and $n_p(a) > n_l(a)$ at any stable private equilibrium with free migration.

Appendix A.2. Proof of Proposition 3.

In this appendix, we establish the expressions for our generalized Henry George Theorem. The expression of the ratio of land values to agglomeration benefits in (27) can be computed as follows. First, note that the denominator provides the sum of income:

$$\varepsilon\int_{a_*}^{\infty} an(a)^{1+\varepsilon}dG(a) = \varepsilon\int_{a_*}^{\infty} w(a)n(a)dG(a) = \varepsilon N\overline{w}_*.$$

The numerator is solved by noting that we can write (15) and (17) together as

$$n_*(a)^\gamma = n_*(a_*)^\gamma + (1-\theta_u)\left[w_*(a) - w_*(a_*)\right],$$

where $w_*(a) \equiv w(a, n_*(a))$ and $w_p(a) \equiv w(a, n_p(a))$ in what follows to alleviate notation. Furthermore, by (17), per-capita urban costs in the marginal city is

$$n_*(a_*)^\gamma = \left(a\frac{\varepsilon}{\gamma}\right)^{\frac{\gamma}{\gamma-\varepsilon}} = \frac{\varepsilon}{\gamma}a_* n_*(a_*)^\varepsilon = \frac{\varepsilon}{\gamma}w_*(a_*).$$

Putting these expressions together leads to the expression

$$\begin{aligned}\gamma n_*(a)^\gamma &= \gamma n_*(a_*)^\gamma + \gamma(1-\theta_u)\left[w_*(a) - w_*(a_*)\right] \\ &= \varepsilon w_*(a_*) + \gamma(1-\theta_u)\left[w_*(a) - w_*(a_*)\right] \\ &= \varepsilon w_*(a) + (\gamma(1-\theta_u)-\varepsilon)w_*(a) - (\gamma(1-\theta_u)-\varepsilon)w_*(a_*) \\ &= \varepsilon w_*(a) + \theta_u w_*(a) - \theta_u w_*(a_*),\end{aligned}$$

where we use the transformation,

$$\gamma(1-\theta_u) - \varepsilon = \gamma\frac{1+\varepsilon}{1+\gamma} - \varepsilon = \frac{\gamma}{1+\gamma} - \frac{\varepsilon}{1+\gamma} = 1 - \frac{\varepsilon+1}{1+\gamma} = \theta_u.$$

Therefore the numerator can be written in an easy to integrate form

$$\int_{a_*}^{\infty} n_*\left[\varepsilon w_*(a) + \theta_u w_*(a) - \theta_u w_*(a_*)\right] \mathrm{d}G(a) = N\left[\varepsilon \bar{w}_* + \theta_u \overline{w}_* - \theta_u w_*(a_*)\right]$$

The ratio of the numerator to the denominator is then

$$\eta_* = \frac{N\left[\varepsilon \overline{w}_* + \theta_u \overline{w}_* - \theta_u w_*(a_*)\right]}{\varepsilon N \bar{w}_*} = 1 + \frac{\theta_u}{\varepsilon}\frac{\overline{w}_* - w_*(a_*)}{\overline{w}_*},$$

which is the expression in equation (27).

The ratio of agglomeration to congestion in the private equilibrium is found through similar steps. The denominator is $\varepsilon N \overline{w}_p$. The numerator can be written in a convenient form using equations (23) and (21). First, note that congestion costs in city a can be written as

$$\gamma n_p(a)^\gamma = \gamma n_p(a_p)^\gamma + \gamma(1-\theta_f)\left[w_p(a) - w_p(a_p)\right].$$

Using equation (21), the population in the least productive city can be written as

$$n_p(a_p)^\gamma = \left[\frac{\varepsilon a_p}{\gamma(1+\rho\gamma)}\right]^{\gamma/(\gamma-\varepsilon)} = \frac{\varepsilon}{\gamma(1+\rho\gamma)} a_p \left[\frac{\varepsilon a_p}{\gamma(1+\rho\gamma)}\right]^{\varepsilon/(\gamma-\varepsilon)} = \frac{\varepsilon}{\gamma(1+\rho\gamma)} w_p(a_p).$$

Congestion costs can then be reduced to

$$\begin{aligned}\gamma n_p(a)^\gamma &= \gamma n_p(a_p)^\gamma + \gamma(1-\theta_f)\,[w_p(a) - w_p(a_p)]\\ &= \frac{\varepsilon}{(1+\rho\gamma)} w_p(a) + \left[\gamma(1-\theta_f) - \frac{\varepsilon}{(1+\rho\gamma)}\right](w_p(a) - w_p(a_p))\\ &= \varepsilon(1-\theta_f) w_p(a) + (1-\theta_f)\theta_u(1+\gamma)[w_p(a) - w_p(a_p)].\end{aligned}$$

Therefore the numerator can be written in an easy to integrate form

$$\begin{aligned}&\int_{a_p}^{\infty} n_p\left[\varepsilon(1-\theta_f) w_p(a) + (1-\theta_f)\theta_u(1+\gamma)(w_p(a) - w_p(a_p))\right] dG(a)\\ &\qquad = N\left[(1-\theta_f)\varepsilon\overline{w}_p + (1-\theta_f)\theta_u(1+\gamma)(\overline{w}_p - \theta_u w_p(a_p))\right].\end{aligned}$$

The ratio of the numerator to the denominator is

$$\eta_p = (1-\theta_f)\left[1 + \frac{\theta_u(1+\gamma)}{\varepsilon}\frac{\overline{w}_p - w_p(a_p)}{\overline{w}_p}\right],$$

which is the expression in equation (28).

Appendix B. Competitive land developers, general case

We present the case with competitive developers and federal taxes and subsidies. Local land developers attract urban dwellers by offering (possibly negative) subsidies, s, such that the net utility is weakly larger than the economy-wide utility level, denoted by u_d (where subscript d stands for developers). They collect aggregate land rents in the site they develop. Each developer takes the redistribution of land rents ρ, possibly to a federal government, as given. A developer who owns site a solves the following maximization problem:

$$\begin{aligned}\max_{n(a),s}\ \pi(n(a),s;a,u_d) &= \gamma(1-\rho)n(a)^{\gamma+1} - n(a)s\\ \text{s.t.}\quad u_d &\le an(a)^\varepsilon - (1+\gamma)n(a)^\gamma + s,\end{aligned}$$

where the first term in π is the aggregate land rent collected and the second term is the cost of the subsidy to attract $n(a)$ people to the site. Local land developers set agents at their reservation utility, i.e., $s = u_d - an(a)^\varepsilon + (1+\gamma)n(a)^\gamma$. Plugging s into the profit and optimizing with respect to $n(a)$, the first-order condition is given by

$$a(1+\varepsilon)n(a)^{\varepsilon}-(1+\rho\gamma)(1+\gamma)n(a)^{\gamma}=u_d. \tag{C-1}$$

Note that, in the presence of fiscal externalties, $\rho \neq 0$, (C-1) and (13) no longer coincide. In that case, the allocation with developers is no longer optimal. The reason is that the developers take into account the fiscal externalities when attracting people to their sites. When ρ is positive, the subsidy that would be optimal from a social perspective would not be most profitable. The central authority eats into the developer's incentives to make land rents large enough. Hence large cities may end up being too small.

Tables and Figures

Table 1
Characterizing the System of Cities.

Types of sites	Any	Any	Homog.	Heterog.	Any	Any
	$\theta_f < \theta_{ul}$	$\theta_f \in (\theta_{ul}, 0]$	$\theta_f = 0$		$\theta_f \in (0, \theta_{ul}]$	$\theta_f > \theta_{ul}$
Largest Cities						
Free Mobility	oversized	oversized	optimal	oversized	oversized	undersized
Local Governments	oversized	undersized	optimal	undersized	undersized	undersized
Smallest Cities						
Free Mobility	oversized	oversized	optimal	undersized	undersized	oversized
Local Governments	oversized	oversized	optimal	oversized	oversized	oversized
Number of Cities						
Free Mobility	too few	too few	optimal	too few	ambiguous	too many
Local Governments	too few	ambiguous	optimal	too many	too many	too many

Notes: This table extends Proposition 2 by characterizing urban systems with different levels of fiscal externality. The fiscal externality gets larger from left to right.

Figure 1. Shape of $f(1-\theta)$ for the proof of Proposition 2.
This figure depicts the relationship between the second ratio in equation (A-2), given by $f(1-\theta_f)$ and $1-\theta_f$. The second ratio peaks at $1-\theta_u$ and is zero at zero and $\gamma/\varepsilon(1-\theta_u)$.

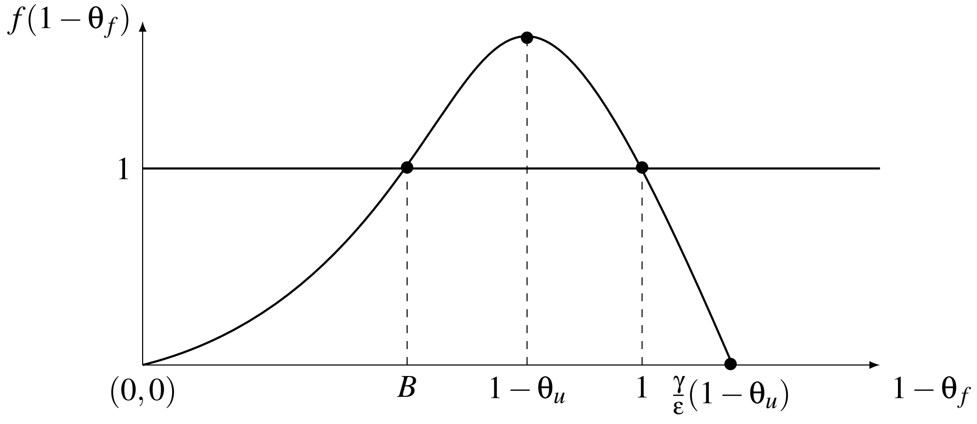
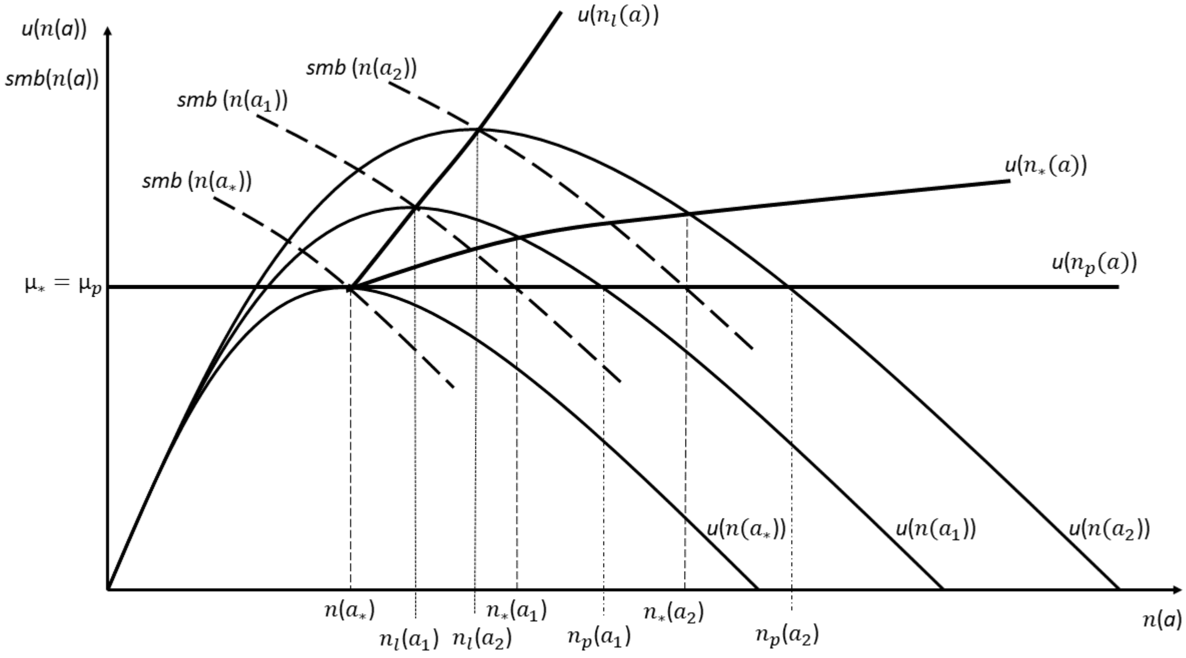


Figure 2. Utility and Production Amenities, in Different Systems of Cities

This figure depicts the average and marginal utility of three cities with respect to population, depicted as a solid line $u(n(a))$ and dashed line $mb(n(a))$, respectively. The central government, local government, and free mobility equilibrium allocate different populations to cities 1 and 2, depicted as $n_*(a)$, $n_l(a)$, and $n_p(a)$, respectively. The different populations cause the average utility in cities 1 and 2 to differ, depicted as $u(n_*(a))$, $u(n_l(a))$, and $u(n_p(a))$, respectively.



$u(n(a))$
$smb(n(a))$
$u(n_l(a))$
$smb\ (n(a_2))$
$smb\ (n(a_1))$
$smb\ (n(a_*))$
$u(n_*(a))$
$u(n_p(a))$
$\mu_* = \mu_p$
$u(n(a_*))$
$u(n(a_1))$
$u(n(a_2))$
$n(a_*)$
$n_*(a_1)$
$n_*(a_2)$
$n_l(a_1)$
$n_l(a_2)$
$n_p(a_1)$
$n_p(a_2)$
$n(a)$