CERGE
Center for Economics Research and Graduate Education
Charles University Prague

Essays in Economic Theory

Maxim Goryunov

Dissertation

Prague, October 2016

Maxim Goryunov

Essays in Economic Theory

Dissertation

Prague, October 2016

Dissertation Committee

Jakub Steiner (CERGE-EI and University of Edinburgh; chair)
Filip Matějka (CERGE-EI)
Sergey Slobodyan (CERGE-EI)

Referees

Michal Fabinger (University of Tokyo)
Philipp Kircher (University of Edinburgh)

Contents

Abstract

In the first chapter of this work, I study the sorting of workers to firms, when firm size is explicitly taken into account. I develop a method to *non-parametrically* identify match production function from data on workers' wages and firms' revenues and posted job vacancies. Under the proposed identification procedure, ordering of workers and firms is identified independently, and can therefore be achieved using potentially different data sets. The model sheds light on the question of exporter wage premium: exporters pay higher wages because they are larger, and higher wages are required to support a larger firm size.

In the second chapter we elaborate on Anas' (2004) impossibility theorem, which states that monopolistic competition or economies of scale alone are insufficient to explain the growth of cities in response to growing population or decreasing trade costs (under constant urban costs); cities shrink. To enhance the realism of assumptions, instead of Anas' normative approach, we introduce migration and developers' equilibria and another sector. Still, "vanishing" remains robust! Ultimately, we argue that the "vanishing" mechanism looks realistic and can have an explanatory power: industries, free of externalities, *should* locate in small towns. Moreover, the comparative statics shows *how* such "manufacturing" towns gradually decline, whereas other cities do not.

In the last chapter, to enrich the usual monopolistic competition model, we combine it with a space of product characteristics, i.e., consumers' "ideal varieties". Unlike Hotelling, in our *partially localized competition*, zones of service among continuously distributed producers intersect due to love for variety. When the equilibrium density of firms is uniform, it reacts positively to growing market size (population), similarly to non-localized monopolistic competition. However, positive/negative price reaction is now determined by increasing/decreasing elasticity of elementary utility (instead of demand elasticity as in non-localized competition). The firm's *range of service* is a new notion introduced in this work. When a firm does not serve all the consumers, the range of service decreases with the expansion of the market.

Abstrakt

V první kapitole této práce studuji způsob, jakým se pracovníci třídí do firem, přičemž je explicitně brána v potaz velikost firmy. Vytvořil jsem metodu, kterou lze neparametricky identifikovat párovací produkční funkci z údajů o mzdách pracovníků a výnosů a volných pracovních míst firem. Podle navrhované identifikační metody řazení pracovníků a firem je identifikováno nezávisle, a proto může být dosaženo použitím potenciálně různých datových souborů. Tento model vrhá světlo na otázku mzdové prémie vývozců: vývozci platí vyšší mzdy, protože jsou větší, kde větší mzdy jsou vyžadovány pro podporu větší velikosti firmy.

Ve druhé kapitole rozpracováváme Anasův (2004) teorém impossibility, který praví, že pouhá monopolistická konkurence nebo výnosy z rozsahu samy o sobě nestačí k vysvětlení růstu měst v reakci na rostoucí populaci nebo pokles obchodních nákladů (při konstantních městských nákladech); města se zmenšují. Pro větší realističnost předpokladů představíme migraci nebo vývojářské rovnováhy a další sektor namísto Anasova normativního přístupu. Výsledek mizících měst přesto přetrvává! Nakonec polemizujeme, že mechanizmus mizení vypadá realisticky a může mít vypovídací hodnotu: průmysly, které nemají žádné externality, by se měli usadit v malých městech. Komparativní statika navíc ukazuje, jak se taková průmyslová města postupně zmenšují, zatímco ostatní města ne.

V poslední kapitole obohacujeme obvyklý model monopolistické konkurence o prostor charakteristik produktů, tzn. ideálů jednotlivých spotřebitelů. Narozdíl od Hotellingova modelu, v naší čsátečně lokalizované konkurenci se zóny služeb mezi spojitě rozloženými producenty neprotínají díky touze po rozmanitosti. V případě, kdy je rovnovážná hustota firem rovnoměrná, reaguje pozitivně na rostoucí velikost trhu (populaci), podobně jako v nelokalizované verzi monopolistické konkurence. Nicméně, pozitivní/negativní reakce ceny je nyní určována rostoucí/klesající elasticitou elementárního užitku (namísto elasticity poptávky v nelokalizované konkurenci). Novým zaměřením studie je také rozsah slu{zeb firmy, pokud nepokrývá celý prostor, který klesá s rostoucím trhem.

Acknowledgments

I would like to express my gratitude to a large number of people who in many different ways contributed to this dissertation.

I am immensely grateful to my supervisor, Jakub Steiner. His encouragement, guidance and challenging questions and discussions shaped my vision of economics and this thesis, in particular, to an extent that can hardly be overestimated. I would also like to thank the other members of my dissertation committee, Filip Matějka and Sergey Slobodyan, for their advice and comments that helped me improve my work.

I am thankful to my co-authors: Sergey Kokovin and Takatoshi Tabuchi, without whom this dissertation would not be out there.

I would like to thank everyone at CERGE-EI - professors, my fellow students and staff members — who made my five years here an unforgettable experience. Part of this research was carried out during my stay at the University of Edinburgh in the winter of 2015. I would like to express my gratitude to the School of Economics for its enormous hospitality.

I benefited from numerous discussions of different parts of this thesis with many other academics. I am grateful to Alp Atakan, Kristian Behrens, Marcus Berliant, Michal Fabinger, Sergey Kichko, Philipp Kircher, Kenneth Mirkin, Yasusada Murata, Jacques-Francois Thisse, Ludo Visschers, Philip Ushchev, as well as participants in different seminars, conferences and workshops, in which I had opportunities to present my work.

I am indebted to the members of Academic Skills center: Andrea Downing, Deborah Nováková and Paul Whitaker, who helped to turn my dissertation into readable text, and to Ludmila Matysková, who prepared the Czech version.

Last but not least, I would like to thank my friends outside of CERGE-EI, economics and academia for their support and encouragement during these years.

Needless to say, all errors remaining in the text are my responsibility.

Florence, Italy
October 14, 2016

Max

Introduction

My dissertation is titled ”Essays in Economic Theory", and although the three chapters differ substantially in their topic and focus, two unifying themes permeate the dissertation: one is conceptual, the other is instrumental.

On the conceptual level, the dissertation is focused on the different instances of market frictions. Since my acquaintance with the Welfare Theorems, I have become at the same time fascinated with and amused by the idea that under certain conditions market systems deliver the first-best allocation. My fascination comes from the elegance of the argument and the unanticipated connection between the self-interested agents premise and its welfare implication. My amusement reposes on the further inference that uniform applicability of this result should have left economists, both in academia and policy-making, unemployed and even unemployable. These and similar ideas fed my interest in studying market imperfections from the early stages of my career, and brought me to the understanding that, to a large extent, Economic Theory is a theory of frictional interactions and market failures.

The methodological tool that connects the three chapters is the theory of monopolistic competition. That it can be considered a theory of market failure can partially explain my passion for it. When I learned the monopolistic competition theory during my master studies, it immediately struck me as an appealing concept for modeling a wide range of markets. The simultaneous presence of local market power, and the absence of direct strategic interactions seemed to me a natural proxy description of the supply side of the economy. For that reason, whenever faced with modeling a market, I always started with the monopolistic competition approach, and it has always served me well so far.

Therefore, each essay of my dissertation can be viewed as an applied study of market frictions and their impact on the equilibrium allocations: on the labor market, in economic geography and in industrial organization.

In the first chapter I study matching of workers and firms on the labor market when both sides of the market are heterogeneous and the search process is impeded by frictions. In contrast to the large body of literature, which relies on a one-to-one matching framework and equates jobs to firms, I explicitly take into account firm size and the firm's ability to choose not only the type, but also the number, of workers it wants to employ. I develop a procedure that allows for non-parametric identification of match production function and vacancy creation cost function from workers' wages and firms' revenues and profits. An important insight from the developed model is that disregard of firm size leads to underestimation of the complementarity between the ability of workers and the productivity of firms, and therefore, of potential output gains from better sorting. In addition, I extend the model into the international trade setting to address the question of the exporting wage premium — an empirical observation that exporting firms tend to pay higher wages. In my model, exporters pay higher wages due to the size effect — they are larger, and under search frictions, higher wages are required to support larger firm size.

In the second chapter (jointly with Sergey Kokovin), we study what monopolistic competition has to say about the evolution of city sizes. We focus on the question of whether internal economies of scale, juxtaposed with the interplay between congestion costs within cities and transportation costs between them, can generate increasing city size, when the population of the system is growing and creation of cities is endogenous. Our answer is negative. We show that, by themselves, internal economies of scale, often assumed to be a characteristic feature of manufacturing industries, do not lead to the creation of diverse cities, but quite the opposite — in a growing world, cities specialize. This result is robust to different mechanisms behind the city size determination, — whether these are free movement or a benevolent local mayor, and to the presence of large service cities in the system. The chapter has recently appeared in *Papers in Regional Science.*

The last chapter (jointly with Sergey Kokovin and Takatoshi Tabuchi) is focused on the properties of monopolistic competition in product characteristics space. To enrich the standard model of monopolistic competition, we introduce a space of product characteristics akin to the early Hotteling (1929) approach. However, unlike Hotteling and the following literature, we maintain a love for variety assumption among consumers, which is central to the modern monopolistic competition literature. This combination of an “ideal type" of product and love for variety in consumption gives rise to partially localized competition: zones of service of different firms intersect. We show that competition intensity, measured as number of firms and the marginal utility of money, increase with growing market size, similar to a standard model of spaceless monopolistic competition. However, the price decreases (increases) if elasticity of utility (instead of demand elasticity in spaceless competition) at the point of consumption of an “ideal variety" is a decreasing (increasing) function. This result stems from the aggregation of heterogeneous consumers into a demand function faced by a firm — aggregate demand does not inherit all the properties of individual demands. In addition, we show that with increasing market size, competition becomes more localized: the segment of space served by a firm decreases as the market expands.

Chapter 1

Sorting When Firms Have Size

1.1 Introduction

The increasing availability of detailed micro level data sets has made us well informed about the large extent of heterogeneity on both sides of the labor market. Recent research has shown that firms differ by size, productivity, capital intensity and many other characteristics. More importantly, the differences are enormous even within narrowly defined industries (Crozet and Trionfetti, 2013); or in other words, a large proportion of these differences can hardly be attributed to any observable characteristic of firms. This suggests that unobserved firm characteristics play an important role in any explanation of firm behavior and outcomes. A similar observation holds for workers. Studies of wage determination and wage inequality have revealed that wage inequality has been growing in recent decades in virtually all countries, and that most of the inequality and its growth cannot be attributed to observable characteristics of workers, even in narrowly defined occupation-sector cells (see, for example, Helpman, Itskhoki, Muendler and Redding, 2012). Therefore, a study of the labor market cannot disregard the heterogeneity in, and interplay between, the unobserved characteristics of both firms and workers.

How does this enormous heterogeneity in the labor market play out in the interaction between its two sides — firms and workers? Are there strong complementarities in production between characteristics of workers and firms? How smooth or frictional is the process of reallocation? Does the market allocate workers to employers in optimal fashion? If it does not, what is the role of different sources of misallocation, e.g. search frictions and market power, and how far from optimal output is the departure? The answer to these and related questions is important for the resolution of numerous economic debates. Study of misallocation at the micro level is important for the macroeconomics literature, as it has a direct impact on aggregate productivity fluctuations and long-run income dynamics. One manifestation of this question, especially relevant and occupying business cycle literature nowadays, is whether the slow recovery after the Great Recession is due to a mismatch between workers and firms.

Understanding the allocation of workers to jobs is particularly vital to international trade. There, researchers are interested in whether exporters pay higher wages due to their own higher productivity or due to hiring better workers. An answer to this question is essential for understanding the implications of trade liberalization for the distribution of wages.

In order to provide a partial answer to these questions, I develop a model that features heterogeneous price-making monopolistically competitive firms, heterogeneous workers and a frictional labor market. The model addresses the implications of the creation of vacancies within firms and choice of size for labor market sorting outcomes. In a nutshell, the model introduces the random search model developed by Shimer and Smith (2000) into the monopolistic competition framework à la Melitz (2003). Workers searching for jobs are randomly matched with vacancies posted by firms. There is no free entry of firms, in contrast to the most standard search models, however, each firm can post as many vacancies as it needs. In other words, there is free entry of vacancies within a firm. Production is linear in the *quantity* of labor, i.e. firm output is a sum of outputs of its workers. Nevertheless, even the most productive firms do not grow indefinitely due to decreasing demand for their products and the local market power they enjoy.

The contribution of the paper is twofold. First, I show that match output, and hence a firm's production function, are identified non-parametrically. The identification strategy I develop utilizes firm level data. One of its advantages is that identification of firms' unobserved characteristics is achieved independently from workers' characteristics in a straightforward and intuitive way. In other words, ranking of firms is identified only from firms' variables, such as revenue and vacancies, and does not depend on the wages they pay. Identification of the match output function is essential for any counterfactual analysis that includes shifting different workers between different jobs. In that context, non-parametric identification is especially important because of the lack of microfoundations behind the structure of the match production function. All the more so, the arguments of the match production function, i.e. workers' and firms' unobserved heterogeneity drivers, are not well understood themselves.

In addition, I show that along with the match output, vacancy creation costs are identified non-parametrically. Understanding the shape of vacancy creation cost is important because it has potential implications for business cycle models: The faster the marginal cost is increasing with vacancy creation, the more incentives a firm has to smooth the hiring process. Thus, convex vacancy creation cost can be one explanation for slower recoveries.

Second, I extend the model into an international trade setting to show how the explicit incorporation of firm size can dramatically change predictions about sorting. Exporters allow for a wider range of quality among their workers and pay them higher wages than non-exporters. This is in dramatic contrast to the result of Bombardini, Orefice and Tito (2014), who show that in a one-to-one matching framework, exporters tend to choose, on average, better workers and have less skill dispersion in their workforce. The different result is driven by the firm size effect. In the present model, exporters are larger, and the cost of supporting larger firm size require larger equilibrium match surplus. Therefore, the set of acceptable workers, i.e. those with whom the match surplus is positive, expands. In addition, exporters pay a higher wage to their workers because the wage is positively related to the match surplus.

Understanding predictions about matching sets of exporters relative to non-exporters is important because it has implications for the effects of trade liberalization on wage inequality. Indeed if, on average, matching sets become tighter when more firms engage in exporting activity, wage inequality will also increase, for two reasons. First, the share of workers who can enjoy the exporter wage premium increases more slowly than the share of exporting firms. Second, the tighter matching sets indicate that matches are closer to perfect and workers' wages are closer to their maximal attainable wages (given the aggregate environment). Conversely, expanding matching sets downward pressure on wage inequality.

This paper fits into the extensive line of research on the estimation of models of sorting on the labor market based on unobserved characteristics of workers and firms, firmly grounded in theory, such as Lopes de Melo (2013) and Lise, Meghir and Robin (2013). However, the paper addresses two major shortcomings of the current literature. First, a large part of the literature imposes a good deal of structure on the model and, in particular, does not allow for varying degrees of complementarity — Hagedorn, Law and Manovskii (2014) being a notable exception. Second, virtually all research in the area so far has not made any distinction between 'a firm' and 'a job'. In other words, in these models, firm boundaries are arbitrary, and a firm with n workers is equivalent to n firms with one worker each. Although theoretically convenient, this approach has a major shortcoming when applied to data. Since no data set has information on profitability or output of a particular worker or workplace in a firm, the identification of firm characteristics is based on information about wages and labor flows. Therefore, the identification strategies tend to be indirect and computationally intensive.

Additionally, the abstraction from the firm size prevalent in the macroeconomic labor literature inevitably misses a potentially important intensive margin of employment adjustment over the business cycle: expansion (as opposed to entry) of firms during booms and their contraction (as opposed to exit) during recessions. Intuitively, a firm that can adjust its labor force size has additional room to maneuver when faced with shocks, and models taking this into account can produce different amplification mechanisms. Hence, the growing literature on the role of firm size in labor market dynamics in macroeconomics. Kaas and Kircher (2014), Elsby and Michaels (2003) and Moscarini and Postel-Vinay (2014) develop macroeconomic models that capture sluggish labor market dynamics, job flows and evolution of the firm size distribution over the business cycle. However, this paper, to the best of my knowledge, is the first attempt to explicitly take into account the role of the firm size for the outcome of the sorting on the frictional labor market.

Apart from these empirical motivations to consider firm size in the labor market sorting models, there is also a purely theoretical reason that deserves attention. As Bagger and Lentz (2015) note, in a one-to-one matching framework, the decision to accept or reject a match relies heavily on a fundamental scarcity. In such a world, the decision to agree upon a match is equivalent to a decision to discontinue searching. However, the relevance of this assumption is not so obvious, since workers can continue to search for opportunities while employed and firms can have many workers. There is a large literature that relaxes the scarcity assumption on the worker side of the model via on-the-job search. This paper can be viewed as a mirror image of that literature. Although retaining the scarcity on the worker side, I relax it on the firm side of the model via explicit introduction of the firm's choice of size.

1.2 Literature Review

Since the seminal study by Abowd, Kramarz and Margolis (1999), it has been believed that one can grasp unobserved characteristics by first running Mincerian regressions of wages on observable characteristics of firms and workers and their respective fixed effects using longitudinal linked employer-employee data sets. Second, examining these fixed effects in particular, correlation between them conditional on being matched has been considered a rough measure of sorting. With the increasing availability of linked employer-employee data sets, this approach has been widely adopted and applied to data sets from a number of countries, with a general conclusion that the correlation coefficient between fixed effects in worker-firm matches is not very large. Moreover, most studies have found it to be either insignificant or even negative[1]. Although in their review of early literature, Abowd and Kramarz (1999) cautioned that "it is important to keep in mind that it is not always possible to make a direct interpretation of the statistical parameters (for individuals or firms) in terms of simple economic model" (p. 2671), the lack of a significant positive correlation between worker and firm fixed effects has been widely interpreted as absence or unimportance of sorting.

Recent research has shown that the identifying assumptions of this reduced form approach are inconsistent with virtually every equilibrium model of sorting, and that the estimated fixed effects do not contain information on underlying unobserved characteristics. In other words, applied to data generated by an equilibrium sorting model, this regression approach yields fixed effect estimates that have no interpretation within the original model. The intuition behind this, uncovered by Eeckhout and Kircher (2011), is that wages are potentially non-monotone in a firm's type: a better firm has to be compensated for hiring workers who are worse than desired by the firm and therefore, a linear model is fundamentally misspecified. In addition to the purely theoretical argument against interpreting the absence of correlation between workers' and firms' fixed effects as the absence of sorting, Lopes de Melo (2013) has shown that the correlation between the fixed effects of a worker and her coworkers is strong, suggesting that similar workers do indeed sort together.

Understanding the limitations of the reduced form approach of two way fixed effects regressions has led to the development of literature on the identification of sorting

[1] See Abowd and Kramarz (2009) for the review of early literature, Gruetter and Lalive (2009) for Austria, Abowd, Kramarz, Pérez-Duarte and Schmutte (2009) for the U.S., and Card, Heining and Kline (2013) for Germany among others.

grounded in theory. The starting point for understanding the assortative matching is Becker's (1973) assignment model with transferable utility. The main insight from this model is the crucial dependence of the sign of sorting on the complementarities between two sides of the market: positive assortative matching (PAM) — mating of likes — arises when the production function is supermodular, i.e. the marginal product of an agent in a match increases with the quality of her partner. Shimer and Smith (2000), Atakan (2006) and Eekhout and Kircher (2010) build on Becker's insight and develop the assignment model to introduce search frictions. They show that in the presence of search frictions the interplay between the degree of complementarities and the level of search friction is decisive for determination of the degree and sign of sorting. The Shimer and Smith (2000) model, being the most natural approach, has become a cornerstone of the literature on sorting in the labor market. However, one of the limitations of the theoretical literature on sorting is its focus on one-to-one matching, and its resulting disregard of the role of firm size. This paper aims to overcome this limitation, introducing firm size into what is essentially the Shimer and Smith (2000) framework.

Lopes de Melo (2013) and Lise, Meghir and Robin (2013) develop structural models of sorting and wage dynamics. Estimation of these models suggests that PAM between workers and firms is present in the data. However, the main limitation of their approach is the strong assumptions imposed on the functional form of the production function, that do not allow the sign and strength of sorting to differ along the domain of worker and firm types.

Hagedorn, Law and Manovskii (2014) take a step further. Building on Shimer and Smith (2000), they develop an identification procedure that allows for non-parametric identification of the production function. Applying their framework to German linked employer-employee data, they show that although complementarities between worker and firm productivity (and hence, PAM) prevail on average, there are regions of local substitutability, and that the market exploits this feature of the production function: reassigning workers to firms in perfectly assortative fashion would reduce total output by 1.43%. However, the empirical literature has inherited the limitation of its theoretical predecessor, namely, the focus on one-to-one matching. The identification procedure I develop in this paper borrows heavily from Hagedorn et al. (2014), but overcomes the limitations of one-to-one matching.

Few papers have studied sorting on the labor market in one-to-many matching framework. Eeckhout and Kircher (2012) expand a frictionless Beckerian approach and show that if firms can choose not only the type (quality), but also quantity of production factors, necessary conditions on the primitives for PAM become stricter. Intuitively, switching to worse workers is not as detrimental for a firm as in a one-to-one matching world, since a lower quality can be compensated by a larger number of workers. Therefore, for this not to happen, the loss in match value must be very high in the best firms, i.e. marginal match output should change very sharply with the type of firm. The interesting question of how the presence of the search frictions affects conditions for PAM is outside the scope of this paper. Instead, I take an agnostic stand on the strength of complementarities and a data driven approach: Given the observed labor market outcomes, I uncover the shape of the primitives.

The only attempt to utilize firm level data in the empirical study of sorting I am aware of is the study by Bartolucci, Devicienti and Monzon (2015). They use a number of definitions of the firm's profit to rank firms, and exploit the patterns of movement of workers between firms to deduce the aggregate measures of the degree and sign of the sorting on the labor market. Their methodology remains valid in the model I develop here. Therefore, this paper can be viewed as providing theoretical support, in terms of a general equilibrium model, to their empirical procedure.

1.3 Model

The economy consists of two sectors: one producing differentiated intermediate inputs using labor, and the other assembling the final good from intermediate inputs. The final good is produced under perfect competition. The intermediate good sector is the crucial building block of the model. Its structure integrates Shimer and Smith's (2000) model of a time consuming job search and Melitz's (2003) approach to firm heterogeneity.

Firms in the intermediate sector require labor for production. Both firms and workers are heterogeneous, yet the production function is linear in the *quantity* of labor. However, the market for intermediate inputs is monopolistically competitive, and local market power constrains optimal firm size. Unemployed workers search for jobs in the intermediate sector and firms post vacancies to hire labor. The labor market is frictional: it takes time to fill a vacancy and to find a job.

1.3.1 Final Good Production

The final or consumption good is assembled from varieties of differentiated inputs under Constant Elasticity of Substitution (CES) technology:

$$Y = \left[\int\limits_{j \in \Omega} q(j)^{\frac{\sigma-1}{\sigma}} dj \right]^{\frac{\sigma}{\sigma-1}}, \qquad \sigma > 1, \tag{1.1}$$

where Ω denotes the set of varieties available for production of the final good, and σ is the elasticity of substitution between varieties. In what follows I assume that there is a measure one continuum of firms, each producing distinct variety j. Thus, Ω is a set of measure one. Given the constant returns to scale technology and perfect competition in the final good sector, I can focus on a representative firm and its demand for inputs. The price of the final good is normalized to one.

The cost minimization problem in the final good sector is standard. It implies the following inverse demand for intermediate inputs:

$$p_\varphi = Y^{1/\sigma} q_\varphi^{-1/\sigma} \tag{1.2}$$

that is taken as given by the producers of the intermediate inputs.

1.3.2 Differentiated Sector

Diverging from Melitz (2003), I assume that the mass of varieties of intermediate inputs is fixed[2]. Firms are heterogeneous, they differ in their type φ which is uniformly distributed on $[0, 1]$ interval. Production of a good requires labor, which is also differentiated. I assume that the economy is populated by a unit mass of workers of type a which is uniformly distributed on $[0, 1]$ interval[3]. I assume the following production technology: given any cumulative labor force distribution $L_\varphi(a)$ within a firm of type φ, the output is

$$q(L_\varphi(a), \varphi) = \int \psi(a, \varphi) dL_\varphi(a). \tag{1.3}$$

[2] One can introduce an entry game similar to the original Melitz model. However, since the main focus of the paper is on the labor market, I leave the entry stage out. Conversely, one can always find the fixed production cost and entry cost levels, such that the measure of stayers is one.

[3] Observe that uniform distribution of types is not a restriction. Any other continuous distribution can be transformed to uniform by effectively "renaming" workers aided by corresponding changes in the match production function. In other words, types are ordinal; they do not measure productivity or skills per se, but only by their effectiveness in production.

Here $\psi(a, \varphi)$ can be understood as the efficiency units of labor worker a provides to a firm φ, or as the standard match output of a firm-worker pair. Then, the aggregate firm output is a sum of the output of all individual pairs. Although this assumption disregards potential complementarities or spillovers between different workers[4], it is in line with most of the current literature that treats aggregate output as a sum of match outputs. I apply similar logic to the within firm production, to facilitate comparison with existing models of frictional labor market sorting, and to highlight the role of firm size not confounded with intrafirm technological spillovers.

I assume that the match output $\psi(a, \varphi)$ is an increasing function, with the underlying structural assumption that ordering of types a and φ is meaningful — a higher level implies a more productive type — and global — being more productive does not depend on the match partner. In other words, this is a model of absolute advantage in the labor market. Although restrictive, the last assumption is prevalent in the matching literature. Importantly, I do not put any restrictions on the cross derivative of the match output function, since the main focus of the paper is on its identification.

Intermediate good producers take the demand (1.2) for their goods as given. Thus, the revenue of a firm producing differentiated variety as a function of the production volume is given by

$$R(q_\varphi) = Y^{1/\sigma} q_\varphi^{\frac{\sigma-1}{\sigma}} \tag{1.4}$$

Observe that revenue is a concave function of firm output. This feature, stemming from the demand structure, limits firm size in this model. Alternatively, one can say that a firm faces decreasing monetary returns to its production scale, and therefore one can easily construct an isomorphic version of the model — with production function concave in total effective labor, and perfect competition between the firms in the intermediate sector.

1.3.3 Labor Market

Now, I turn to the core of the paper — labor market structure. Firms' and workers' behavior on the labor market is crucial in the identification of the match output function.

[4] There are a few exceptions addressing intrafirm worker productivity interdependence. Bombardini, Gallipoli, and Pupato (2012) study intrafirm complementarities level as a source of industry comparative advantage. Helpman, Itskhoki, and Redding (2012) introduce congestion into production technology. However, since production function generally depends on the whole labor distribution within a firm, every tractable approach to it is bound to impose restrictive structural assumptions.

Furthermore, the labor market is frictional, and frictions are the only source of movement of workers between different firms, and therefore — the source of identification.

Time is discreet. In every period workers can be either employed or unemployed. Employed workers receive wage income, and unemployed workers enjoy utility equivalent to flow income $b(a)$. Firms post vacancies, and unemployed workers search for a job. The chances of finding a job and filling a vacancy are governed by the labor market tightness θ, which is defined as the vacancy-to-unemployment ratio. The meeting rate is given by $m(\theta)$ for a firm and $\theta m(\theta)$ for a worker, the latter representing the matching function. This indirectly implies a standard assumption of a constant returns to scale matching function. I additionally assume that $m(\theta)$ is decreasing in θ and $\theta m(\theta)$ is increasing in θ, which is equivalent to the assumption that the number of matches increases both with the number of vacancies and with the number of unemployed workers.

The meeting is random: neither firms nor workers can target a potential partner's type. The match is consummated voluntarily upon a meeting, when types of both partners are perfectly observable. There is no on-the-job search in the baseline model, and a worker stays in the match until its separation. The matches are dissolved exogenously with probability δ.

Households are assumed to be risk-neutral suppliers of labor of a given skill a. They maximize the expected lifetime income flow, discounted at an interest rate r. Denote $U(a)$ the value function of unemployed worker of type a, $V(a;\varphi)$ the value function of a worker a employed at firm φ. I impose symmetry across firms of a given type, which allows me to ignore the potential dependence of value functions and wages on firm employment. The value functions of a worker obey the following two Bellman equations[5]:

$$rU(a) = b(a) + \theta m(\theta) \int \gamma(\varphi) \max\{V(a;\varphi) - U(a), 0\} d\varphi, \qquad \forall a \tag{1.5}$$

$$rV(a;\varphi) = w(a,\varphi) + \delta(U(a) - V(a;\varphi)) \qquad \forall (a,\varphi) \tag{1.6}$$

Here $\gamma(\varphi) = \frac{v(\varphi)}{\int v(\varphi) d\varphi}$ is the distribution of vacancies across firm types. It governs the chances of meeting a firm of any given type. It is straightforward that a worker engages in a match if the value of being employed exceeds the value of being unemployed. The

[5] Due to monopolistic competition in the intermediate goods sector, firms there obtain positive profits that workers can potentially have claims on. However, unless these claims depend on the state of employment, they do not affect employment decisions. For that reason, I omit them from Bellman equations for workers' monetary flow to ease notation, effectively assuming that a separate class of entrepreneurs enjoys all profits from the intermediate goods sector.

interpretation of these equations is rather standard. Equation (1.5) states that the flow value of unemployment consists of income in unemployment and expected gain in value from meeting a firm. Equation (1.6) represents the flow value of employment at firm φ as wage $w(a, \varphi)$ at this firm and potential loss in value from separation. Generally wages depend not only on the type of worker and firm, but also on the composition of the labor force within a particular firm. I ignore this dependence because I focus on the symmetric steady state equilibrium, in which all firms of the same type have the same labor force structure.

Thus, the behavior of workers is straightforward: They look for a job and take any that brings higher income flow. As I show later, this is equivalent to a simple reservation wage rule. Denote $\mathcal{A}(a)$ the set of acceptable matches for a worker of type a, i.e. $\mathcal{A}(a)$ is a subset of firm types a worker a is willing to work for given the equilibrium wage:

$$\mathcal{A}(a) = \{\varphi : V(a, \varphi) - U(a) > 0\}.$$

Now I turn to firms' behavior on the labor market. Denote $J(L_\varphi, \varphi)$ a value of firm of type φ and labor force $L_\varphi(a)$. Firms maximize their present value, which is equal to the discounted stream of profits. In order to hire workers, firms choose a measure of vacancies v to post. The Bellman equation for the firm problem is:

$$J(L_\varphi, \varphi) = \max_{v, L'} \frac{1}{1+r} \left\{ R(q(L_\varphi, \varphi)) - \int w(a, \varphi) dL_\varphi(a) - c(v) + J(L', \varphi) \right\} \tag{1.7}$$

subject to the definition of revenue (1.4) and the law of motion of within firm employment

$$0 \leq L'(a) \leq (1-\delta)L(a) + m(\theta)v \int_0^a \frac{u(a')}{u} \mathcal{I}\{\varphi \in \mathcal{A}(a')\} da', \tag{1.8}$$

where $u(a)$ is the measure of unemployed workers of particular type a, and $u = \int_0^1 u(a')da'$ is total unemployment. $c(v)$ is the vacancy posting flow cost with $c'(v) > 0$, $c''(v) \geq 0$. The change in within firm employment consists of labor attrition due to separation, and new hires obtained from filled vacancies. $\mathcal{I}\{\cdot\}$ is the indicator function, and the term $\mathcal{I}\{\varphi \in \mathcal{A}(a')\}$ ensures that a worker takes a job if offered. The inequality in the labor law of motion implies that a firm can shadow any of its labor force without cost if necessary, creating an asymmetry in labor adjustment cost. Although artificial, this assumption is standard in the search literature. It is not crucial for my results, since in a steady state there is no voluntary match destruction.

Lastly, because hires are not interchangeable with workers outside the firm, workers have bargaining power. I assume that wages are determined through the generalized Nash bargaining solution in the spirit of Stole and Zwiebel (1996), thus

$$(1-\beta)(V(a;\varphi)-U(a)) = \beta\frac{dJ(L,\varphi)}{dl(a)} \tag{1.9}$$

where β is the bargaining power of a worker. Put briefly, worker and firm receive fixed shares of the match surplus, with β being the share of the worker. Due to the risk-neutrality on both sides, the total match surplus can be viewed as a monetary gain of size $S = \frac{dJ(L,\varphi)}{dl(a)} + V(a;\varphi) - U(a)$. Here the meaning of derivative $\frac{dJ(L,\varphi)}{dl(a)}$ is the marginal benefit for a firm from hiring a worker of a particular type. Effectively, the firm considers an increase in its value from a match with a worker, relative to the non-consummation of this particular match. A worker's gain from the match $V(a;\varphi) - U(a)$ is the share β (his bargaining power) of the match surplus. This is the intuition behind (1.9).

Note that in general the solution to the bargaining problem would depend not only on the firm's and worker's types, but also on the whole distribution of employment in the firm, and, in particular, on the firm's size. However, since I focus on the steady state of the economy, firm type φ captures all the latter, and therefore, the equilibrium wage depends only on the pair of types.

Finally, labor balance identities should hold, i.e. the sum of employed and unemployed workers should be equal to the total population:

$$M\int L_{\varphi}(a)d\varphi + \int_{0}^{a} u(a')da' = a. \tag{1.10}$$

Recall that the distribution of workers by type in the population is uniform, and therefore the mass of people with a type weakly below a is equal to a.

1.4 Labor Market Equilibrium and Identification of Sorting

In this section I provide partial characterization of the firms' optimal behavior. The established properties of the firms' behavior will be central to the development of a strategy for the identification of the model primitives. As is standard for models of frictional sorting, I use a properly defined match surplus function, as a gain from consummating the match relative to the outside option. Then, I describe how the match surplus feeds into wages and hiring decisions. The established interdependence of surplus, wages and vacancy creation allows for the identification of the production function from data on wages and firm revenue.

I focus on the steady state equilibrium of the economy. First, define a surplus function

$$s(a,\varphi) = \frac{\sigma}{\sigma-\beta}\frac{dR(\varphi)}{dl(a)} - rU(a) \tag{1.11}$$

The next proposition establishes that this is a proper definition of the surplus function in the sense that matches are consummated whenever it is positive. This can be understood intuitively: an additional worker brings to the negotiation table the value of his marginal product, $\frac{dR(\varphi)}{dl(a)}$ in flow terms. At the same time, his presence increases output, damping marginal revenue. This in turn leads to a decrease in the value of the marginal product of other workers. Hence, the firm improves its position in bargains with other workers it employs in the period. Thus, the multiplier $\frac{\sigma}{\sigma-\beta} > 1$ takes care of this pecuniary externality in the negotiation process. The smaller β, i.e. weaker bargaining position of workers naturally leads to less importance of this externality. At the limit, if the workers have no bargaining power, the match gain is equal to the marginal revenue. The negative part of the surplus is straightforward: the worker forgoes her value in unemployment and the firm has nothing to lose, since the vacancy cost is sunk at the point of wage negotiation.

Proposition 1. *(i) The hiring decision is governed by the surplus function with matches being consummated whenever* $s(a,\varphi) \geq 0$

(ii) The outcome of wage bargaining yields the following wage rule:

$$w(a;\varphi) = \beta s(a,\varphi) + rU(a) \tag{1.12}$$

(iii) The vacancy creation policy of a firm is indirectly defined by

$$c'(v) = \frac{m(\theta)}{r+\delta} \int (1-\beta) s^{+}(a,\varphi) \frac{u(a)}{u} da \tag{1.13}$$

where $s^{+}(\cdot) = \max\{s(\cdot), 0\}$

Moreover, with this result at hand, the Bellman equation for an unemployed worker (1.5) can be reformulated as follows:

$$rU(a) = b(a) + \beta \frac{\theta m(\theta)}{r+\delta} \int s^{+}(a,\varphi) \gamma(\varphi) d\varphi \tag{1.14}$$

The intuitive interpretation of the results presented in Proposition 1 is clear in light of the surplus function definition. The wage is nothing but a standard Nash-bargaining type surplus sharing rule that assigns β share of the surplus to a worker. Although one should bear in mind that the parties bargain over the surplus over the whole period of the relationship, due to the focus on the steady state, this is equivalent to sharing flow surplus every period.

The last point of Proposition 1 requires that the optimal policy of a firm equates the marginal cost of a vacancy on the left hand side to its expected marginal benefits. Indeed, on the right hand side of the equation (1.13) the chance of meeting a worker, $m(\theta)$, is multiplied by a firm's average share of surplus resulting from a meeting (whether a match follows or not), represented as an integral over worker types with the distribution of unemployed workers being the relevant one. In addition, multiplier $\frac{1}{r+\delta}$ accounts for the total discounted flow over the expected length of the relationship. The equation (1.14) has a similar meaning, only from the worker's viewpoint.

It is worth emphasizing the two assumptions of the model that allow for extension of the job search model from one-to-one to one-to-many matching framework and neat equilibrium characterization. First, as I stressed earlier, there are no direct complementarities between the workers in the production function, and output is the sum of marginal products of the workers. Second, the CES aggregation of the intermediate goods into the final good in (1.1) guarantees that the marginal value-product of a worker is proportional to her marginal product, with the proportionality coefficient depending only on the firm's type. These two assumptions taken together allow for neat characterization of the solution to the firm's problem.

The identification procedure I develop hinges on the established properties of optimal behavior of firms and workers. I now discuss how they allow for identification of worker ranking, value function in unemployment, vacancy creation cost and match output function.

1.4.1 Identifying Worker Types and Unemployment Value

The worker side of the model is similar to the standard one-to-one frictional search setup of Shimer and Smith (2000). Thus, the identification of worker types developed by Hagedorn et al. (2014) carries over to the model presented in this paper. The validity of this identification strategy is warranted by the following proposition:

Proposition 2. *Let the value in unemployment, $b(a)$, be non-decreasing in worker type a. Then,*

- *value in unemployment $U(a)$ is increasing in worker type a, and hence, wage $w(a, \varphi)$ and value in employment $V(a; \varphi)$ are increasing in a;*
- *minimum and maximum wages attainable by a worker of a given type a are increasing in a.*

In addition, if for a worker of type a there is a firm type φ that does not hire her in equilibrium, then the minimal wage she attains is equal to the flow value of unemployment: $\min\limits_{\varphi} w(a, \varphi) = rU(a)$.

The intuition behind Proposition 2 is quite straightforward. To a given firm, a better worker has a larger value of marginal product $\frac{dR(\varphi)}{da} = \frac{\sigma-1}{\sigma} p_\varphi \psi(a, \varphi)$, which is reflected in the wage ranking within a firm. Given the uniformity of ranking within a firm and non-decreasing flow value in unemployment, better workers have better prospects when unemployed. The last part of the Proposition states that, if workers do not accept wage offers from some firms, they follow the simple reservation wage rule in accepting offers with the reservation wage being equal to flow value in unemployment.

The identification of worker ranking from wage data is based on the derived monotonicity properties. These properties also hold in a one-to-one matching model as in Hagedorn et al. (2014) and their identification strategy applies. For the sake of completeness I reiterate their argument in the remainder of this subsection.

If a firm matches with all worker types, the wages it pays would provide a global ranking of the workers. However, this can hardly be expected, especially if the search frictions are sufficiently small and complementarities in the production function are sufficiently strong. Nevertheless, wage ranking within a firm provides a ranking of workers within that firm. With sufficient mobility of workers across firms, one can employ a transitivity argument. Consider a worker a moving from firm φ_1 to firm φ_2. Then, any worker $\tilde{a}$ with a higher wage in firm φ_2, i.e. $w(\tilde{a}, \varphi_2) > w(a, \varphi_2)$, should have a higher rank than any worker $\tilde{\tilde{a}}$ in firm φ_1 with wages $w(\tilde{\tilde{a}}, \varphi_1) < w(a, \varphi_1)$, and vice versa. In other words, if, according to wage ranking within firm φ_2, $\tilde{a} > a$ and, according to wage ranking within firm φ_1, $a > \tilde{\tilde{a}}$, transitivity implies that $\tilde{a} > \tilde{\tilde{a}}$. With sufficient mobility of workers across firms, aggregation of interfirm ranking, aided by this transitivity argument, identifies the global ranking of workers in a linked employer-employee dataset.

Two complications might arise in empirical implementation of this procedure. First, the measurement error in wages can distort the observed workers' ranking within the firm. To address this problem, Hagedorn et al. (2014) augment the aggregation procedure by assigning weights to worker pairs within the firm. The particular structure of weights depends on the distributional assumption about the measurement error. They work with a normal distribution, imposing independence across workers and firms. Under this assumption, weights have an intuitive structure: a higher wage difference in an observed worker pair leads to a higher incremental value in the aggregation objective function (effectively Bayesian probability) if they are ranked according to the wage differential.

Second, the exact aggregation of ranks within a firm is computationally complex. The results of Proposition 2 on global ranking in maximal and minimal attainable wages help to improve the procedure by providing an initial ranking that should be close to the exact ranking (and is not exact only due to the measurement error). Hagedorn et al. (2014) show that one can initialize the algorithm with global ranking by maximal or minimal wage, and use single worker moves for improvement. This procedure yields an accurate solution without being as computationally demanding as the original problem.

The last part of Proposition 2 allows for the identification of the unemployment value as a minimal attainable wage. However, given the relatively short time span of linked employer-employee data sets usually used in a sorting estimation, straightforward empirical implementation could be problematic. Hagerdorn et al. (2014) put forward a solution based on the fact that ranking of workers is identified. By the continuity argument, similarly ranked workers must have similar reservation wages. Thus, one can group together similarly ranked workers and consider them as being of the same type. This approach dramatically expands the number of observations available for a given worker type and yields sufficiently precise estimates of reservation wages and unemployment values.

1.4.2 Identification of Vacancy Posting Cost

Insights from Proposition 1 allow for the identification of the vacancy posting cost function when a researcher has data on the number of vacancies within a firm, in addition to the wage data. Observe that with a Nash bargaining result (1.12), the surplus can be identified from wage data:

$$s(a,\varphi) = [w(a,\varphi) - \min_{\varphi} w(a,\varphi)]/\beta \tag{1.15}$$

In other words, the surplus is proportional to the wage premium over the reservation wage of a worker. Here, I used the fact that flow value in unemployment is identified by minimal attainable wage using the procedure developed in the previous subsection. Now, I rewrite the vacancy creation policy (1.13) in the following way:

$$(r+\delta)c'(v) = m(\theta)\int \mathcal{I}\{s(a,\varphi)>0\}\frac{u(a)}{u}da \times \int \frac{(1-\beta)s^{+}(a,\varphi)\frac{u(a)}{u}}{\int \mathcal{I}\{s(a',\varphi)>0\}\frac{u(a')}{u}da'}da \tag{1.16}$$

with $\mathcal{I}$ being an indicator function. The two terms on the right hand side of (1.16) have direct empirical counterparts. The first term represents the chances of a vacancy meeting a worker multiplied by the share of acceptable workers in the unemployment pool. Together, this constitutes a probability that the vacancy is filled at the end of the period. Thus, the empirical counterpart of the first term is the ratio of the number of new hires to the number of posted vacancies. The second term is the firm's share of the surplus from a match averaged across new matches. With the surplus identified from (1.15), it is proportional to the average wage premium of new hires over their respective reservation wages.

With the marginal posting cost identified, one can test for convexity of the vacancy creation cost, i.e. for increasing marginal cost. The degree of vacancy posting cost convexity has important implications for macroeconomic models. Although the model does not feature business cycle fluctuations, the role of the shape of the vacancy creation cost function in firm dynamics can be understood intuitively. Indeed, if the cost function was found to be convex, it would imply that firms have incentives to smooth vacancy creation over the recoveries, i.e. to distribute vacancy creation over a longer period of time, leading to a slower recovery process. On the other hand, constant marginal cost would imply that firms immediately adjust their labor force to the optimal level.

The identification of the vacancy posting cost function depends crucially on the linearity of the relationship between the surplus and wage premium, which is the result of the particular assumption about the bargaining process. However, in any model where the wage depends positively on the match surplus, the relationship between wage premium and surplus would be monotone. This assumption seems a natural outcome of a wage setting process. Therefore, the proposed procedure for vacancy cost identification is robust to alternative specifications of the bargaining arrangement. Although this identification strategy would not correctly identify the exact functional form of vacancy posting cost, with wages monotone in the match surplus, it identifies a monotone transformation of the marginal cost of vacancy posting. Therefore, the test for constant marginal cost would not be misspecified, and would still discriminate correctly between linear and convex cost functions as long as the vacancy creation technology is the same for all firms.

1.4.3 Identifying Firm Types and Production Function

The main focus of the literature is on the identification of sorting. In this subsection, I show how the model structure allows one to identify the firm ranking and production function with the help of the firm level data. Since most of the previous literature has equated firms and jobs, the identification procedures developed so far can rarely make use of firm level data. The only exception I am aware of is Bartolucci et al. (2015), who use firm data to recover aggregate measures of strength and sign of sorting on the labor market. The explicit introduction of the firm into the model allows for much simpler identification of the details of sorting outcome from an additional source of information.

I start with the following proposition:

Proposition 3. *In equilibrium, better firms enjoy higher profits, i.e. firm value $J(\varphi)$ is increasing in φ. Additionally,*

(i) If $c''(v) > 0$ then $\varphi_i > \varphi_j$ implies that either $R_i > R_j$, or $v_i > v_j$, or both.

(ii) If the marginal vacancy posting cost is constant, $c'(v) = c$, $\varphi_i > \varphi_j$ implies $R_i > R_j$.

Turned on its head, the proposition implies that the profit ordering pins down type ordering. Yet due to potential measurement error in or misreporting of profits, the ranking of firm types from profits alone might not be measured efficiently. The second part of Proposition 3 provides an additional source of identification that might be useful in practical applications. It states that ordering of revenues and vacancies identify firms' ranking as well as profits. The more aligned profits, revenue and vacancies rankings are, the easier it is for the model to reconstruct firms' ranking confidently.

One of the advantages of this identification strategy lies in the fact that two rankings are identified using different sources of information: workers' ranking is identified from individual wage data, whereas firms' ranking is from firm level data. In addition to making the identification of the firms' ranking more straightforward and intuitive, relative to the current literature, it evades potential biases in the sorting estimation, stemming from the fact that firm type is identified from wages, and hence the types of its workers.

If the correlation between revenue and vacancies is not perfect, but sufficiently high, the researcher can use the analogous tactics that were applied in the identification of flow unemployment value. Firms with similar revenue and vacancies, yet an opposing ranking of the two, can be grouped together as firms of the same type.

The last step is the identification of the production function. Observe that, using surplus definition (1.11), we can find the match production function as

$$\psi(a,\varphi) = \frac{\sigma-\beta}{\sigma-1} Y^{\frac{1}{\sigma-1}} R_{\varphi}^{\frac{1}{\sigma-1}} \left[s(a,\varphi) + rU(a) \right].$$

From this, together with the identification results developed above, it follows that the production function can be identified, up to a constant multiplier, from wages and revenues, by the following equation:

$$\tilde{\psi}(a,\varphi) = R_{\varphi}^{1/(\sigma-1)} \left[w(a,\varphi) - (1-\beta) \min_{\varphi} w(a,\varphi) \right] \tag{1.17}$$

Note that the production function is identified non-parametrically. Thus, it allows for flexibility in the sign and degree of complementarity on the domain of function. With the production function identified, one can investigate the degree of complementarities locally and globally. In addition, one can ask how much of total output can be gained by worker reallocation between firms or jobs reallocation (changes in firm sizes), i.e., how detrimental search frictions are.

The developed procedure for identification of the production function relied on knowledge of the elasticity of the substitution parameter σ. An alternative approach would be to augment the wage equation and to use a linear-regression technique to estimate the augmented version of it. This would allow a researcher to identify the elasticity of substitution simultaneously with the production function. Therefore, it would provide, in addition, an indirect check of the model's validity: the estimated value of the elasticity of substitution should lie in the region agreeable with the literature[6].

The idea behind the alternative production function identification is somewhat straightforward. Rather than invert the wage equation (1.12) for the match output, one can write it as follows

$$\ln(w(a,\varphi) - (1-\beta)\min_{\varphi} w(a,\varphi)) = \chi + \frac{1}{\sigma - 1}\ln R_{\varphi} + \ln \psi(a,\varphi) \tag{1.18}$$

This is a standard log-linear equation that can be estimated with ordinary least squares methodology. The disadvantage of this identification technique is that it requires variation in the firm revenue on the level of type. In the empirical implementation any data set provides two sources of such variation. Firstly, recall that for identification of types we grouped similar workers and similar firms together. Consider a worker-firm pair (i,j) and note by $a(i)$ and $\varphi(j)$ respectively the worker and the firm type assigned to them during the identification. Then the wage equation for econometric estimation can be written as

$$\ln(w_{ijt} - (1-\beta)\min_{\varphi} w(a(i),\varphi)) = \chi + \frac{1}{\sigma - 1}\ln R_{jt} + \ln \psi(a(i),\varphi(j)) + \varepsilon_{ijt}, \tag{1.19}$$

where ε_{ij} keeps track of measurement error in wages. Thus, grouping similar firms would allow identification of the firm production function and the elasticity of substitution. However, the variation in revenue at the firm level must be small by construction, leading to very imprecise estimates

Arguably more importantly, there is inevitable time variation in firm revenue stemming from the business cycle. Although this sort of fluctuation is likely to be the main source of identification of σ in practice, the model so far does not account for productivity fluctuation. More work is needed to understand to what extent aggregate productivity movement would alter the identification procedure developed. However, if they are small relative to the labor market adjustment velocity, one can conclude that business cycle

[6] However, the power of this test is rather low, since consensus on the acceptable values of the elasticity of substitution has not yet been achieved. A survey by Hillbery and Hummels (2013) reports elasticity values in the range from 0.9 to 34.4.

fluctuations should not distorte the identification too much.

Observe that the structure of the model suggests a specification of the wage equation similar to that of Hagedorn et al. (2014), but different from the one usually considered in the literature. Rather than decomposing log-wages into two-way fixed effects components, it suggests looking at the firm specific component of the wage premium over worker's reservation wage, and not of the wage itself. This is an implication of strategic wage bargaining at the firm level

1.4.4 Discussion of the Size Effect

What difference does the firm size make to the production function identification? There are two channels through which it plays its role. First, the interaction between the market power and the firm size effectively creates disparity between match output and its marginal value. The same effect would be experienced in the presence of concave production function, which would lead to disparity between marginal and average products. Second, vacancy creation within the firm might lead to a different marginal posting cost between firms. Indeed, if the vacancy creation cost function is convex, different firm size unequivocally leads in equilibrium to different marginal vacancy cost for different firms. However, most of the models of one-to-one matching assume a constant entry cost for vacancies, independent of their type. Thus, this unaccounted variation in vacancy cost can become a cause of misidentification of the production function.

To illustrate how these effects may play out, consider the identification of production function in Hagedorn et al. (2014). Their model is particularly close to the one developed in this paper, lacking only the firm size component. The separate existence of vacancies outside firms creates a value to an unfilled vacancy, and the outcome of the wage bargaining accounts for this value. Hence, the production function is identified from the following equilibrium condition:

$$f(a,\varphi) = \frac{w(a,\varphi) - \beta r V_v(\varphi) - (1-\beta) r U(a)}{\beta} \tag{1.20}$$

with $V_v(\varphi)$ being the value of unfilled vacancy of type φ. In contrast to free entry of vacancies into the economy, voluntary creation of vacancies within firms leads to the value of a vacancy being equal to its marginal cost within a firm, but not on the aggregate level.

Therefore, if the data generating process is described by the model developed here, the

Hagedorn et al. (2014) identification procedure will identify the following transformation of the production function:

$$\tilde{f}(a,\varphi) = \xi R_{\varphi}^{-\frac{1}{\sigma-1}}\psi(a,\varphi) + (r+s)c'(v_{\varphi}) \tag{1.21}$$

with ξ being a multiplier reflecting the size of the economy.

The second summand in (1.21) accounts for the unaccounted variation in marginal vacancy cost across firms described above. The intuition behind its appearance is the following. With the match surplus identification coming from the worker side of the model, and therefore being unaltered by the presence of the outside option for the firms, the value of the vacancy shifts the identified match output up. Although the vacancy cost does not enter the bargaining procedure in my model, it would be premature to claim that vacancy creation costs do not play a role in wage bargaining, since this result depends on the intricacies of the nature and timing of the vacancy cost, as well as the bargaining protocol. However, the importance of this effect should not be overstated. Since this effect does not influence the cross derivative of the production function, it is irrelevant to analysis of complementarities, and to the results of counterfactuals that do not substantially change the distribution of the firm sizes. Even in the latter case, for this effect to be important, the vacancy posting cost function should bear a high degree of convexity.

The firm size effect comes from a multiplier $R_{\varphi}^{-\frac{1}{\sigma-1}}$ in (1.21), and exactly accounts for the disparity between match output and its value (or between marginal and average products in the alternative specification). In other words, the one-to-one matching model equates the marginal product and its value, however, it identifies the latter. Importantly, this disparity affects the cross derivative of the production function. One would expect that the multiplier is decreasing in the firm size, therefore, conclusions from the models that do not take it into account might understate the degree of complementarities between workers' skills and firms' productivity in the economy. This effect might be an important driver behind the modest gains from re-sorting workers found in the literature so far.

Importantly, the one-to-one matching model, and hence identification procedure, can be considered a limiting case of the model and identification procedure I develop in this paper. In particular, when $\sigma \to \infty$, aggregation of the intermediate goods into the final goods (1.1) becomes linear, i.e. intermediate sector firms produce perfect substitutes, and firm boundaries effectively disappear. In particular, as can be seen from (1.21), both identification strategies identify the same production function (up to an additive shifter). In this empirical sense, the model I develop can replicate a one-to-one matching model.

1.5 Extension: Exporter Wage Premium

In this section I show how the model can be useful above and beyond addressing the identification of sorting. In particular, I develop an extended version of the model that is relevant to international trade.

Since the seminal work by Melitz (2003), firms' heterogeneity and firm size distribution has become an important explanatory aspect of new trade models and applications. However, the workers' heterogeneity is rarely addressed in these models. Grossman, Helpman and Kircher (2013) study sorting of workers into firms in a different framework. They employ the Heckscher-Ohlin trade model, and focus mostly on sorting between, rather than within, industries. In addition, search frictions in their model do not alter the sorting pattern. The closest work in spirit to mine is the model of Helpman et al. (2010). Their model generates positive assortative matching in a similar setup, due to the functional forms they utilize. Furthermore, in their model, better workers are paid more only because they are employed by more productive firms, i.e. personal productivity affects the wage only through the chance of being hired by a better firm, and the exporters pay higher wages solely due to their higher productivity.

My model easily allows for an extension into international trade because it combines a workhorse model of the matching of heterogeneous types with the standard model of heterogeneous firm sizes. Therefore, it is natural to think about such an extension and the effect international trade has on wages in this framework. As I show later, there is room for an exporting wage premium even if the exporters do not differ from non-exporters in their own productivity (type).

I briefly outline the extended model here, relegating a more detailed description to the Appendix. Consider a world of two symmetric countries, with economies consisting of final and intermediate goods sectors, as described in Section 2. The intermediate inputs can now be traded across the border with impediments a la Melitz (2003). The exports involve fixed cost of numeraire f_x, which is idiosyncratic to a firm, and iceberg cost τ common to all firms, i.e. to ship one unit of the good into the foreign country, τ units of it must be shipped out of the country of origin. Under these assumptions about trade cost, symmetric countries and introduced demand structure, the revenue of a firm becomes:

$$R_\varphi = [Y(1+I\tau^{1-\sigma})]^{1/\sigma} q_\varphi^{\frac{\sigma-1}{\sigma}} \tag{1.22}$$

Here I is an indicator of exporting activity, i.e. $I = 1$ if a firm decides to export and $I = 0$ otherwise. Although in this new environment, the distribution of unemployment and firm sizes would be different from that in the closed economy, Proposition 1 continues to hold. However, though it is hard to track the general equilibrium effects of trade opening on unemployment and production, one could make an interfirm comparison in an open economy. Due to the idiosyncratic cost of exporting, there might be two firms of the same type (production function) one of which is exporting while the other is not. The following proposition summarizes the differences between such two hypothetical firms.

Proposition 4. *Consider two firms, i and j, such that φ_i=φ_j. Assume that in equilibrium firm j exports and firm i does not, i.e. $I(\varphi_j) = 1$ and $I(\varphi_i) = 0$. Then*

(i) $q_j > q_i$ and $R_j > R_i$, i.e. the exporting firm is larger measured by output and revenue;

(ii) if type a of workers is hired by firm i, it is hired by firm j, i.e. firm j has a weakly larger matching set

In addition, if the vacancy posting cost function is convex, $c''(\cdot) > 0$, then

(iii) $v_j > v_i$, i.e. exporting firm posts more vacancies;

(iv) $w_j(a, \varphi) > w_i(a, \varphi)$, i.e. the exporting firm pays a higher wage to any given worker type

Parts (i) and (iii) of the proposition are rather standard for the trade literature. Since the seminal work by Bernard and Jensen (1999), it has been confirmed empirically and theoretically that exporting firms are larger than non-exporting ones. The result (ii) is less straightforward, and deserves special explanation. Start with a hypothetical situation in which these two firms have the same output. Due to the availability of a foreign market, the conversion of output into revenue is higher for the exporting firm. i.e. its total and marginal revenue are higher. This drives up the surplus from a match with any given worker, and the exporting firm has incentives to expand, both in terms of vacancy creation and type acceptance. Expansion puts downward pressure on the surplus, but does it in a uniform fashion across workers. However, since a larger firm has to create more vacancies, in equilibrium the average surplus from its new hires has to be larger as well. Together with the fact that a firm cannot change the surplus from matches with different workers differently, the surplus from any given match should be higher, implying a larger matching set for an exporting firm. From this immediately follows the last part of the proposition, as wages are tightly connected to the match surplus. Thus, the model highlights the different foundation for an exporter wage premium: the cost of the supporting firm size. In this model, large firms have to create more vacancies at higher marginal cost. The results of the Proposition, especially of the second part of it, rely heavily on this assumption. However, as noted in the previous subsection of the paper, this particular assumption can be tested in the future.

The result on comparison of the matching sets is drastically different from that of the model with one-to-one matching in Bombardini et al. (2014). In a world where a firm can match with one worker only, exporting increases the importance of a right match, shifting up and narrowing down the acceptance set of the exporting firm relative to a non-exporting firm. This difference highlights how explicit incorporation of firm size into the search models of labor markets can substantially change the predictions of the models.

1.6 Conclusion

This paper develops an equilibrium model of matching between workers and firms where firms, as opposed to jobs, have size. In other words, firms make decisions not only about the extensive margin — what types of workers to hire — but also about the intensive margin — how many workers to hire. I also show theoretically how equilibrium conditions resulting from the optimizing behavior of workers and firms allow for identification of the model primitives such as match output and vacancy creation cost functions. I show that for the identification, one needs data on workers' wages and firms' revenues and vacancies, which are usually observable in the modern linked employer-employee data sets. Importantly, identification of the production function, which is the cornerstone in addressing the question of sorting, is performed non-parametrically.

The proposed identification procedure permits the quantification of the role of search friction in its interplay with the complementarities in production and with firm size. In order to quantify the role of frictions on the extensive margin of hiring, one can compute the change in total output resulting from optimal reassignment of workers to firms, conditional on the observed number of jobs within each firm. Additionally, to assess the role of search frictions with regard to firm size, one can look at the loss in the aggregate output relative to the globally optimal assignment of workers to firms. I believe the empirical quantification of these effects will be an important step in the further advancement of this line of research.

Next, I extend the model to allow for international trade. This exercise shows the importance of consideration of firm size for predictions about equilibrium sorting. In particular, in this model, exporters have larger matching sets than non-exporters; i.e. they hire more types of workers, in contrast to the prediction of the one-to-one matching model of Bombardini et al. (2014). In addition, this formulation sheds new light on the exporter wage premium: I show that the necessity of supporting larger firm size forces exporters to pay higher wages.

This paper is a first step in the study of the role of the firm size in sorting on the labor market. The further advancement of this line of research requires the empirical assessment of the model and quantification of the role of the firm size. However, prior to that, the model should be enriched to include prominent features of the data, such as job-to-job transitions and on-the-job search. First, allowing workers to search for a job while employed would relax the scarcity assumption imposed on the workers' side of the labor market in the same way this paper has relaxed the scarcity assumption on the firms' side. Second, I expect that this extension will dramatically improve the performance of the model when faced with data.

Bibliography

Abowd, J.M., Kramarz, F. (1999) The analysis of labor markets using matched employer-employee data. In Ashenfelter O. and Card D. (Eds.). *Handbook of labor economics 3B* (pp. 2629-2710). Amsterdam: Elsevier North Holland.

Abowd, J.M., Kramarz, F., Margolis D.N. (1999). High wage workers and high wage firms. *Econometrica*, 67(2), 251-333.

Abowd, J.M., Kramarz, F., Pérez-Duarte, S., Schmutte, I. (2009). *A formal test of assortative matching in the labor market.* NBER Working Paper No. 15546

Atakan, A. (2006). Assortative matching with explicit search cost. *Econometrica*, 74(3), 667-680.

Bagger, J., Lentz, R. (2015). *An Empirical Model of Wage Dispersion with Sorting.* Unpublished manuscript.

Bartolucci, C., Devicienti, F., Monzon, I. (2015). *Identifying Sorting in Practice.* Unpublished manuscript.

Becker, G.S. (1973). A theory of marriage: Part I. *Journal of Political Economy*, 81(4), 813-846.

Bernard, A. B., Jensen, J. B. (1999). Exceptional exporter performance: cause, effect, or both?. *Journal of international economics*, 47(1), 1-25.

Bombardini, M., Gallipoli, G., Pupato, G. (2012). Skill Dispersion and Trade Flows. *American Economic Review*, 102(5), 2327-48.

Bombardini, M., Orefice, G., Tito, M. D. (2014). *Does Exporting Improve Matching? Evidence from French Employer-Employee Data.* Unpublished manuscript.

Card, D., Heining, J., Kline, P. (2013). Workplace Heterogeneity and the Rise of West German Wage Inequality. *The Quarterly Journal of Economics*, 128(3), 967-1015.

Crozet, M., Trionfetti, F. (2013). Firm-level comparative advantage. *Journal of International Economics*, 91(2), 321-328.

Eeckhout, J., Kircher, F. (2010). Sorting and Decentralized Price Competition. *Econometrica*, 78(2), 539–574.

Eeckhout, J., Kircher, F. (2011). Identifying sorting—in theory. *Review of Economic Studies*, 78, 872-906.

Eeckhout, J., Kircher, F. (2012). *Assortative Matching with Large Firms: Span of Control over More versus Better Workers.* Unpublished manuscript.

Elsby, M. W., Michaels, R. (2013). Marginal jobs, heterogeneous firms, and unemployment flows. *American Economic Journal: Macroeconomics*, 5(1), 1-48.

Grossman, G. M., Helpman, E., Kircher, P. (2013). *Matching and sorting in a global economy.* NBER Wrking Paper No. 19513.

Gruetter, M., Lalive R. (2009). The importance of firms in wage determination. *Labor Economics*, 16(2), 149-160.

Hagedorn, M., Law, T. H., Manovskii, I. (2014). *Identifying Equilibrium Models of Labor Market Sorting* Unpublished manuscript.

Helpman, E., Itskhoki, O., Redding, S. (2010). Inequality and unemployment in a global economy. *Econometrica*, 78(4), 1239-1283.

Helpman, E., Itskhoki, O., Muendler, M.A., Redding, S. (2012). *Trade and inequality: From theory to estimation.* NBER Working Paper No. 17991.

Hillbery, R., Hummels, D. (2013) Trade Elasticity Parameters for a Computable General Equilibrium Model. In Dixon P. and Jorgenson D. (Eds.). *Handbook of Computable General Equilibrium Modeling 1B* (pp. 1213-1269). Amsterdam: Elsevier North Holland.

Kaas, L., Kircher, P. (2014). *Efficient firm dynamics in a frictional labor market.* Unpublished manuscript.

Lise, J., Meghir, C., Robin, J. M. (2013). *Mismatch, sorting and wage dynamics.* NBER Working Paper No. w18719.

Lopes de Melo, R. (2013). *Firm wage differentials and labor market sorting: Reconciling theory and evidence.* Unpublished Manuscript.

Melitz, M. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6), 1695-1725.

Moscarini, G., Postel-Vinay, F. (2014). *Wage Posting and Business Cycles: a Quantitative Exploration.* Unpublished manuscript.

Shimer, R., Smith, L. (2000) Assortative matching and search. *Econometrica*, 68(2), 343-369.

1.A Proof of Proposition 1

I prove the proposition by verifying the following guess: the equilibrium wage is given by:

$$w(a; L_\varphi, \varphi) = \xi_0 rU(a) + \xi_1 \frac{dR_\varphi(q)}{dl(a)} \tag{1.23}$$

i.e. the wage is a linear combination of the unemployment value of a worker and his marginal value product. This leads to a firm's wage bill

$$\int w(a; L_\varphi, \varphi) dL_\varphi(a) = \xi_0 r \int U(a) dL_\varphi(a) + \xi_1 \frac{\sigma - 1}{\sigma} R_\varphi(q) \tag{1.24}$$

With this at hand, we can move to the firm's problem defined by (1.7) and (1.8). The conditional maximization can be written as follows:

$$J(L, \varphi) = \max_{v, L'} \left\{ \frac{1}{1+r} \left[(1 - \xi_1 \frac{\sigma - 1}{\sigma}) R_\varphi(q) - \xi_0 r \int U(a) dL(a) - c(v) + J(L', \varphi) \right] \right.$$
$$\left. \int \lambda(a)((1-s)l(a) + m(\theta) v \frac{u(a)}{u} - l'(a)) da + \int \mu(a) l'(a) da \right\}$$

Taking the first order conditions we obtain:

$$\frac{c'(v)}{1+r} = m(\theta) \int \lambda(a) \frac{u(a)}{u} da \tag{1.25}$$

$$\frac{1}{1+r} \frac{dJ(L', \varphi)}{dl'(a)} = \lambda(a) - \mu(a) \tag{1.26}$$

Since for every a such that $l'(a) > 0$ $\mu(a) = 0$, $\lambda(a)$ defines current marginal value of an additional worker of given type. Thus, the first order condition with respect to v requires that the cost of vacancy were equal to marginal gains from it.

Now we can employ the envelope theorem:

$$\frac{dJ(L, \varphi)}{dl(a)} = \frac{1}{1+r} \left[(1 - \xi_1 \frac{\sigma - 1}{\sigma}) \frac{dR_\varphi(q)}{dl(a)} - \xi_0 rU(a) \right] + \lambda(a)(1-s) \tag{1.27}$$

Fix on steady state and such a so that $l(a) = l'(a) > 0$. For those values we can rewrite (1.27) as

$$\frac{dJ(L, \varphi)}{dl(a)} = \frac{1}{r+s} \left[(1 - \xi_1 \frac{\sigma - 1}{\sigma}) \frac{dR_\varphi(q)}{dl(a)} - \xi_0 rU(a) \right] \tag{1.28}$$

To uncover the left hand side of (1.9), observe that from (1.6) follows:

$$V(a;L,\varphi) - U(a) = (w(a;L,\varphi) - rU(a))/(r+s) \tag{1.29}$$

We can combine the last equation with (1.28) and (1.9), obtaining

$$(1-\beta)[\xi_0 rU(a) + \xi_1 \frac{dR}{dl(a)} - rU(a)] = \beta[(1 - \xi_1 \frac{\sigma - 1}{\sigma}) \frac{dR}{dl(a)} - \xi_0 rU(a)] \tag{1.30}$$

with a method of indeterminant coefficients yielding

$$\xi_0 = (1-\beta) \qquad \text{and} \qquad \xi_1 = \frac{\sigma\beta}{\sigma - \beta} \tag{1.31}$$

Hence, the first claim of the proposition.

With this result we can go further. Observe that (1.5) can be rewritten as

$$rU(a) = b(a) + \frac{\theta m(\theta)}{r+s} \int \gamma(\varphi)\beta(\frac{\sigma}{\sigma-\beta}\frac{dR}{dl(a)} - rU(a))^+ d\varphi \tag{1.32}$$

Analogously, (1.25) can be rewritten as

$$c'(v) = \frac{m(\theta)}{r+s} \int \frac{u(a)}{u}(1-\beta)(\frac{\sigma}{\sigma-\beta}\frac{dR}{dl(a)} - rU(a))^+ da \tag{1.33}$$

1.B Proof of Propositions 2 and 3

By contradiction, suppose that for some $a' > a$ $U(a') < U(a)$. This implies that for any φ

$$s(a,\varphi) = \frac{\sigma\beta}{\sigma-\beta} B(\varphi)\psi(a,\varphi) - rU(a) < s(a',\varphi)$$

with $B(\varphi) = \frac{\sigma-1}{\sigma} Y^{1/\sigma} q_\varphi^{-1/\sigma}$.

From this immediately follows that

$$rU(a) - b(a) = \beta \frac{\theta m(\theta)}{r+s} \int s^+(a,\varphi)\gamma(\varphi)d\varphi < rU(a') - b(a'),$$

which, together with the assumption of non-decreasing income flow in unemployment, contradicts the assertion. Thus, $U(a)$ is increasing in its argument. The part about wage and value in employment for a given φ directly follows from the equilibrium wage rule

(1.12) and value in employment (1.6). Minimal attainable wage

$$\underline{w} = \min_{\varphi : s(a,\varphi) \geq 0} w(a, \varphi) = rU(a),$$

and thus, increasing in a as well. Maximum attainable wage

$$\overline{w} = \max_{\varphi : s(a,\varphi) \geq 0} w(a, \varphi) = \max_{\varphi : s(a,\varphi) \geq 0} \left\{ \frac{\sigma\beta}{\sigma - \beta} B(\varphi)\psi(a, \varphi) + (1 - \beta)rU(a) \right\}$$

is increasing by the envelope theorem.

Analogously, assume by contradiction that for some $\varphi_i > \varphi_j$ $q(\varphi_i) < q(\varphi_j)$. Then, for any worker a:

$$s(a, \varphi_j) = \frac{\sigma - 1}{\sigma - \beta}[YM]^{1/\sigma} q^{-1/\sigma}(\varphi_j)\psi(a, \varphi_j) - rU(a) < s(a, \varphi_i)$$

Again, after integration yields:

$$\frac{(r + s)c'(v_j)}{(1 - \beta)m(\theta)} = \int s^+(a, \varphi_j)\frac{u(a)}{u} da < \int s^+(a, \varphi_i)\frac{u(a)}{u} da = \frac{(r + s)c'(v_i)}{(1 - \beta)m(\theta)}.$$

If the posting cost function is linear, this is a contradiction. Assumption of convex cost implies the first past of the result. Then, immediately $R(\varphi) = Y^{1/\sigma} q^{\frac{1-\sigma}{\sigma}}(\varphi)$ behaves analogously to $q(\varphi)$.

The firm value $J(\varphi)$ is trivially increasing in the firm's type. This follows from the simple argument, akin to the revealed preference. Consider two firms of types $\varphi' > \varphi$. First, firm φ' can choose to produce an amount of output equal to the equilibrium output of firm φ, and hence have the same revenue, using exactly the same combination of types. However, since workers are more productive in firm φ' than in φ, it will require fewer workers, and thus will need to post fewer vacancies. The last step is to show that although firm φ' will pay higher wages to individual workers, the total wage bill will still be smaller than that of firm φ. Given the wage rule (1.12), the total wage bill is

$$\frac{\beta(\sigma - 1)}{\sigma - \beta} R_\varphi + (1 - \beta)r \int U(a) dL_\varphi(a),$$

which is straightforwardly smaller for firm φ' under the described scenario. Thus, I have shown that firm φ' can generate the same revenue as firm φ at smaller expenses. This implies that its equilibrium profit, and value $J(\varphi')$, is larger.

1.C Extended Model Structure and Proof of Proposition 4

I now allow the world to have to identical countries. Each country is the same as the country described in Section 2. Final good and labor markets are country specific, whereas intermediate goods can be traded across the border with impediments. Now, both home and foreign produced varieties of intermediate good can be used in the final good production. The production function becomes:

$$Y = \left[\int\limits_{j\in\Omega^H} q(j)^{\frac{\sigma-1}{\sigma}} dj + \int\limits_{j\in\Omega^F} q(j)^{\frac{\sigma-1}{\sigma}} dj \right]^{\frac{\sigma}{\sigma-1}}, \qquad \sigma > 1, \tag{1.34}$$

where Ω^H is the set of the intermediate goods produced in the home country and Ω^F is the set of the intermediate goods imported from the foreign country. This production function generates demand for the intermediate good of the same structure as before: $p_j = Y^{1/\sigma} q_j^{-1/\sigma}$. In a symmetric equilibrium, final goods output is the same in both countries.

An intermediate goods producer can choose whether to sell all produced quantity on the home market, or to export some of it to the foreign country. Markets are assumed to be segregated, so that firms can charge different prices in different countries. To enter the foreign market, a firm must pay a fixed cost f_j per period in the market, where f_j is drawn from some distribution $F(\cdot)$ independently across firms. Additionally, shipping the good to the foreign country involves an iceberg cost τ, i.e. to deliver and sell a unit of the good to the foreign country, the firm must ship τ units from the home country.

I break the firm's problem down into two steps. First, I describe the optimal way to distribute sales of given amount q of the good between two countries and how much revenue it will generate. The answer to this question is the solution to the following net revenue maximization problem:

$$\max_{q_H, q_F} Y^{1/\sigma} q_H^{\frac{\sigma-1}{\sigma}} + Y^{1/\sigma} q_F^{\frac{\sigma-1}{\sigma}}$$

$$\text{s.t.} \qquad q_H + \tau q_F = q$$

The straightforward solution is to distribute the output so that $q_F = q_H \tau^{-\sigma}$, and therefore, revenue that can be generated from the given amount q of the output is

$$R(q) = [Y(1 + I\tau^{1-\sigma})]^{1/\sigma} q^{\frac{\sigma-1}{\sigma}},$$

where I stands for the indicator of exporting. Now, taking into account the revenue generating function, the firm must decide whether to export, how much output to produce and what type of workers to employ in production. The slightly modified firm's objective function (1.7) becomes:

$$J(L_\varphi, \varphi) = \max_{v,L',I} \frac{1}{1+r} \left\{ R(q(L_\varphi, \varphi), I) - \int w(a,\varphi) dL_\varphi(a) - c(v) - If_j + J(L', \varphi) \right\} \tag{1.35}$$

subject to the hiring constraint (1.8) and $I \in \{0, 1\}$. Bellman equations for workers on the labor market do not change. Since exporting does not alter the structure of the firm problem, i.e. it can be solved for each I with revenue function scaled up proportionally and then maximum value chosen, the result of Proposition 1 applies, and the firm's vacancy creation policy remains the same. Now I prove Proposition 4.

Consider two firms i and j of the same type $\phi_i = \phi_j$ but due to different fixed exporting cost draws, only firm j is exporting. Start by contradiction. Suppose that $q_i > q_j$. Then, $\frac{dR_j}{dl(a)} > \frac{dR_i}{dl(a)}$ and $s(a, \varphi_j) > s(a, \varphi_i)$ for all types a. Following the optimal vacancy posting rule (1.13) would imply that firm j accepts more types and posts (weakly) more vacancies. Given that the production function of two firms are the same and the interfirm labor force size is proportional to the number of vacancies posted, we arrive at a contradiction. Hence, the exporting firm has larger output and larger revenue.

However, the output advantage of firm j cannot be too large. If $q_j > q_i$ to the extent that $\frac{dR_j}{dl(a)} < \frac{dR_i}{dl(a)}$, by analogous reasoning we arrive at a contradiction. Therefore, the exporting firm has (weakly) higher marginal revenue from and match surplus with every worker. This guarantees that the exporting firm has a larger matching set and posts more vacancies. The last assertion also follows because wages depend on firm type only through the worker's share of the surplus.

Chapter 2

“Vanishing Cities:” Can Urban Costs Explain Deindustrialization?

Co-authored by Sergey Kokovin.

2.1 Introduction

Studies of contemporary urban development show a rather typical tendency of many manufacturing facilities to locate in small towns or rural areas. Examples include many sectors from food production to heavy industries such as auto manufacturing, Toyota city being one of the most illuminating examples. Generally, nowadays the physical stages of manufacturing have become one of the least urbanized activities (see, for instance, Holmes and Stevens, 2004, or Kolko, 2010, on the comparison between manufacturing and service industries). Instead, big cities are becoming more and more “deindustrialized”, specializing in exporting services rather than goods. Exportable services include governance of territories by governments, governance of multi-plant firms by headquarters, research, blueprint production, education, etc. Small cities, in contrast, rely on manufacturing and tend to decrease in size. Can we say that the decreasing size of small cities is a *natural* outcome of market evolution, driven by noticeably reduced trade costs? We present a possible explanation of how “... a massive concentration of economic activities within a fairly small number of urban regions... has triggered a process of counterurbanization”

(Tabuchi et al., 2005).[1]

Generally, urban theory and economic geography (see, for example, Fujita and Thisse, 2013) have much to say about the "agglomeration forces" driving both firms and workers into cities, and about countervailing "dispersion forces." Among the latter, commuting costs, land rent, and other diseconomies of scale understandably restrict city size. Alternatively, the well-known Krugman's Core-Periphery model uses agricultural population as a dispersion force. It predicts that large trade costs can support many small cities, whereas *decreasing trade costs force agglomeration into a few large cities.* This view has become popular (see Combes et al., 2008).

The opposite tendency — evolution towards smaller and smaller cities — is also predicted by market theory in several settings. Describing competition between two cities or regions, Helpman (1998) has explored the tension between an agglomeration force stemming from a preference for variety and a dispersion force stemming from a limited housing supply (urban cost). Treating two cities somewhat similarly, Tabuchi (1998) introduces competition for land as a dispersion force. Both models reach similar results on *arising dispersion*: equal distribution of population between two regions when transportation costs become low enough.

More recent advances in the field leave two regions aside and consider more general *systems* of cities instead, searching for their equilibrium number (Tabuchi et al., 2005, Tabuchi and Thisse, 2011). Some find both agglomeration and dispersion tendencies of evolution. Among these studies of city systems, we adopt and modify Anas' (2004) approach. He studies a normative setting where world (or country) population is given, and a social planner maximizes the per-capita welfare by choosing a number of symmetric cities, taking into account consecutive equilibrium. The agglomeration force amounts to economies of scale, whereas the dispersion force stems from urban "commuting" costs. Anas' main theorem describes optimal cities under growing world population as follows: their number increases but their individual sizes *decrease*, and eventually *drop down* to a technologically admissible minimum, sufficient for producing only one variety of manufacturing good (mono-city). The explanation is that the benefit of living in a big city (close to many producers) *decreases* when more and more varieties are imported from other cities in a growing world, while commuting costs remain the *same*. Anas interprets his surprising result as a failure of Krugman's economic geography: without

[1] See also Behrens and Bougna (2013), who report that manufacturing industries are less geographically concentrated in Canada.

externalities, a simple monopolistic competition mechanism is *insufficient to drive the growth of cities* in response to growing populations and/or decreasing trade costs. To better understand the evolution of cities, we first check the robustness of Anas' mechanism and then highlight additional issues: comparative statics of stable equilibria and stable equilibria with developers (the latter concept appearing potentially useful in other models of economic geography).

We find Anas' (2004) modeling strategy appealing and use it as a baseline. In a centripetal role, production scale economies seem sensible as an agglomeration force behind regional development (see empirical evidence on U.S. regional specialization and localization in Kim, 1995). As to the centrifugal role, Krugman's assumption of agricultural industry as a dispersion force has been widely criticized as anachronistic. Instead, commuting, land prices, and other urban costs are perceived as very important for city residents in modern economies (Tabuchi, 1989). We follow this line of reasoning.

However, we suspect that Anas' (2004) normative setting and restrictive assumptions drive his unexpected result. Does it remain valid in more realistic settings? Although Anas' global optimum is an equilibrium in the sense that utility is equalized across existing cities, it ignores the question of stability, which we are focused on. Therefore, instead of a normative model with central planning, we explore two positive alternatives: (1) a stable equilibrium, in which each citizen can voluntarily choose a city to live in, or can settle in a new city (understanding how production and trade will respond to her choice); (2) a stable equilibrium with developers, in which each city decides whether to invite additional citizens or not (also understanding the production/trade consequences). These two versions resemble two cases in the theory of clubs. The migration setting resembles the "open clubs" theory, where everybody can join regardless of the will of city residents. The developers' equilibrium can be related to a "closed clubs" setting, where the admission decision is made by the current members.[2] Otherwise, the modeling remains as in Anas (2004): one sector, general equilibrium, Dixit-Stiglitz preferences, iceberg trade costs, and a technological minimum for a city size called "village". [3]

We start by describing our results with *stable equilibria*. They are multiple; not uniquely determined by preferences and costs. Therefore, we study the "zone of equilib-

[2] We cannot directly rely on club theory, because our clubs-cities are *interacting*: they influence each other through trade, not only through competition for membership.

[3] Although we ignore integer problem throughout the paper, we employ the notion of the minimal technological size to have a well defined equilibrium under any parameters. It can be defined as the minimum number of people necessary to produce one variety of the manufacturing good.

ria". Propositions 5 and 6 describe "vanishing" cities. The zone of parameters (L, ϕ, n) admissible for stable cities turns out to be bounded. This implies that cities must disappear in 3 cases: (i) when the current number n of cities grows larger than a certain uniform bound n^*, or (ii) when the world population L becomes greater than a certain uniform bound L^{d*}; or (iii) when trade freeness ϕ becomes larger than a certain uniform bound ϕ^{d*}.

Therefore, given other parameters, whatever the historical city system is, growth of L, ϕ or n eventually causes our city system to abruptly switch to complete dispersion, due to individual migration. This vanishing effect under migration pressure looks closer to reality than Anas' globally optimal cities, with both exhibiting a striking contrast to Krugman's agglomeration outcome under growing freeness ϕ. An intuitive explanation of the result is the following: Krugman's dispersion force based on agricultural demand decreases with trade freeness, whereas *our dispersion force is the urban cost, and does not change with trade freeness.* At the same time, the agglomeration force is weakening. Similarly, the dispersion force does not change with an increasing world population, but the agglomeration force weakens, due to an increasing share of imported varieties in consumption, meaning that domestic production becomes less important to consumers. The outcome is the same: dispersion. However, besides this limiting case, detailed comparative statics can be more interesting: What precedes the abrupt disintegration of a city system into villages? Proposition 7 states that, under a growing population, stability restrictions can cause city size to either gradually shrink or to abruptly collapse.

To obtain additional predictions, we must limit equilibria multiplicity and define a reasonable *selection* among equilibria. Our stable equilibrium with developers imposes additional restrictions on cities. We assume that citizens are able to restrict entry to their city, or to attract new residents by granting small privileges, and thereby increase average welfare. Such collectively rational behavior is represented by a benevolent city government called the "local developer" or “city mayor” (unlike Anas' global planner, but like entities that maximize the price of land by trying to conform to the citizens' wishes). We explore two versions of such equilibria: "myopic" and "wise", both displaying similar outcomes. It turns out that the zone of (symmetric) stable equilibria with developers is a curve $N(L)$ within stable equilibria, bounded near the origin. At higher L it disappears and the result is disintegration of cities (Proposition 8). Additionally (whenever cities are multiple), the equilibrium city size *gradually decreases* in response to the increasing world population or trade freeness, before dropping down to its technological minimum. This version of the model differs from Anas' global optimization in that its local government ignores the interests of other cities. However, the main conclusion remains the same: the tendency of (manufacturing, industry-specific) towns to decrease and eventually collapse down to “villages” or mono-cities.

Further, for our general explanatory plan it is important to extend our model to two sectors, with decreasing trade cost in only one of them. This is done by combining two (or many) sectors with the Cobb-Douglas upper-tier utility. Since the lower-tier utilities are CES, this guarantees fixed budget shares for both sectors. Then, as we show in a special section, all predictions about a single sector remain valid even in the presence of another sector in which trade costs do not change.

Now we compare the dispersion result uncovered to a group of models of urban systems that display similar effects. There are three main distinctions of this strand of literature from our setup. First, these models have worked with quasilinear preferences, thus, bear a partial equilibrium flavor. Second, the agricultural population serves as a dispersion force, sometimes combined with urban cost. Third, the production side is modeled following Forslid and Ottaviano (2003) as "footloose entrepreneurs" (variable costs are bourn in numerarie but fixed cost requires manufacturing labor). In particular, Tabuchi and Thisse (2006) consider a model with quadratic quasi-linear utility, two regions, two manufacturing sectors and an agricultural sector, with urban costs as an additional dispersion force. As a result, when one good is perfectly mobile, the corresponding industry is partially dispersed, whereas the other is agglomerated, thus showing regional specialization. This conclusion satisfies our need to distinguish physical manufacturing from intellectual production, however, the two-region world is too stylized. A further step towards displaying asymmetric cities in economic geography is an important model of "urban hierarchies" by Tabuchi and Thisse (2011). They consider a model of many manufacturing sectors, each endowed with its own technology and trade costs, and many possible locations on the circumference. When transport costs steadily decrease, some cities expand at the expense of the others by attracting a growing number of industries, while some cities decrease in size or disappear from the space-economy. Though in a different setting (no urban cost), such cities' specialization and related diverse evolution resemble our conclusions.

We find even more similarity in Tabuchi, Thisse and Zeng (2005), who consider migration of “footloose entrepreneurs” in a multi-regional economy with partial equilibrium and quadratic preferences as in Ottaviano et al. (2002). Urban cost is assumed to be increasing in the city population. When the number of cities is forced to be constant, some types of cities may grow and other cities simultaneously shrink with trade costs. When the number of cities is allowed to vary, the city size first increases (only under some parameters) and then decreases in response to decreasing transport costs. This result shows similarity to ours on the decreasing stage. We believe that the possible difference in monotonicity under large transportation costs stems from the assumption of an agricultural population rather than from linear demand structure.

We stress that we study evolution of various cities in a different setting: general equilibrium à la Dixit-Stiglitz, one-factor technology with increasing returns without fixing the world's total mass of firms (presuming it is less stylized than Forslid and Ottaviano's approach). Thereby we show that the vanishing effect is not the consequence of very specific assumptions like partial equilibrium and simplified cost function, but holds also in other settings. Still, the main difference from the previous approach is that cities in this model do not include all industries; they specialize either in knowledge-intensive intellectual products, or in manufacturing. We postulate this feature, leaving any models with complete specialization that arise as an equilibrium outcome for future research. Instead, the goal of our study is to show that some industries may become dispersed, while other industries remain agglomerated—just because of the simple tendency revealed by Anas: decreasing trade costs in the former and stable ones in the latter.

Overall, our study generally supports the prediction that *cities comprised of industries lacking intra- or inter-industry externalities have a tendency to decrease and eventually reach their minimal technological size.*

The rest of the paper is organized as follows: Section 2 presents our baseline model. Section 3 studies migration equilibria. Section 4 deals with developers' equilibria. Section 5 considers two sectors extension of the model, and the final section concludes. All proofs are relegated to the Appendix.

2.2 Model: System of Cities with Migration

We introduce a model of a city system very close to that of Anas', except for: (1) possibly asymmetric cities and (2) a migration process instead of a social planner setting. We start with the description of the internal city structure and then embed it into a system of cities.

City. Traditionally, we consider monocentric and circular cities endowed with a Central Business District (CBD) where production and trade take place. The only production input is labor supplied by consumers – citizens. Each consumer needs one unit of land and possesses a unit of time which she spends commuting to her workplace (CBD) and laboring. The cost of commuting is s units of time per unit of distance, therefore, a consumer living at distance x from CBD spends sx units of time for commuting and supplies $h(x) = 1 - sx$ units of labor for production. Wage per unit of working time in the CBD is denoted as w.

Suppose there are N residents in a given city. Then, because of unit land requirement, the radius of the city becomes $r = \sqrt{N/\pi}$. Given individual labor supply and uniform distribution of citizens within the city, overall labor supply for production H is given by

$$H(N) = \int_0^r 2\pi x h(x) dx = \pi r^2 - 2\pi r^3/3 = N - kN^{3/2}, \tag{2.1}$$

where $k \equiv 2s/3\sqrt{\pi}$ is a constant, summarizing commuting cost. We denote the average (per-citizen) labor supply in the city as

$$\theta(N) \equiv H(N)/N.$$

In addition, we assume zero opportunity value of land and free reallocation within the city.

Now, we explain how redistribution of rent makes income proportional to the labor supply. Suppose that the city size is r. We have normalized the land rent on the edge of the city to zero. Since consumers face the same price vector, free reallocation of consumers should lead to disposable income equalization among them. In addition, we assume that the local government collects the land rent and distributes it equally among citizens in a form of lump-sum transfer. Moreover, independently of the structure and size of other cities, land rent always remains within the city where it is collected. Thus, the disposable income of every citizen in the city of size N is

$$I(N) = \theta(N)w \equiv (1 - k\sqrt{N})w,$$

which is decreasing in size.[4]

[4] To see details of rent redistribution, see that at any location within a city the sum of rent cost and commuting cost must be the same. Hence, the rent at any point x must be $R(x) = s(r - x)w$, and total rent in the city is $TR = \int_0^r 2\pi xs(r - x)wdx = \pi swr^3/3 = kwN^{3/2}/2$.

System of cities and goods. Suppose there are n cities with population masses $(N_1, N_2, ..., N_n)$. There is only one differentiated good in the economy. Each variety of the good is produced by only one firm residing in some one city. All cities trade with each other through some common "hub", i.e., transport costs for each pair of cities is the same. The market for varieties is monopolistically competitive, and entry to the market is free.

Consumers have identical Constant Elasticity of Substitution (CES) preferences over the set of varieties:

$$U = \left[\sum_{i=1}^{n} \int_0^{m_i} x_{ki(j)}^{(\sigma-1)/\sigma} dj \right]^{\sigma/(\sigma-1)}, \tag{2.2}$$

where $x_{ki(j)}$ is a single purchase of variety j produced in city i and consumed in city k. Parameter $\sigma > 1$ denotes the elasticity of substitution. A consumer of type k maximizes in x_k her utility (2.2) subject to the budget constraint

$$\sum_{i=1}^{n} \int_0^{m_i} p_{ki(j)} x_{ki(j)} dj \leq I(N_k), \tag{2.3}$$

where $p_{ki(j)}$ is the price of variety j produced in city i and consumed in city k. Labor is the numeraire, $I(N_k)$ denotes income. Taking the first-order conditions and expressing the Lagrange multiplier from the budget, we obtain the consumer *demand function* $\mathbf{X}(\cdot)$ in the form:

$$\mathbf{X}_{ki(j)}(p_{ki(j)}, I_k, P_k) = p_{ki(j)}^{-\sigma} I(N_k)/P_k^{1-\sigma} \qquad P_k = \left[\sum_{i=1}^{n} \int_0^{m_i} p_{ki(j)}^{1-\sigma} dj \right]^{1/(1-\sigma)}, \tag{2.4}$$

with P_k being a price index. It is "perfect", as a price of one unit of utility, in the sense that the indirect utility of a consumer in the city k is

$$V_k = I(N_k)/P_k. \tag{2.5}$$

Production. Each producer is a price-maker for her variety. As is standard in monopolistic competition literature, we assume that a producer has fixed labor cost F to set up a plant and marginal labor requirement c of production. Trade within a city is costless, whereas trade with other cities requires iceberg transportation costs. This means that supplying one unit of a good from city i to city k requires $\tau_{ki} = \tau > 1$ units of the good when $i \neq k$ but $\tau_{ii} = 1$. Under these assumptions each producer j in city i

has a profit function

$$\pi_{i(j)} = \sum_{k=1}^{n} [p_{ki(j)} - \tau_{ki} c w_i] \mathbf{X}_{ki(j)}(p_{ki(j)}, I_k, P_k) N_k - F w_i \tag{2.6}$$

which she maximizes with respect to prices subject to demand functions (2.4) taking the price indexes as given. As is standard, under CES preferences, such a profit function is concave and has a unique maximum. Then, the symmetry of producers leads to *symmetric* pricing by all firms from a given city. This allows us to drop index j from further discussion. Producer optimization leads to:

$$p_{ki} = \frac{\sigma \tau_{ki} c w_i}{\sigma - 1} \qquad \pi_i = \left(\sum_{k=1}^{n} \frac{\tau_{ki} c x_{ki} N_k}{\sigma - 1} - F \right) w_i. \tag{2.7}$$

Free entry into the market drives firms' profit in every city to zero. Combining zero-profit conditions (2.7) with labor market clearing yields an equilibrium mass of varieties in every city:

$$m_i = \frac{N_i \theta(N_i)}{F \sigma} \quad \forall i. \tag{2.8}$$

Finally, equilibrium wages, w_i, and corresponding prices, p_{ki}, and incomes, $I(N_k)$, can be obtained from market clearing for a representative variety produced in each city:

$$\sum_{k=1}^{n} \frac{\tau_{ki} p_{ki}^{-\sigma} I(N_k) N_k}{P_k^{1-\sigma}} = (\sigma - 1) F / c \quad \forall i \tag{2.9}$$

Trade equilibrium associated with a system of n cities of sizes $(N_1, N_2, ..., N_n)$ is defined as a bundle $\{x_{ki}\}_{i=\overline{1,n}}^{k=\overline{1,n}}$ of consumption values, a bundle of prices $\{p_{ki}\}_{i=\overline{1,n}}^{k=\overline{1,n}}$, vector of varieties masses $\{m_i\}$ and vector of wages $\{w_i\}$ such that: (i) consumption values solve consumers' problems (2.2) subject to budget constraint (2.3) under given prices, wages and available varieties; (ii) prices solve producers' problems (2.6) given demand function (2.4), price indexes and wages; (iii) firms earn zero profit (free entry); (iv) labor market and market for every variety clear.

We do not discuss the existence of such general equilibria, pointing out later on the existence of symmetric ones (whose behavior under small perturbations we study). Further, every trade equilibrium delivers indirect utility $V_k = \theta(N_k) w_k / P_k$ to any consumer in city k. Suppose the world population amounts to L consumers.

As is standard, we call n cities of sizes $(N_1, N_2, ..., N_n)$ a **migration equilibrium** if

(1) $\sum_{i=1}^{n} N_i = L$ and (2) related trade equilibrium yields the same level of indirect utility across cities: $V_k = V_i \quad \forall k, i$.

Naturally, the symmetric distribution of population across an arbitrary number of cities is a migration equilibrium. Indeed, the case of symmetric cities implies symmetric trade equilibrium. In this case cities are interchangeable, and utility is equalized across cities. However, our goal is to understand when such symmetric migration equilibrium is *stable*, in the sense that small perturbations are not amplified. Therefore, we shall consider mainly symmetric (or close to symmetric) population distributions across cities. Let us reserve notation (n, N) for the symmetric equilibrium with n cities of size N, so that $L = nN$.

2.3 Migration Stability

In this section we discuss the stability of any symmetric equilibrium (n, N) against small perturbations in population distribution. First, we define two stability conditions based on different kinds of perturbations: migration to the countryside (unpopulated locations) and migration to other cities.

1. Consider a system of slightly asymmetric cities. Starting from n cities of size N, suppose that a new $(n+1)$-st city of size ε is created, with one of the old cities taking size $\tilde{N} = N - \varepsilon$. If in new trade equilibrium indirect utilities V_i evaluated at point $\varepsilon \approx 0$ satisfy $V_{n+1}(\varepsilon) < V_n(\tilde{N})$, we say that this migration equilibrium (n, N) is (strictly) **stable against dispersion**; otherwise it is not.

2. Consider a system: 1-st city of size $N_1 = N + \varepsilon$, 2-nd city of size $N_2 = N - \varepsilon$ and $n - 2$ cities of size N. If in related trade equilibrium incremental utility $\frac{dV_1(N+\varepsilon)}{d\varepsilon}$ evaluated at the point $\varepsilon \approx 0$ is negative, we say that migration equilibrium (n, N) is (strictly) **stable against agglomeration**, otherwise it is not.

When a symmetric migration equilibrium (n, N) satisfies both stability conditions, we call it a **stable equilibrium**.

The first requirement of stability is that a small shift of population from a city into a previously unpopulated area does not create incentives for mass movement to this newly created town. The second requirement is that small movements of populations from one city to another do not make the target city more attractive. We must add that, in reality, there can be more sophisticated deviations from the stable state: any vector of changes

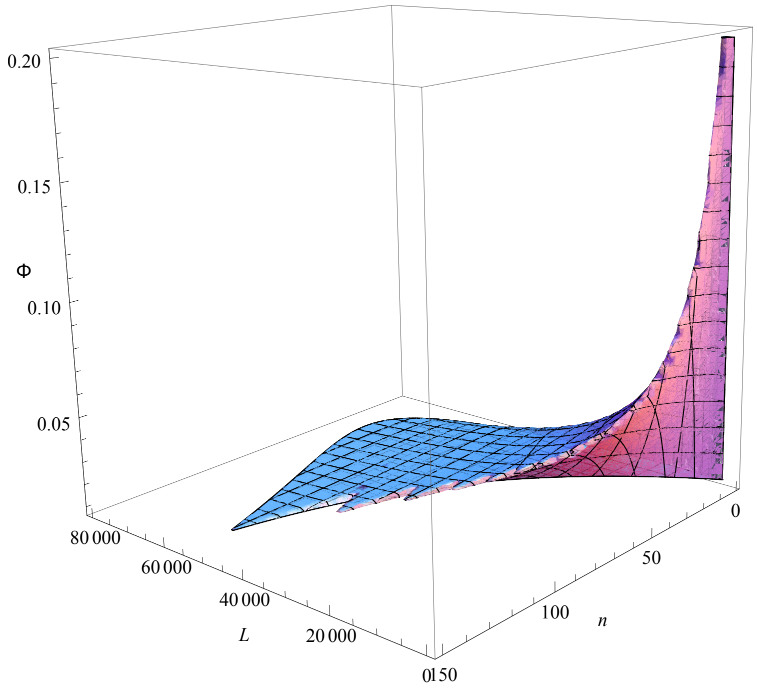

Figure 2.1: Region of stable equilibria.

in populations; thereby real stable equilibria could be narrower that what we call stable. However, our definition is sufficient to show that the stability zone is bounded.

Now we formulate the conditions when a symmetric equilibrium is stable in both senses.

Lemma 1. (Stability conditions) *(1) Symmetric migration equilibrium* (n, N) *is stable against dispersion if and only if*

$$1-k\sqrt{\frac{L}{n}} > \left[\frac{1+(n-1)\tau^{1-\sigma}}{n\tau^{1-\sigma}}\right]^{-\frac{1}{\sigma}-\frac{1}{\sigma-1}}. \tag{2.10}$$

(2) Symmetric migration equilibrium (n, N) *is stable against agglomeration if and only if*

$$\frac{2\sigma-1}{(\sigma-1)\left(\sigma-1+\sigma\frac{1+(n-1)\tau^{1-\sigma}}{1-\tau^{1-\sigma}}\right)} < \frac{k\sqrt{L}}{2\sqrt{n}-3k\sqrt{L}}. \tag{2.11}$$

It may be interesting to look at the shape of stable combinations (L, ϕ, n) from Lemma. This region is displayed in Fig. 2.1 for specific values $\sigma = 11$, $k = 0.005$. *All* combinations inside this shaded area generate stable equilibria, because our Lemma gives necessary and sufficient conditions.

In Fig. 2.1, we see that the zone of stable equilibria is *bounded* in all three dimensions L, n, ϕ. This zone has a complex saddle-type shape; it looks like a hill with a grotto underneath (see our subsequent figures for details). Our further plan is to prove such boundedness for *any* parameter values k, σ. In some sense, this means studying the comparative statics of sections of the shaded area from Fig. 2.1. Specifically, to correctly resolve Anas' question about "vanishing cities", we now describe how the *region of stable migration equilibria* changes with the population of the whole system and/or with trade frictions.

First of all, we simplify the notation by (conventionally) introducing trade freeness $\phi \equiv \tau^{1-\sigma} \in [0, 1]$. This measure is decreasing in the elasticity of substitution, σ, and higher ϕ implies freer trade. For any number n of cities, the condition of stability against dispersion can be reformulated in two alternative ways:

(1) given trade freeness ϕ, the total population is bounded from above as

$$L(n) \leq L^d(n) = \frac{n}{k^2}\left[1 - \left(1 + \frac{1-\phi}{n\phi}\right)^{-\frac{1}{\sigma} - \frac{1}{\sigma-1}}\right]^2;$$

(2) given total population L, freeness of trade is bounded from above as

$$\phi(n) \leq \phi^d(n) = \frac{1}{1 + n\left[\left(1 - k\sqrt{L/n}\right)^{-\frac{2\sigma-1}{\sigma(\sigma-1)}} - 1\right]}.$$

Similarly, under any number of cities n, the condition of stability against agglomeration requires that:

(1) given freeness of trade ϕ, the total population is bounded from below:

$$L(n) \geq L^a(n) = \frac{n}{k^2}\left[\frac{2}{\sigma + 2 + \frac{(\sigma-1)\sigma n\phi}{(2\sigma-1)(1-\phi)}}\right]^2;$$

(2) given total population L, freeness of trade is bounded from below:

$$\phi(n) \geq \phi^a(n) = \frac{1}{1 + n\frac{\sigma(\sigma-1)}{(2\sigma-1)\left(\frac{2\sqrt{n}-3k\sqrt{L}}{k\sqrt{L}} - \sigma + 1\right)}}.$$

In other words, two kinds of stability conditions provide upper and lower bounds on parameter values under which stable equilibria may exist. Using these bounds, the following proposition shows that population growth or trade liberalization (increasing freeness) must lead to an absence of stable equilibria.

Proposition 5. (No stable cities in large/free world)

(1) Maximal stable population $L^d(n)$ is uniformly bounded from above; i.e., there exists such $L^{d*} < \infty$, that any equilibrium $(n,\ L/n)$ is unstable against dispersion whenever $L > L^{d*}$;

(2) Maximal stable freeness $\phi^d(n)$ is uniformly separated from one; i.e., there exists such $\phi^{d*} < 1$, that any equilibrium $(n,\ L/n)$ is unstable against dispersion whenever $\phi > \phi^{d*}$.

In other words, *if the world is large enough or trade is free enough, the only stable outcome is the dispersion of the population* to "villages", i.e. locations of minimal admissible size. The remaining question is the boundedness of the region of stability (Fig.2.1) in dimension n. In other words, we ask whether stable symmetric equilibria with large number of cities n exist. The next proposition precludes this possibility and, therefore, gives additional credibility to the "vanishing cities" theory.

Proposition 6. (No stable equilibria with many cities). *Under any admissible parameters (L, ϕ, k, σ), there exist some $\bar{n}$ such that any equilibrium with bigger number $n > \bar{n}$ of cities is unstable.*

Although our stability conditions reduce the amount of possible equilibrium configurations to a bounded zone in (ϕ, L, n) space, there is still a continuum of stable equilibria for admissible parameter values. Indeed, any existing city creates a lock-in effect, preventing creation of new cities. This result is akin to that of Fujita, Krugman and Mori (1999), who have found continuum of equilibria in a city system with an explicit linear space structure. Fujita et al. employ evolutionary dynamics for selection among equilibria, however, due to the complexity of their space structure, they have been forced to resort to numerical simulations. On the contrary, the simplified spacial structure of our model allows us to describe all stable equilibria and to provide analytical selection by simplified evolutionary dynamics.

Comparative statics. How does the system of cities *change* when the population grows or trade costs decrease? Do the cities in our model grow, gradually decrease, or collapse? We first explain numerical simulations, interpret them, then develop them into a proposition.

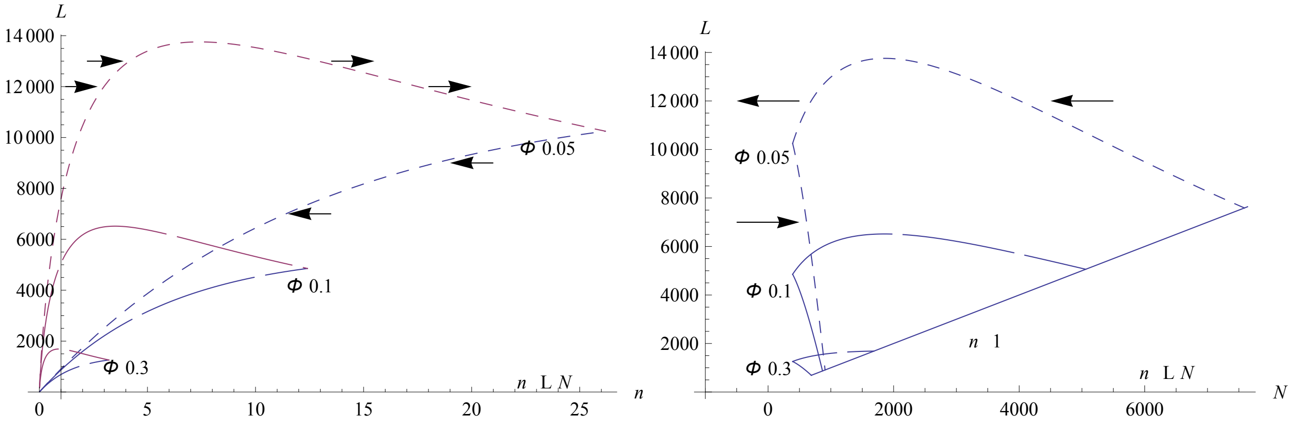

Figure 2.2: How the region of migration stable equilibria shrinks with respect to trade freeness ϕ.

Fig. 2.2 plots the regions of migration stable equilibria in (n, L) coordinates under changing trade freeness: $\phi = 0.05$, 0.1, 0.3. In essence, the left panel of Fig. 2.2 contains some sections of a stable zone from Fig. 2.1 in (n, L) plane. The right panel inverts this zone into (N, L) space, and the straight line, $n = 1$, cuts away cases with less than one city (the same as the vertical line with abscissa coordinate 1 does in the left panel). The main observation is that *the larger trade freeness is — the smaller in terms of area the zone of stable equilibria is.* Specifically, the boundaries of the zone shrink towards the origin, making the stable area smaller.

First consider the direction of changes outside the stability zone: Does the family of cities move towards stability or towards collapse? The arrows outside the zone show the migration tendency. In the right panel we observe that above the upper boundary, the tendency works to *decrease* size N of a city (which is shown in the left panel as increasing $n = L/N$). The arrow in the right side of the right panel says that when a too-large unstable city decreases its population, it can reach the region of stability. A similar stable result is shown by the lower left arrow, which is *below the kink* breaking the left boundary. Instead, the upper-left arrow (*above the kink*) in this panel says that, under sufficiently high population, a too-small city further loses its population and *collapses* into a village. This (upper left) boundary of the stable zone is unstable.

With this in mind, we use this figure to express our intuitions about possible changes in the city system. To grasp the possible impact of a growing population L on cities under given ϕ, consider the following thought experiment. Assume, for instance, $\phi = 0.3$. Suppose we start with only one settlement ($n = 1$) and with a small population $L = 2$: Adam and Eve. What happens? This historical point of urbanization lies below the critical value $L^a(2)$, i.e., *below* the lower bound of related stability zone. So, it cannot happen that the couple live apart, in different villages. Instead, they must agglomerate. Now let the population grow. Then the agglomeration tendency remains: everyone lives in the same city. The picture tells us that this single-city pattern of urbanization will persist until the population exceeds approximately 600. From this point onwards, our growing population can either remain in this city or try to settle a new one. A two-cities equilibrium becomes possible when the population reaches approximately 1,100 but there is no force to shift the system to that other equilibrium. However, when the population exceeds 1,900, any n-city system loses stability; it abruptly collapses into villages consisting of 1 citizen each (the minimal technological size).

Observe that different levels of trade cost make a qualitative difference. Under $\phi = 0.1$ or $\phi = 0.05$, on the upper border of the stability region there is a possibility that a growing population may result in a *gradually growing number of cities* instead of abrupt dispersion, at least on the ascending wing of this region (though each city decreases in size). Generally, under typical parameters, this fairy tale and related picture support the idea that *either gradual or abrupt decreases in city sizes* are possible in response to growing population or/and trade freeness. We formulate this tendency as a proposition.

Proposition 7. *Assume that during growth of the population under fixed other parameters, the number of cities remains stable until the system reaches the border of the stability zone. Then, the shape of the border governs the evolution of city sizes: further evolution can display either gradually decreasing city size or abrupt collapse of cities, but not an increase.*

Proof. In our stability condition (2.10), and in Fig. 2.2, we see that it is the dispersion condition (not the agglomeration one) that limits the size of the stable world from above. Therefore, this condition governs the comparative statics. Its violation triggers the increase in the number of cities. From the formula we see that the related city size $L^d(n)/n$ is a decreasing function. □

2.4 Developers' Stability

Although the notion of migration stability allows us to reduce the number of plausible equilibria and ensures our impossibility results, the remaining multiplicity of stable equilibria is somewhat disturbing. Therefore, in this section we develop another notion of stability as a selection criterion: stability against actions by a "developer" (local government) who aims to maximize the representative citizen's utility in her city and who has some power to invite in or push out citizens. Of course, it would be fair to call this actor "benevolent local government" or "city mayor," but if we believe that the benefits of a particular city structure to citizens can be capitalized in the land rent, there should not be any difference between a mayor's behavior and that of a developer. However, this decision-making differs from Anas' global planner. It also differs from simple migration — because it considers all intra-city benefits from inviting a new citizen in or forcing a citizen to exit — instead of considering personal benefits to a migrant. Nevertheless, we still impose the requirement of stability against dispersion or creation of a new city, because the developer cannot force citizens to stay in the city. We consider two cases with similar outcomes: a wise developer and a myopic developer.

Wise developer. We assume in this paragraph that each developer correctly predicts all changes in the trade equilibrium that will occur after a new citizen is invited from some other city.

(Symmetric) *stable equilibrium of wise developers* is a system of cities (n, N) such that it is a strict local Nash equilibrium among n developers choosing their city sizes. It means that there is an $\hat{\varepsilon} > 0$ such that all possible local ε-perturbations of the city population ($\varepsilon < \hat{\varepsilon}$) bring strictly negative value changes to a developer.

This notion does not consider the possibility of new cities and other asymmetric situations. By our assumption of wise predictions, the changes in the equilibrium trade and welfare coincide with predictions that we made when studying migration, because the wise developer understands that she can attract a new citizen only from some other city. Further, she expects a citizen to join some other city (not to die) once expelled from hers. This allows us to show that the new concept of equilibrium is a selection from the previous concept. Namely, under any ϕ, the wise developer's equilibrium is the lower border of related equilibria zone displayed in Fig. 2.2. Interestingly, a wise developer would chose a system with the largest number of cities and the smallest city size among migration stable equilibria.

Indeed, if the migrant goes out and thereby *decreases* welfare in the destination city, by symmetry, she *increases* welfare in her city of origin. Only when the derivative of welfare with respect to migration is zero can the situation be a wise developers' equilibrium. Thus, we come to

Proposition 8. (Wise developers) *(i) A system of cities (n, N) is a stable equilibrium of wise developers only if it satisfies stability against agglomeration as equality:*

$$L = L^a(n) = \frac{n}{k^2}\left[\frac{2}{\sigma + 2 + \frac{(\sigma-1)\sigma n\phi}{(2\sigma-1)(1-\phi)}}\right]^2; \tag{2.12}$$

(ii) It belongs to stable equilibria. Thereby, the developer's equilibrium remains possible within same three bounds: world population, trade freeness and number of cities—all must be small enough.

Myopic developer. Although the introduced notion is a rational one, we find our requirements for the local government to have perfect foresight too demanding. Indeed, the setup requires a developer to predict changes not only in her city, but in all cities throughout the country as well. To relax this requirement, we introduce the notion of a myopic developer. More precisely, we assume that each developer is myopic (boundedly-rational) when predicting outside trade consequences caused by excluding or inviting a citizen. This means that when maximizing welfare in her city, she expects no response from all relevant variables in other cities: population, price indices and wages. Suppose there are $n-1$ cities of size N, whereas #1 developer's city has size N_1 (to be optimized). Then, given symmetry in $n-1$ other cities, the (trade) equilibrium conditions for price index and wage in developer's city can be formulated as:

$$P_1^{1-\sigma} = \frac{N\theta(N)}{F\sigma}\left(\frac{\sigma\tau c}{\sigma-1}\right)^{1-\sigma}(n-1) + \frac{N_1\theta(N_1)}{F\sigma}\left(\frac{\sigma c w_1}{\sigma-1}\right)^{1-\sigma} \tag{2.13}$$

$$\frac{(n-1)\tau\theta(N)N[\sigma\tau c w_1/(\sigma-1)]^{-\sigma}}{P_i^{1-\sigma}} + \frac{\theta(N_1)N_1 w_1[\sigma c w_1/(\sigma-1)]^{-\sigma}}{P_1^{1-\sigma}} = (\sigma-1)F/c \tag{2.14}$$

This form of equilibrium conditions is standard. However, the developer's optimization with respect to N_1 is different, since she takes N and P_i as given. Denote elasticities of price index and wage (perceived by the developer) in city #1 with respect to local population N_1 as $\varepsilon_d^{P_1}$ and $\varepsilon_d^{w_1}$, respectively.

Definition. We call a symmetric equilibrium (n, N) *stable against a myopic developer* if elasticity of indirect utility (perceived by developer) is equal to zero $\varepsilon_d^{V_1} \equiv \varepsilon^\theta + \varepsilon_d^{w_1} - \varepsilon_d^{P_1} = 0$ evaluated at $N_1 = N$ and the equilibrium stable against dispersion.

Proposition 9. (Myopic developers) *(i) A system of cities (n, N) is a stable equilibrium* *of myopic developers only if it satisfies the following condition:*

$$L = L^m(n) = \frac{n}{k^2}\left[\frac{2}{\sigma + 2 + \frac{(\sigma-1)\sigma n\phi}{(2\sigma-1)}}\right]^2; \tag{2.15}$$

(ii) It belongs to migration-stable equilibria. Thereby, the developer's equilibrium remain possible within same three bounds: world population, trade freeness and number of cities — must all be small enough.

The established condition for stability against actions by a myopic developer is similar to the condition of stability against agglomeration (2.11). It differs only by the multiplier $(1-\phi)$ in the last term of the denominator. Thus, the costlier trade is, the more developer stability behavior resembles that of stability against agglomeration (and that of a wise developer). The intuition is straightforward: cities affect each other through trade only. Therefore, the higher trade costs are, the less impact a developer's city has on other cities, and the smaller the developer's mistake in assuming a lack of change in other cities will be. Moreover, the developer's myopia pushes the system towards fewer cities, i.e, larger size: $L^m(n) > L^a(n)$.

Now we present comparative statics analysis of the stability against a developer's action graphically. Fig. 2.3 is a copy of Fig. 2.2 supplemented with the line of developer's stable equilibria $L^m(n)$ and its counterpart $\tilde{L^m}(N)$. Observe that when the world population L grows, the related point on the solid curve of the developer's equilibria moves to the right in the left panel. Its counterpart shifts to the left in the right panel, which describes the same equilibrium in terms of the city size $N = L/n$. Such behavior means that *the number of cities increases in response to population growth*, whereas *the city size decreases.*

A similar conclusion follows for trade freeness, only the comparison does not go along each solid curve, but across three curves. When freeness increases, the point of equilibrium (for any size of the world L) goes to the right in the left panel and to the left in the right panel. This again means that *the number of cities increases in response to decreasing trade costs*, whereas *the city size decreases.* Indeed, $L^m(n)/n$ is again a decreasing function of ϕ, meaning that the new equilibrium must have smaller city size and, hence, a larger number of cities. Thus, we have come to the proposition which was the purpose of introducing the developers' equilibria.

Proposition 10. *Consider the growth of population L under fixed other parameters, or*

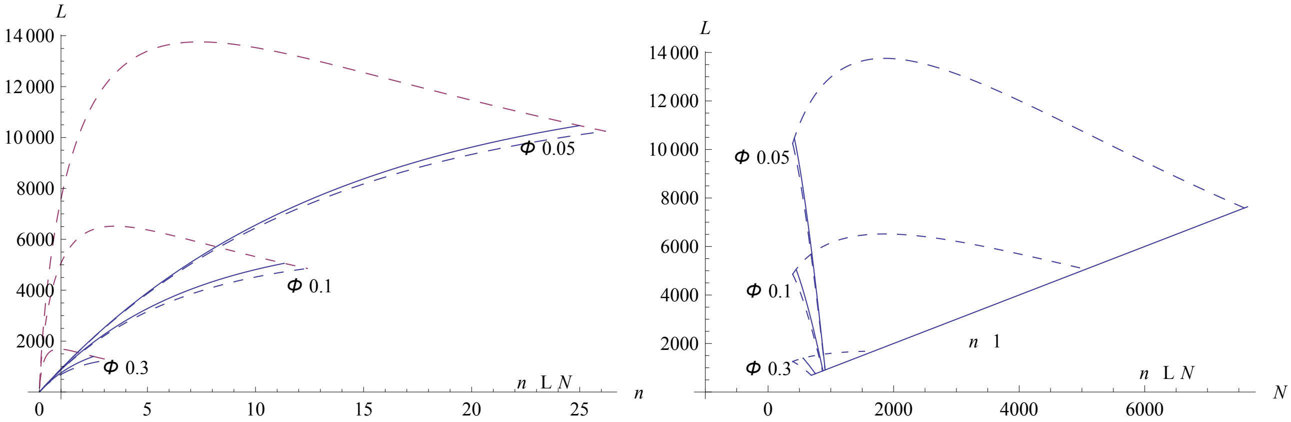

Figure 2.3: Developer stable configurations in coordinates (n, L) or (N, L) (under $k = 0.001$, $\sigma = 11$).

decreasing trade cost τ under fixed other parameters. In each case, both wise and myopic developers' equilibrium displays either gradually decreasing city size or abrupt collapse of cities, but not an increase in city size.

2.5 Extension to Two Sectors

A thoughtful reader has noticed the discrepancy between our theoretical framework and its empirical interpretation as a divergent evolution of manufacturing cities in contrast with other cities. Describing the "deurbanization" of a particular manufacturing industry among others, we have hitherto dealt with a general equilibrium model with one sector only. Now, we show how our setup may be embedded into a multi-sector framework, and maintain qualitatively similar results.

Model. Assume our small-city system remains the same. However, now it also trades through the same hub with a region called "capital city" (our results would also hold true for *several* fixed big cities). This city completely belongs to another sector; it produces some aggregate good S, using its own sector-specific labor. We prefer to interpret this good as "tradable services," including blueprints, research, governance, etc. However, the capital residents share same preferences as our provincial residents. Any citizen here or there consumes two goods: composite good U produced in our provincial cities (defined as 2.2) and good S produced in the capital city. Preferences for the two goods have a standard Cobb-Douglas form:

$$\check{U} = \left(\left[\sum_{i=1}^{n} \int_{0}^{m_i} x_{kji}^{(\sigma-1)/\sigma} dj \right]^{\sigma/(\sigma-1)} \right)^{\mu} S_k^{1-\mu}$$

where $\check{U}$ denotes overall utility of consumption and the Cobb-Douglas parameter μ satisfies $0 < \mu < 1$. Utility is maximized under the budget constraint

$$\sum_{i=1}^{n} \int_{0}^{m_i} p_{kji} x_{kji} dj + P_s S_k \leq I(N_k)$$

that includes the outside good and its final price P_s, which may include the cost of transportation to our cities. In what follows, we treat the second commodity S as a homogeneous good produced under constant returns (another possibility would be a diversified good with some CES preferences and price index P_s). Then, spending some mass L_s of sector-specific labor with unit productivity, the capital produces as much as L_s. It consumes $C_s = (1-\mu)L_s$ out of L_s, because of the well-known two-stage budgeting rule, resulting from Cobb-Douglas-CES preferences. The reminder of L_s goes for export and, after iceberg reduction during transportation, becomes the aggregate provincial import $S = \sum_{k=1}^{n} S_k N_k = \mu L_s / \tau_s$. Here τ_s is the transport coefficient for good S. Thus, S is pinned down as a constant. Further, the trade balance between the capital and the province should determine the terms of trade P_s/P through equation

$$\sum_{i=1}^{n} \int_{0}^{m_i} p_{sji} y_{sji} dj = P_s S$$

where y_{sji} denotes exports of manufacturing from the province, i.e., our provincial output net of provincial consumption and net of transportation losses. The profit function of a provincial firm now takes into account exports to the capital, but no additional analysis is needed, because the profit-maximizing pricing rule (2.7) remains the same. The internal structure of our provincial cities remains the same. Our firms and developers take the price P_s of "services" as given.

An *allocation* includes: a mass n of (symmetric) provincial cities, a mass m of varieties, a consumption vector $(\mathbf{X}, \mathbf{S})$ in each of the cities, a consumption vector in the capital, and production and prices of all goods. An allocation is an *equilibrium* if it satisfies natural conditions for provincial cities: rational behavior of consumers, producers, and balanced budget constraints, plus similar conditions for the capital city (we do not need to detail the consumption and production in the capital because of well-known two-stage budgeting under Cobb-Douglas-CES combination: budget shares for both goods in both worlds always remain constant).

Analysis. Instead of previous indirect utility (2.5), now, under Cobb-Douglas preferences, the indirect utility in any provincial city i is standardly determined (up to a constant multiplier) as income divided by the weighted product of prices, as follows:

$$V_i = \frac{\theta(N_i)w_i}{P_i^{\mu} P_s^{1-\mu}}. \tag{2.16}$$

Now since the shipment of good S is fixed, we can normalize its unites so that $S = 1$. With this normalization and a citizen's disposable income $\theta(N_i)w_i$, the trade balance between the province and the capital determines the price of services:

$$P_S = (1-\mu)\sum_i^n N_i\theta(N_i)w_i \tag{2.17}$$

We focus now on the stability of manufacturing cities against dispersion to villages. From (2.17) we see that if a new zero-size town is established, the price of the outside good is unaffected. Further, under CES preferences, the size of firms is fixed, and thus the number of our varieties (2.8) and their prices (2.7) are unaffected by the presence of the capital. Therefore, the stability condition analogous to (2.10) now requires $w_{n+1} < \theta(N)(P_{n+1}/P_n)^{\mu}$. The stability condition analogous to (2.10) now becomes

$$1 - k\sqrt{\frac{L}{n}} > \left[\frac{1+(n-1)\tau^{1-\sigma}}{n\tau^{1-\sigma}}\right]^{-\frac{1}{\sigma}-\frac{\mu}{\sigma-1}} \tag{2.18}$$

Lemma 2. *If the stability condition (2.18) is violated in a one-sector world (for $\mu = 1$), it is also violated in a two-sector world (for $\mu < 1$).*

Indeed, since the bracketed term in (2.18) is greater than one, when its power increases ($\mu \downarrow$) the right-hand side increases and the inequality can become violated.

Corollary 1. *All propositions about bounded zone of stable migration equilibria and developers equilibria in a one-sector world (for $\mu = 1$) remain valid in a two-sector world (for $\mu < 1$). The zone of stable equilibria must shrink.*

Economically, due to trade costs, the price index in a newly established small city is larger than that in the old ones, this difference being an agglomeration force. Now, in the presence of $\mu < 1$, this force fades due to a lower share of manufacturing in the expenditure of a citizen. Thus, $\mu < 1$ reinforces the dispersion result. This argumentation bridges our Introduction with the developed theoretical setup.

Testable prediction. Importantly, the two-sector version of the model sets a question for empirical work. Indeed, the comparative statics in μ suggest a testable prediction: under Cobb-Douglas preferences, parameter μ represents the share of income spent on manufacture. Recall also (see Fig. 2.3 and related discussion) that the cities' size decreases in response to decreasing trade costs, and growing world population. A new topic is comparative statics in parameter μ. It should push cities' size in the same direction as trade freeness ϕ, because both work as dispersion forces against agglomeration. Thus, in a cross-section comparison of industries, one would expect to observe a *negative* correlation between the urbanization level of each sector and the share of consumer spending on its product. To conduct this task, one needs to control only for technological differences in transportation and employment. Cost parameters (c, F) do not affect city sizes. To the best of our knowledge, this hypothesis is novel, and can be tested in future work.

Although we do not explicitly model the reasons for "exportable services" being produced exclusively in the capital or few cities, it is not difficult to imagine a number of potential explanations for that. First of all, *indivisibilities* may preclude dispersion of specific activities, like large scale governance which is characteristic for capital cities or major cities. Second, *communication externalities* may lead to concentration of R&D activities and education in large cities. Finally, strong *complementarities*, or the importance of good matching between two sides of the market may require a thick market, which is supposedly the case for arts and movies industries in New York and Los Angeles (see Florida et al., 2012). Combining such post-industrial cities with industrial ones in a unified framework is our goal in future work.

2.6 Conclusion

We have revisited Anas' (2004) framework that predicts how, in response to growing population or decreasing trade costs, some cities can gradually decline and disappear. These are the "industrial" cities, which do not need externalities. To enhance the realism of the model, instead of Anas' normative approach, we consider migration and developers' equilibria. Still, *the vanishing effect proves to be robust*, because it rests on decreasing agglomeration force (home-market effect) and stable dispersion force (urban cost). Also, it turns out to be robust to the presence of other ("non-industrial") sectors in the economy, which are not affected by decreasing trade costs in the "industrial" sector. Additionally, we reveal details of comparative statics: how city size changes monotonically with the trade costs, the population of the world and preferences between the two sectors. In particular, goods with better transportation technology are more likely to be produced in smaller cities.

Generally, we interpret this vanishing effect as a realistic outcome a in post-industrial world, in which many industries indeed relocate to small towns or rural areas. Competing with other studies of "urban hierarchies" and city specialization, this model is one possible explanation for deindustrialization. To give an example, our study may shed some light on the coexistence of large metropolitan areas, such as Tokyo, and small specialized towns, such as Toyota city, the nature of industrial mix being an important determinant of city size. An extension of this approach would be a full-fledged general equilibrium model of qualitatively different sectors residing in different cities and their joint evolution.

Bibliography

Anas, A. (2004). Vanishing Cities: What does the new economic geography imply about the efficiency of urbanization? Journal of Economic Geography, 4: 181-199.

Behrens, K., Bougna, T. (2013). An anatomy of the geographical concentration of Canadian manufacturing industries. *CIRPEE discussion paper* 13-27.

Combes, P.-P., Mayer, T., Thisse, J.-F. (2008). *Economic Geography: The integration of regions and nations.* Princeton university press

Florida, R., Mellander, C., Stolarick, K. (2012). Geographies of scope: an empirical analisys of entertainment, 1970-2000. *Journal of Economic Geography*, 12: 183-204.

Forslid, R., Ottaviano, G. (2003). An analytically solvable core-periphery model. *Journal of Economic Geography*, 3(3): 229-240.

Fujita, M., Krugman, P., Mori, T. (1999). On the evolution of hierarchical urban systems. *European Economic Review*, 43: 209-251.

Fujita, M., Thisse, J.-F. (2013). *Economics of agglomeration: Cities, industrial location, and globalization.* Cambridge: Cambridge university press.

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

Holmes, T. J., Stevens, J. J. (2004). Spatial distribution of economic activities. In North America. In V. Henderson and J.-F. Thisse (eds.) *Handbook of regional and urban economics*, volume 4, 2797-2843. Amsterdam: North-Holland.

Kim, S. (1995). Expansion of markets and the geographic distribution of economic activities: The trends in U.S. regional manufacturing structure, 1860-1987. *The Quaterly Journal of Economics*, 110(4): 881-908.

Kolko, J. (2010). Urbanization, agglomeration and coagglomeration of Service Industries. In E. Glaeser (ed.) Agglomeration economics. Chicago: National Buro of Economic Research and The Chicago University Press, 151-180.

Ottaviano, G., Tabuchi, T., Thisse, J.-F. (2002) Agglomeration and trade revisited. *International Economic Review*, 43(2): 409-435.

Tabuchi, T. (1998). Urban agglomeration and dispersion: A synthesis of Alonso and Krugman. *Journal of Urban Economics*, 44: 333-351.

Tabuchi, T., Thisse, J.-F. (2006). Regional specialization, urban hierarchy, and commuting costs. *International Economic Review*, 47(4): 1295-1317.

Tabuchi, T., Thisse, J.-F. (2011). A new economic geography model of central places. *Journal of Urban Economics*, 69: 240-252.

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

2.A Proofs

Proof of Lemma 1. (1) Consider a city system perturbed by a new small city of size ε, $(n+1, N, N, ..., N-\varepsilon, \varepsilon)$. Since first $n-1$ cities are symmetric, we concentrate on trade equilibrium, which is symmetric for those cities and take labor in the first city as numeraire. This implies wage $w_i = 1$ for all these $i = \overline{1, n-1}$. Using pricing rule (2.7) that uses constant markups, this system has at most six distinct prices (domestic and export prices for normal city, the same for the disturbed city, and for new city):

$$p_{ii} = \frac{\sigma c}{\sigma - 1}, \; p_{i'i} = \frac{\sigma \tau c}{\sigma - 1}, \; p_{nn} = \frac{\sigma c w_n}{\sigma - 1},$$

$$p_{i'n} = \frac{\sigma \tau c w_n}{\sigma - 1}, \; p_{n+1,n+1} = \frac{\sigma c w_{n+1}}{\sigma - 1}, \quad p_{i',n+1} = \frac{\sigma \tau c w_{n+1}}{\sigma - 1}. \tag{2.19}$$

This system can be aggregated into price indexes, evaluated at point $\varepsilon = 0$, so that we distinguish only "big" cities from "new" ones (by symmetry, we obtain $w_n = 1$):

$$P_i = \left[\frac{N\theta(N)}{F\sigma} \left(\frac{\sigma c}{\sigma - 1} \right)^{1-\sigma} (1 + (n-1)\tau^{1-\sigma}) \right]^{1/(1-\sigma)} \tag{2.20}$$

$$P_{n+1} = \left[\frac{N\theta(N)}{F\sigma} \left(\frac{\sigma \tau c}{\sigma - 1} \right)^{1-\sigma} n \right]^{1/(1-\sigma)} \tag{2.21}$$

At trade equilibrium, the utilities also depend on wages. To find the wages, we recall constant firm size and use market clearing equations (2.9) for varieties produced in n big cities and in $(n+1)$-st city, which is small:

$$\frac{(1 + (n-1)\tau^{1-\sigma})\theta(N)N[\sigma c/(\sigma - 1)]^{-\sigma}}{P_i^{1-\sigma}} = (\sigma - 1)F/c \tag{2.22}$$

$$\frac{n\tau\theta(N)N[\sigma \tau c w_{n+1}/(\sigma - 1)]^{-\sigma}}{P_i^{1-\sigma}} = (\sigma - 1)F/c. \tag{2.23}$$

Taking a ratio of these two conditions, we obtain the equilibrium (shadow) wage in $(n+1)$-st city:

$$w_{n+1}^{-\sigma} = \frac{1 + (n-1)\tau^{1-\sigma}}{n\tau^{1-\sigma}} \tag{2.24}$$

Recall that $\theta(0) = 1$. Therefore, we can rewrite the comparison of utilities in small and big cities $V_{n+1} \leq V_n$ as a comparison of real incomes $w_{n+1}/P_{n+1} \leq \theta(N)/P_n$. Substituting the definition of $\theta(N)$, w_{n+1} and the ratio of price indexes from above we get result (2.10).

(2) Consider a city system perturbed (as in the definition of agglomeration stability) by small migration ε from the second city to the first, $(n,\, N+\varepsilon, N-\varepsilon, N, ..., N)$. We apply the same method. There are again at most six distinct prices and we can write the equilibrium equations for the price indexes and wages with labor in cities $i = \overline{3, n}$ being a numeraire good:

$$P_1 = \left[\frac{(N+\varepsilon)\theta(N+\varepsilon)}{F\sigma} \left(\frac{\sigma c w_1}{\sigma - 1} \right)^{1-\sigma} + \frac{(N-\varepsilon)\theta(N-\varepsilon)}{F\sigma} \left(\frac{\sigma \tau c w_2}{\sigma - 1} \right)^{1-\sigma} \right. \\ \left. + \frac{N\theta(N)}{F\sigma} \left(\frac{\sigma \tau c}{\sigma - 1} \right)^{1-\sigma} (n-2) \right]^{1/(1-\sigma)} \qquad (2.25)$$

$$P_2 = \left[\frac{(N+\varepsilon)\theta(N+\varepsilon)}{F\sigma} \left(\frac{\sigma \tau c w_1}{\sigma - 1} \right)^{1-\sigma} + \frac{(N-\varepsilon)\theta(N-\varepsilon)}{F\sigma} \left(\frac{\sigma c w_2}{\sigma - 1} \right)^{1-\sigma} \right. \\ \left. + \frac{N\theta(N)}{F\sigma} \left(\frac{\sigma \tau c}{\sigma - 1} \right)^{1-\sigma} (n-2) \right]^{1/(1-\sigma)}$$

$$P_i = \left[\frac{(N+\varepsilon)\theta(N+\varepsilon)}{F\sigma} \left(\frac{\sigma \tau c w_1}{\sigma - 1} \right)^{1-\sigma} + \frac{(N-\varepsilon)\theta(N-\varepsilon)}{F\sigma} \left(\frac{\sigma \tau c w_2}{\sigma - 1} \right)^{1-\sigma} \right. \\ \left. + \frac{N\theta(N)}{F\sigma} \left(\frac{\sigma c}{\sigma - 1} \right)^{1-\sigma} (1 + (n-3)\tau^{1-\sigma}) \right]^{1/(1-\sigma)}$$

$$\frac{(N+\varepsilon)\theta(N+\varepsilon) w_1 [\sigma c w_1/(\sigma-1)]^{-\sigma}}{P_1^{1-\sigma}} + \frac{\tau (N-\varepsilon)\theta(N-\varepsilon) w_2 [\sigma \tau c w_1/(\sigma-1)]^{-\sigma}}{P_2^{1-\sigma}} \\ + (n-2) \frac{\tau N \theta(N) [\sigma \tau c w_1/(\sigma-1)]^{-\sigma}}{P_i^{1-\sigma}} = (\sigma - 1)F/c \quad (2.26)$$

$$\frac{\tau(N+\varepsilon)\theta(N+\varepsilon)w_1[\sigma\tau c w_2/(\sigma-1)]^{-\sigma}}{P_1^{1-\sigma}} + \frac{(N-\varepsilon)\theta(N-\varepsilon)w_2[\sigma c w_2/(\sigma-1)]^{-\sigma}}{P_2^{1-\sigma}}$$

$$+(n-2)\frac{\tau N\theta(N)[\sigma\tau c w_2/(\sigma-1)]^{-\sigma}}{P_i^{1-\sigma}} = (\sigma-1)F/c$$

Differentiating this system of equation w.r.t. ε we aim to sign $\frac{dV_1}{d\varepsilon}$ at the symmetric point $\varepsilon = 0$.

Note that: (1) at the symmetric point $w_1 = w_2 = 1$; (2) by symmetry and definition of N_2 we have $\frac{dw_2}{d\varepsilon} = -\frac{dw_1}{d\varepsilon}$ and similar equality applies to price indexes; finally, $\frac{dP_i}{d\varepsilon} = 0$, i.e. for any third city ($i \neq 1, 2$) the effect of the population increase in city #1 is canceled out by the effect from exactly same decrease in city #2. Denote the elasticity of any variable X with respect to ε as $\mathcal{E}^X \equiv \frac{dX}{d\varepsilon}\frac{\varepsilon}{X}$ and totally differentiate equations (2.25) and (2.26) with respect to ε we obtain:

$$(1-\sigma)(1+(n-1)\tau^{1-\sigma})\mathcal{E}^{P_1} = (1-\tau^{1-\sigma})(1+\mathcal{E}^{\theta}+(1-\sigma)\mathcal{E}^{w_1}) \tag{2.27}$$

$$\sigma(1+(n-1)\tau^{1-\sigma})\mathcal{E}^{w_1} = (1-\tau^{1-\sigma})(1+\mathcal{E}^{\theta}+\mathcal{E}^{w_1}+(\sigma-1)\mathcal{E}^{P_1}) \tag{2.28}$$

We are interested in the elasticity of indirect utility, which can be expressed as $\mathcal{E}^{V_1} = \mathcal{E}^{\theta} + \mathcal{E}^{w_1} - \mathcal{E}^{P_1}$. The solution to the system of elasticity equations delivers:

$$\mathcal{E}^{w_1} - \mathcal{E}^{P_1} = (1+\mathcal{E}^{\theta})\frac{2\sigma-1}{(\sigma-1)\left(\sigma-1+\sigma\frac{1+(n-1)\tau^{1-\sigma}}{1-\tau^{1-\sigma}}\right)}$$

Therefore, the stability condition $\mathcal{E}^{V_1} \leq 0$ can be rewritten as

$$\frac{2\sigma-1}{(\sigma-1)\left(\sigma-1+\sigma\frac{1+(n-1)\tau^{1-\sigma}}{1-\tau^{1-\sigma}}\right)} \leq -\frac{\mathcal{E}^{\theta}}{1+\mathcal{E}^{\theta}}.$$

Recall that $\theta(N) = 1 - k\sqrt{N}$ and, hence, $\mathcal{E}^{\theta} = -\frac{k\sqrt{N}}{2(1-k\sqrt{N})}$. Substituting $\mathcal{E}^{\theta}$ into previous inequality delivers result (2.11). Q.E.D.

Proof of Proposition 5. (1) We start with the behavior of $L^d(n)$ when n goes to infinity. Brief inspection reveals that this limit is of type $\infty \times 0$ indeterminacy. However, we can apply l'Hospital's rule to rearrangements:

$$\lim_{n\to\infty}\left[k\sqrt{L^d(n)} = \frac{1-\left(1+\frac{1-\phi}{n\phi}\right)^{-\frac{2\sigma-1}{(\sigma-1)\sigma}}}{1/\sqrt{n}} = \frac{\frac{2\sigma-1}{\sigma(\sigma-1)}\left(1+\frac{1-\phi}{n\phi}\right)^{-\frac{2\sigma-1}{(\sigma-1)\sigma}-1}\left(-\frac{1-\phi}{n^2\phi}\right)}{-\frac{1}{2n^{3/2}}}\right],$$

and get

$$\lim_{n\to\infty}\left[k\sqrt{L^d(n)}\right] = 0.$$

Then, by continuity $\lim_{n\to\infty} L^d(n) = 0$. This fact can be interpreted as the existence of some $\bar{n}$ such that $\forall n \geq \bar{n} \quad L^d(n) \leq L^d(1) > 0$. By the extreme value theorem, $L^d(n)$ on interval $[1,\bar{n}]$ attains its maximum L^{d*} (which is finite) and, therefore, it is bounded by the value of this maximum on the whole interval $[1,+\infty)$. Then, for L greater than the universal critical population L^{d*}, the equilibrium (n, N) is unstable for all n.

(2) The proof of the second part is similar. First, applying l'Hospital's rule to the expression for $\phi^d(n)$, it is possible to show that $\lim_{n\to\infty}\phi^d(n) = 0$. Further, $\phi^d(n)$ is separated from one for any final n, and attains some maximum $\phi^{d*} < 1$ at some finite n, as in the previous argument. Thus, ϕ^{d*} is separated from 1, so that on the entire interval $[1,+\infty) \ni n$ any equilibrium $(n,\ L/n)$ is unstable against dispersion. Q.E.D.

Proof of Proposition 6. As we have shown in the proof of Proposition 5, critical $L^d(n)$ approaches zero at a speed $1/n$. Therefore, consider the following limit and apply to it l'Hospital's rule:

$$\lim_{n\to\infty}\left[k\sqrt{nL^d(n)} = \frac{1-\left(1+\frac{1-\phi}{n\phi}\right)^{-\frac{2\sigma-1}{(\sigma-1)\sigma}}}{1/n} = \frac{\frac{2\sigma-1}{\sigma(\sigma-1)}\left(1+\frac{1-\phi}{n\phi}\right)^{-\frac{2\sigma-1}{(\sigma-1)\sigma}-1}\left(-\frac{1-\phi}{n^2\phi}\right)}{-\frac{1}{n^2}}\right] = \frac{(2\sigma-1)(1-\phi)}{\sigma(\sigma-1)\phi}$$

Similarly,

$$\lim_{n\to\infty}\left[k\sqrt{nL^a(n)} = \frac{\frac{2}{\sigma+2+\frac{(\sigma-1)\sigma n\phi}{(2\sigma-1)(1-\phi)}}}{1/n} = \frac{2n}{\sigma+2+\frac{(\sigma-1)\sigma n\phi}{(2\sigma-1)(1-\phi)}}\right] = \frac{2(2\sigma-1)(1-\phi)}{\sigma(\sigma-1)\phi}$$

Combining these limits together and applying continuity we obtain $\lim_{n\to\infty} n(L^a(n) - L^d(n)) > 0$. This implies that $\bar{n}$ exists such that $\forall n \geq \bar{n} \quad L^a(n) > L^d(n)$, or equivalently, there is no L such that $L^a(n) \leq L \leq L^d(n)$, which is the necessary condition for migration stable equilibrium. Q.E.D.

Proof of Proposition 9. Let us perform comparative static exercise w.r.t. N_1. Taking elasticity of the equilibrium conditions, we find

$$(1-\sigma)(1+(n-1)\tau^{1-\sigma})\mathcal{E}_d^{P_n} = 1+\mathcal{E}^{\theta}+(1-\sigma)\mathcal{E}_d^{w_n} \tag{2.29}$$

$$\sigma(1+(n-1)\tau^{1-\sigma})\mathcal{E}_d^{w_n} = 1+\mathcal{E}^{\theta}+\mathcal{E}_d^{w_n}+(\sigma-1)\mathcal{E}_d^{P_n} \tag{2.30}$$

Here the subscript d emphasizes that elasticity is perceived by the developer. The elasticity equations for the stability against agglomeration (2.27) and (2.28) look very much alike, only with $(1-\phi)$ multiplier on the right hand side. Therefore, the costlier the trade, the more developer stability behavior resembles that of stability against agglomeration. The intuition is straightforward: cities affect each other through trade only. Therefore, the higher trade costs are, the less impact a developer's city has on other cities, and therefore, the smaller the developer's mistake in assuming no change in other cities will be. Solving these equations and evaluating the elasticity of indirect utility, we obtain:

$$\mathcal{E}_d^{V_1} = \frac{2\sigma-1}{(\sigma-1)(2\sigma-1+\sigma(n-1)\phi)}(1+\mathcal{E}^{\theta})+\mathcal{E}^{\theta} \tag{2.31}$$

Straightforward algebra yields the result. Q.E.D.

Proof of Proposition 10. We start with the case of a wise developer. First, observe that due to Bernoulli's inequality

$$L^d(n) = \frac{n}{k^2}\left[1-\left(1+\frac{1-\phi}{n\phi}\right)^{-\frac{2\sigma-1}{\sigma(\sigma-1)}}\right]^2 \leq \frac{n}{k^2}\left[\frac{(2\sigma-1)(1-\phi)}{\sigma(\sigma-1)n\phi}\right]^2 \tag{2.32}$$

Second, observe that $L^a(n)$ is a unimodal function with

$$arg\max_n L^a(n) = \bar{n} = \frac{(\sigma+2)(2\sigma-1)(1-\phi)}{\sigma(\sigma-1)\phi} \tag{2.33}$$

Thus, for $n > \bar{n}$ $L^a(n)$ is decreasing. Moreover, we now show that for such n $L^a(n) \geq L^d(n)$. Indeed,

$$L^a(n) = \frac{n}{k^2}\left[\frac{(2\sigma-1)(1-\phi)}{\sigma(\sigma-1)n\phi}\right]^2\left[\frac{2}{\frac{(\sigma+2)(2\sigma-1)(1-\phi)}{\sigma(\sigma-1)n\phi}+1}\right]^2 \geq \frac{n}{k^2}\left[\frac{(2\sigma-1)(1-\phi)}{\sigma(\sigma-1)n\phi}\right]^2 \geq L^d(n)$$

where the first inequality follows from $n > \bar{n}$ and definition of $\bar{n}$ (2.33) and the second inequality follows from (2.32). Recall that stability against dispersion requires $L(n) < L^d(n)$, therefore, wise developer stable equilibrium is possible only on the increasing part of $L^a(n)$. Thus, with population L growing, the number of cities n grows as well, and since $L^a(n)/n$ is a decreasing function, the city size N declines. Further, $L^a(n)$ decreases with the increase in freeness ϕ. Again, because the equilibrium must be on the increasing part of the curve, decreasing ϕ given L leads to an increase in the number of cities n. The city size N declines.

Proof for the myopic developer works along the same lines. Q.E.D.

Chapter 3

Hotelling Meets Chamberlin: Spatial Monopolistic Competition

Co-authored by Sergey Kokovin and Takatoshi Tabuchi.

3.1 Introduction

Motivation. Addressing consumers' heterogeneity in the markets for differentiated products, we observe that individual consumers typically favor different "ideal goods", e.g. favorite type of beer or coffee — and, further, often choose something different from time to time. In other words, each consumer's *love for variety* struggles with her *love for ideal product type.* This conflict results in a non-equal mixture of ideal and non-ideal varieties in the consumption bundle of an individual. Somewhat similarly, consumers in a city quite often buy food from the nearest shop but also use other shops from time to time. Such behavior generates an intersection of the shops' ranges of service. On a country-wide scale, we also observe intersecting trade areas of various firms, though closer clients are served more frequently. Overall, in many real markets, the partially-localized preferences of consumers give rise to *partially-localized competition.*

Equally important in this respect is the question of the market structure: Why do some seemingly similar markets show very different degree of product differentiation, e.g. why is more than half of the beer market in the US covered by only three brands, whereas no one brand has even ten percent of the beer market in the UK. It is important to understand which features of the consumer's partially-localized preferences or market geography may account for such different outcomes.

These questions are not exactly new in economic theory. Almost simultaneously, Hotelling (1929) and Chamberlin (1933) introduced two competing ideas: consumers' ideal points and their love for variety. The subsequent Hotelling-style tradition of modeling spatial markets maintains the homogenous good assumption (see Lancaster (1966), Salop (1979), Vogel (2008)). In this case, each consumer is served by a single firm, i.e, *ranges of service do not intersect* because each firm competes only with its adjacent neighbor for the borderline consumer. By contrast, Chamberlinian tradition (which became the mainstream market concept after Dixit-Stiglitz (1977) and has taken a central place in new theories of trade, geography, and growth since then) assumes a "horizontally" differentiated good without space. This makes the *ranges of service completely coinciding*: every firm competes equally with every other firm. In our view, a representative consumer's "love for variety" concept in usual monopolistic competition theory remains subject to doubts and objections. We believe that the real life love for variety stems *mainly* from consumer heterogeneity. We would like to make this intuition explicit, akin to the theory of product differentiation under discreet choice (see Anderson et al. (1992) for a review).

Another question is robustness. Given that the market structure we have in mind appears to be different to those widely studied, would standard theoretical conclusions remain or change drastically after the introduction of heterogeneous consumers? Does market equilibrium behave the way we are used to in response to a change in endogenous variables such as the market size?

Aiming to answer these questions, we introduce a model that fills the gap between two polar views on competition. The present paper bridges the (free-entry version of) Hotelling (1929) and Chamberlin (1933) approaches in a simple but general way. It includes both these polar special cases, differing in essence by two parameters: love for variety (absent in Hotelling) and distance cost (absent in Chamberlin). Our construction aims to keep all features of the mainstream monopolistic competition theory present but to replace the representative consumer with heterogenous consumers. Introducing the simplest rich model of this kind, we compare it with other models bridging space and love for variety, and point out various important extensions for subsequent work. At the center of this paper is the question of how a spatial dimension (consumer heterogeneity) changes the nature and comparative statics of market competition.

Setting. Consumers are continuously distributed with some given density along a circumference, similar to the Salop (1979) model. This space can represent a geographical space, or a space of consumers' tastes, i.e., ideal points among varieties of the differentiated good. In contrast to inelastic demand in Salop and discrete-choice papers, here each consumer combines various quantities of ideal and non-deal varieties in her consumption bundle, due to love for variety. Consumers are identical in preferences, have the same (unspecified additive) utility function and income, but differ in their locations. Naturally, everybody prefers varieties (firms) located closer than those farther away and buys more of a variety located closer to her ideal point. This feature is described by some "cost of distance" linearly introduced into two versions: either as monetary cost to transport the good or as disutility of distance to ideal point. The former version is better suited for geographical interpretation of the model (and also for intermediate production goods), whereas disutility of distance has more bite in a setting with product characteristics space of consumer goods.

As to (the continuum of) firms, their number and location, unlike in Hotelling, is not given but endogenous. Following Chamberlin (1933) and Dixit-Stigliz (1977), our market exhibits free entry and increasing returns in producing a differentiated good. Homogeneous firms simultaneously choose their prices and their locations, taking as given the density of consumers and current local intensity of competition everywhere. Gross demand of a firm is the aggregate of the demands of all consumers within its range of service (where distance costs allow for positive demand). Market equilibrium in the general version of the model consists of three curves in the consumer space: (i) the density distribution of firms, (ii) their prices, (iii) competition intensity (marginal utility of money). The volumes of individual demand for all firm-consumer pairs and ranges of service can be derived from these variables. However, in the basic version of the model with uniform density of consumers and symmetric (uniform) distribution of firms, such an equilibrium boils down to three scalars: mass of firms, price, and competition intensity. Actually, when the circular space of consumers shrinks to a point or distance cost shrinks to zero, at the limit one arrives at the standard model of monopolistic competition. Otherwise, the new model enables richer predictions, especially an endogenous *range of service*, at least under those utilities which have a finite derivative at zero (choke-price); for instance linear-quadratic utility. Without a choke-price, e.g. under constant elasticity of substitution (CES) preferences, the whole consumer space is always served by every firm.

Results. First, we analyze the basic setting, trying to reveal similar effects of competition as in usual monopolistic competition theory (see Zhelobodko et al., 2012, henceforth ZKPT). Should a growing population make firms more numerous and larger, and simultaneously push their prices down? The technical achievement in this direction is convenient reformulation of the aggregated demand faced by a firm into "consumer surplus" of elementary utility at maximal local consumption. Then, the uniform spatial model turns out *equally simple* and tractable to usual monopolistic competition. This news may inspire theorists wishing to expand the usual analysis of international trade to the influence on consumer tastes and firms core competencies.

For both versions of our model (monetary cost and disutility of distance) we obtain general propositions of comparative statics. Namely, under incomplete coverage of space by a firm's service (choke-price), population growth (e.g., opening trade) always leads to more numerous firms, higher competition and smaller individual consumption of each variety, but price behavior and firm size both depend on the elasticity of elementary utility. Under the natural DEU (decreasingly-elastic utility) condition, prices go down, otherwise the opposite outcome takes place. This outcome reminds us of the necessary and sufficient condition from ZKPT for "pro-competitive effect" of the market size or trade, which is increasing (in the absolute value) elasticity of the inverse demand function (IED). However, now IED is replaced by the DEU condition, which appears to be a direct application of ZKPT to the integrated demand, because of the crucial simplification described. These effects are similar, which generally supports the robustness of monopolistic competition modeling: spatial generalization does not destroy it.

Moreover, although the two conditions do not seem to be directly related, intuitive interpretations of them are strikingly similar: expanding market size generates pro-competitive effects when demand is not too convex.

However, instead of similarity, a theorist should be more interested in the direction in which spatial monopolistic competition *differs* from usual modeling. How does the heterogeneity of integrated demands change the shape of gross demand function? As we know, generally, heterogeneity combined with integration makes the demand more convex (see such a conclusion for income heterogeneity in Osharin et al., 2014). Indeed, consider a simple example of quadratic elementary utility $u(q) = q - 0.5q^2$ which generates linear demand function $q = 1 - P$ of total price $P = p + t$ which depends on cost-of-distance coefficient t. One may check that the range of service will be $(1-p)/t$ and the gross demand of a firm is quadratic: $Q = 0.5(1-p)^2/t$, more convex than initial linear demand.

It is the elasticity and convexity of the *gross* demand that govern the behavior of firms and generates many market effects. So, turning from a usual consumer-homogeneous monopolistic competition model to a spatial one may essentially change some theoretical predictions. Upturning this idea, we may say that when empirical estimates of gross demand are found to be a quadratic function in the form $Q = 0.5(1-p)^2$, then under the assumption of spatial monopolistic competition, the elementary utility revealed must also be quadratic, not cubic, as one would conclude assuming spaceless competition. A similar conclusion applies to numerous econometric estimates of CES demand: the magnitude of the underlying elementary elasticity of substitution σ between varieties must be different from one reported under the assumption of spaceless competition. Furthermore, in some cases, introducing space and the related demand convexification may destroy the usual assumption of concave profit. In this case multiple equilibria arise, as our preliminary inquiry in this direction suggests.

Second, we analyze the case with complete coverage of space by service. It is necessary not only to include the CES case into the study, but also to be able to show that usual monopolistic competition is really a limiting case of a spatial one when distance cost fades away. This case is substantially less tractable. However, we show that, as expected, as transportation costs diminish to zero, the model converges to the standard spaceless monopolistic competition model of ZKPT. We also show that the general conclusion persists: the market is pro-competitive when demand is not too convex. However, unfortunately in this case we do not have an exact border line case between two market modes. It remains an empirical question in which mode — partial coverage or complete coverage — each particular market operates.

3.2 The Uniform Model

In this section we set up a simplified version of our model. We assume that consumers are distributed uniformly over a circumference of unit length. A point of the circumference can be viewed as a geographic location or a specific product in the product characteristic space. Firms are free to choose any point of the circumference to enter. For now we constrain our attention to the case where firms are also distributed uniformly over the circumference. In what follows, we label it a uniform equilibrium. The concept of uniform equilibria may be criticized, because they need not be stable and because non-uniform consumer distribution is unlikely to give rise to uniform distribution of firms. However, without this basic model, more complicated equilibria are difficult to comprehend. Moreover, to support an approximately-uniform, or at least continuous distribution of firms, we introduce, in a reduced form, an external *dispersion force* that pushes one firm away from another. It represents land prices and other congestion forces common to economic geography but not modeled here explicitly. In this case, the tendency towards dispersion of firms looks more plausible.

In what follows we consider two versions of spatial models: (i) monetary cost of distance and (ii) disutility of distance. In the first version, the adjustment cost for consuming products produced further away from the consumer's location enters the budget constraint. This formulation is common in economic geography; it is also fits well to the case when our "consumer" is actually a firm that consumes some intermediate good, incurring costs for adjusting the good to fit its exact needs. The second version has more bite in the case of consumption goods; here "distance" from one's favorite variety has some disutility value. We now turn to the formal description of our model.

3.2.1 The Model Setup

Consumers and varieties. We assume identical consumers, each possessing one unit of a numeraire good (for instance, labor). As in Hotelling (1929), any consumer type is characterized by her bliss point x in some space Ω, i.e., her beloved variety of the differentiated good. The types are uniformly distributed with density L along the circular space of product characteristics, the circumference $\Omega = [-1/2, 1/2]$ of length 1 (Salop's "race-track economy" is a proxy for "long" linear interval). Following the Chamberlinian tradition, each variety is produced by a single firm and each firm produces single-product. There is a continuum of firms. A firm's type, denoted $y \in [-1/2, 1/2]$ refers to its location on the circumference, i.e., its targeted type of consumers, whereas (endogenous) density μ_y is the measure of firms in the same location. As has been stated, in this section the density $\mu_y \equiv \mu > 0$ is assumed to be constant at each point $y \in \Omega$. In addition, we assume mill pricing by the firms, i.e. a firm at y charges gate price p_y for its product. Again, with our focus on the symmetry, price distribution is also uniform with $p_y \equiv p > 0$.

Ranges of service of various firms do intersect with each other, because consumers love variety. However, they love various varieties unequally. The bliss-point variety is slightly preferred to other varieties. For instance, one can imagine a consumer occasionally using many restaurants in her city but preferring not to go too far. More generally, either adjusting the non-ideal variety to consumer's tastes is costly, or carrying a purchase home from a remote shop is costly. Specifically, we assume adjustment costs $q \cdot \tau(\theta)$ for buying q and carrying it home from distance θ, where $\tau(\cdot)$ is an increasing function of distance, which for the simplicity of exposition we assume to be linear. Hence, remote varieties will be consumed in smaller amounts. In particular, extremely remote varieties may become ignored, not consumed. Therefore, in equilibrium each consumer x has an (endogenous) range of varieties (firm types) that she wishes to buy, $\hat{\theta} \in (0, 1/2]$ denotes the length of the range of service, also constant for every consumer. An equilibrium may result in a small range $\hat{\theta} < 1/2$ which means “incomplete coverage” of the circumference Ω by each firm's service. Another possibility is “complete coverage by service” $\hat{\theta} = 1/2$, in the case when the cost of distance is small enough to buy products (in different quantities) from all firms.

Now we can formulate the consumer problem. Given the (uniform) price distribution p and firm distribution μ, the consumer seeks to maximize her utility subject to budget constraint:

Monetary cost

$$\max_{q_{xy}>0} \ \mu \int_\Omega u(q_{xy}) dy$$
$$\text{s.t. } \mu \int_\Omega (p + \tau(x,y)) q_{xy} dy = 1$$

Disutility of distance

$$\max_{q_{xy}>0} \ \mu \int_\Omega u(q_{xy}) - q_{xy}\tau(x,y) dy$$
$$\text{s.t. } \mu \int_\Omega p q_{xy} dy = 1$$

A consumer at x buying quantity q_{xy} from a firm at y receives direct consumption utility $u(q_{xy})$ in both versions of the model. The elementary utility function $u(\cdot)$ is assumed to be increasing, thrice differentiable and concave, thus, generating love for variety. In addition, $u(0) = 0$, i.e. the presence of a variety does not generate any utility if the consumer does not consume it. This normalization allows for neat representation of the comparative statics results. Total utility is additive in elementary utilities over the whole range of varieties. This unspecified additive utility will allow us to relate comparative statics market effects to features of preferences and to contrast the results with the current literature on non-spatial monopolistic competition (see ZKPT). Adjustment cost function $\tau(x, y)$ depends on the distance between x and y and represents either the monetary or utility cost per unit of consumption of worse than ideal variety, by assumption $\tau(x, x) = 0$. For simplicity, we assume the transportation cost to be linear in distance. Given that our space is circumference, it implies

$$\tau(x,y) = t \min\left\{|x-y|, 1-|x-y|\right\}.$$

In the first formulation, the costs enter budget constraint, therefore, we refer to this version as the monetary cost model. In the second version, these costs enter utility directly, and we refer to that version as the disutility of distance model.

Producers. The solution to the consumer problem gives rise to the location specific demand functions $d_{xy}(p_y, p, \mu)$, i.e. given the price of variety y and collection of other prices p and density of firms μ how much product a consumer at x buys from a firm at y. Each producer takes the demand functions and number of competitors as given and prices her variety to maximize profit. As is standard in the monopolistic competition literature, we assume that the producer incurs constant marginal cost m of production and fixed cost F to operate on the market. In addition, we introduce a dispersion force in a reduced form into the model. We assume that fixed cost $F = F_y = F(\mu_y) = F(\mu)$ is a non-decreasing function of the density of firms at y.

In most of our analysis we treat the fixed cost component as constant (independent of the number of firms), assuming that dependency is weak enough not to affect comparative statics results. Nevertheless, we introduce a dispersion force for two reasons. First, the dispersion forces are conceptually important, especially in the case relevant to economic geography. The concentration of the activity in a particular point raises the price of land and increases congestion costs. Second, on theoretical grounds, the presence of a dispersion force counters the potential instability of continuous uniform equilibria. The latter consideration will be especially relevant in future development of the model, since preliminary results show that continuous equilibria are not always stable.

Formally, a producer at y chooses the price to solve the following profit-maximization problem:

$$\max_{p_y \geq 0} \pi(p_y, p, \mu) = \max_{p_y \geq 0} (p_y - m) \int_{\Omega} d_{xy}(p_y, p, \mu) dx - F(\mu)$$

Equilibrium. Entry into the market is free. Because of the entry, profits must vanish at each location:

$$\pi(p_y, p, \mu) = 0.$$

Symmetric equilibrium is a bundle $\{p, \mu, d_{xy}(p_y, p, \mu)\}$ of price, density of firms and location-specific demand functions, that satisfy all consumer and producer optimization conditions, and the free-entry condition. This general definition is valid for both versions of the model. In the following analysis of each of the setups, we refine the equilibrium definition accordingly to simplify exposition in each case.

3.2.2 Monetary Transportation Costs

We begin our analysis of features of the spatial monopolistic competition with the monetary transportation cost version of our model. First, we derive the demand function. Standardly, consumer optimization implies:

$$q_{xy}(u'(q_{xy}) - \lambda_x(p + \tau(x, y))) = 0.$$

Here, λ_x is the Lagrange multiplier of the budget constraint of a consumer at x, which can be interpreted as the intensity of competition for this consumer. Due to symmetry, $\lambda_x = \lambda$ for all x. From the complimentary slackness of the consumer's optimality condition it can be seen that if $u'(0)$ is small enough $q_{xy} = 0$ for some y sufficiently far from x. On the other hand, it might be the case that $q_{xy} > 0$ for every pair x and y (indeed, it has to be the case, if $u'(0) = \infty$). We refer to the former as partial coverage, because firms do not serve each and every consumer, and to the latter as full coverage, by analogous reasoning. As we show later, this distinction is quite important because the comparative statics of the equilibrium differs substantially between these two cases. Essentially, the length of coverage can be found as:

$$\hat{\theta}(p) = \max\left\{\frac{1}{t}\left(\frac{u'(0)}{\lambda} - p\right), 1/2\right\}$$

Partial Coverage

We start with the analysis of the case of partial coverage of consumers by firms. Because we consider uniform equilibria when firms are identical up to rotation we can focus on the firm at $y = 0$. Let the elementary demand function $D(p) = u'^{-1}(p)$ whenever the inverse of marginal utility exists and zero otherwise. With this notation $q_{xy} = D(\lambda_x(p + \tau(x, y))$, and the firm's profit can be written as:

$$\Pi(p, \lambda) = 2(p - m)L \int_0^{\hat{\theta}(p)} D(\lambda p + \lambda\tau(\theta, 0))d\theta - F(\mu)$$

When maximizing the profit, producers take intensity of competition λ as given. Here we use the variable $\theta = |x - y|$ of consumer-producer distance, i.e., distance of any consumer-type $\theta \in [0, \hat{\theta}]$ from a firm located at 0. Aggregate quantity sold by the firm is the sum of quantities sold to consumers between $-\hat{\theta}(p)$ and $\hat{\theta}(p)$. Density L of consumers at each location factorizes the total output of the firm dedicated to all consumers everywhere (L is also total population).

An integral of (inverse) derivative can be simplified. Namely, for the case of linear cost function $\tau(\theta) = t\theta$ we consider D (whose argument runs from minimal "price" λp to maximal "price" $\lambda p + \lambda t\hat{\theta}$), and argue that integrating D is the same as integrating its inverse u' whose argument runs from 0 to maximum value $q_0 = D(\lambda p)$, which is the maximal purchase occurring near the bliss-point. Intuitively, instead of integrating quantities over the consumers in the space, we integrate them over the price range. Technically, this amounts to substitution of variables: $q = D(\lambda p + \lambda t\theta)$, or a change of the axis of integration in the price-quantity space. Thus, any firm's gross output Q can be represented as

$$Q = 2L\int_0^{\hat{\theta}(p)} D(\lambda p + \lambda t\theta)d\theta = \frac{2L}{\lambda t}\int\limits_{D(\lambda p)}^{0} qd(D^{-1}(q) - \lambda p) = -\frac{2L}{\lambda t}\int\limits_{0}^{D(\lambda p)} zdu'(z) =$$

$$= \frac{2L}{\lambda t}\left[-D(\lambda p)u'(D(\lambda p)) + \int\limits_{0}^{D(\lambda p)} u'(z)dz\right] = \frac{2L}{\lambda t}[u\left(D(\lambda p)\right) - \lambda pD(\lambda p)]$$

which is similar to "consumer surplus" in spaceless IO models and decreases in p. In fact, it is the surplus of a consumer located exactly at the firm's location. It must be noted that this simplified structure of aggregate demand relies on the assumption of linear transportation costs.

Thus, under linear distance cost $\tau(\theta) = t\theta$ and uniform equilibrium, any producer's profit can be rewritten without an integral, simply as

$$\Pi(p, \lambda) = (p - m)\frac{2L}{\lambda t}[u(D(\lambda p)) - \lambda pD(\lambda p)] - F(\mu). \tag{3.1}$$

Differentiating our profit (3.1) w.r.t. p we arrive at the FOC:

$$\Pi_p = \frac{2L}{\lambda t}\left[u(D(\lambda p)) - \lambda pD(\lambda p) - (p - m)\lambda D(\lambda p)\right] = 0 \tag{3.2}$$

Furthermore, differentiating the previous expression (3.2) we get the producer's second order condition for profit maximization:

$$\Pi_{pp} = \frac{2L}{\lambda t}\left[-(p - m)\lambda D'(\lambda p) - 2D(\lambda p)\right] < 0$$

which we assume to hold strictly in equilibrium. Thus, producer's optimality condition $\Pi_p(p, \lambda) = 0$ together with the free entry condition $\Pi(p, \lambda) = 0$ determine equilibrium pair of price and competition intensity (p, λ). From them, other equilibrium quantities of interest, i.e. consumption, density of firms and range of service can be obtained via the consumer's optimal choice and budget constraint discussed above.

Comparative statics. We have characterized equilibrium in the case of partial coverage. Now we turn to the question of interest: how does the equilibrium react to changes in market size? In particular, whether increasing market size or decreasing transportation costs and the associated increase in competition leads to lower prices. In what follows, we find it helpful to define the additional equilibrium variable $q_0 = D(\lambda p)$, which represents the consumption of an ideal variety. The next proposition establishes that the elasticity of the elementary utility function $\varepsilon_u(q) = \frac{qu'(q)}{u(q)}$ plays a defining role in the comparative statics behavior: decreasing elasticity of utility leads to pro-competitive effects, whereas increasing $\varepsilon_u(q)$ leads to anti-competitive effects of the increasing market size.

Before proceeding to the proposition, we would like to address a concern that $\varepsilon_u(q)$, as defined, is not immune to affine transformations of the elementary utility function. Given our assumption of separable additive aggregate utility, an affine transformation of $u(\cdot)$ must not change the equilibrium outcome. However, our normalization assumption is $u(0) = 0$, which we used in the derivation of the aggregate demand for the firm's product. Without it, the result of our comparative statics analysis would depend on the more cumbersome $\tilde{\varepsilon_u}(q) = \frac{qu'(q)}{u(q)-u(0)}$. To ease the notation, we stick to our normalization, and elasticity of utility as the quantity of interest.

Proposition 11. *Consider the version of the model with monetary costs of transportation and with partial market coverage. Then an increase in the market size L, or a decrease in the transportation cost t, leads to: (i) an increase in the intensity of competition λ; (ii) a decrease in purchases of the ideal variety q_0; (iii) a decrease (an increase) in the prices whenever $\varepsilon_u(q)$ is a decreasing (increasing) function. In addition, expanding market size L leads to increasing μ, i.e. more firms entering the market, and decreasing $\hat{\theta}$, i.e. the competition being more localized.*

Proof. First, observe that L/t enters firm's equilibrium conditions only as a ratio, therefore, the results for the transportation cost follow immediately from the result for the market size. Thus, we focus only the market size effects. Now we rewrite the firm's first order and zero profit conditions in (p, q_0) instead of (p, λ) variables, using the fact that $\lambda = \frac{u'(q_0)}{p}$. The firm's first order condition becomes:

$$u(q_0) - u'(q_0)q_0 - (p-m)\frac{u'(q_0)q_0}{p} = 0,$$

and zero profit condition:

$$(p-m)\frac{2pL}{u'(q_0)t}[u(q_0) - u'(q_0)q_0] = F(\mu).$$

Denote $\varepsilon_u(q_0) = \frac{u'(q_0)q_0}{u(q_0)}$ the elasticity of the utility function at q_0. With this standard notation and some algebra, the equilibrium conditions become:

$$\frac{1}{\varepsilon_u(q_0)} = 2 - \frac{m}{p} \qquad (p-m)^2\frac{2L}{t}q_0 = F(\mu).$$

Observe that from the first equation it follows that in equilibrium $1/2 < \varepsilon_u(q_0) < 1$. We now totally differentiate both equations, obtaining:

$$-\frac{\varepsilon_u'(q_0)}{\varepsilon_u^2(q_0)}\frac{dq_0}{dL} = \frac{m}{p^2}\frac{dp}{dL}$$

and

$$2(p-m)Lq_0\frac{dp}{dL} + (p-m)^2L\frac{dq_0}{dL} + (p-m)^2q_0 = 0.$$

Observe that from the first equation it follows that price p and quantity q_0 co-move when the elasticity of utility is decreasing, and move in the opposite direction when the elasticity of utility is increasing. Combining them, we get:

$$\left[p - m - 2q_0\frac{\varepsilon_u'(q_0)p^2}{\varepsilon_u^2(q_0)m}\right]\frac{dq_0}{dL} = -\frac{(p-m)q_0}{L} \tag{3.3}$$

Using the fact that $\varepsilon_u'(q_0) = \left(\frac{q_0u'(q_0)}{u(q_0)}\right)' = \frac{u'(q_0)u(q_0)+q_0u''(q_0)u(q_0)-q_0u'^2(q_0)}{u^2(q_0)}$ we can rewrite the expression in the square brackets.

$$p - m - 2q_0\frac{\varepsilon_u'(q_0)p^2}{\varepsilon_u^2(q_0)m} = p - m - 2q_0\frac{u'(q_0)u(q_0) + q_0u''(q_0)u(q_0) - q_0u'^2(q_0)}{q_0^2u'^2(q_0)}\frac{p^2}{m} =$$

$$= p - m - 2q_0\left(\frac{1}{q_0\varepsilon(q_0)} + \frac{u''(q_0)}{u'(q_0)}\frac{1}{\varepsilon(q_0)} - \frac{1}{q_0}\right)\frac{p^2}{m} =$$

$$= p - m - 2\left(\frac{1}{\varepsilon(q_0)} - 1\right)\frac{p^2}{m} - 2q_0\frac{u''(q_0)}{u'(q_0)}\frac{1}{\varepsilon(q_0)}\frac{p^2}{m} =$$

$$= p - m - 2\left(1 - \frac{m}{p}\right)\frac{p^2}{m} - 2q_0\frac{u''(q_0)}{u'(q_0)}\frac{1}{\varepsilon(q_0)}\frac{p^2}{m} =$$

$$= (p-m)(1 - 2\frac{p}{m}) - 2q_0\frac{u''(q_0)}{u'(q_0)}\frac{1}{\varepsilon(q_0)}\frac{p^2}{m} = (p-m)\frac{-1}{\varepsilon_u(q_0)}\frac{p}{m} - 2q_0\frac{u''(q_0)}{u'(q_0)}\frac{1}{\varepsilon(q_0)}\frac{p^2}{m} =$$

$$= -\frac{1}{\varepsilon_u(q_0)}\frac{p^2}{m}\frac{u''(q_0)}{u'(q_0)}\left[\frac{p-m}{p}\frac{u'(q_0)}{u''(q_0)} + 2q_0\right]$$

The term in front of the bracket is clearly positive. We can now use the firm's second order condition, which in our variables can be expressed as $\frac{p-m}{p}\frac{u'(q_0)}{u''(q_0)} + 2q_0 > 0$, which is exactly the bracketed term. Thus, the bracketed term in (3.3) is positive. This implies that $\frac{dq_0}{dL} < 0$, i.e. consumption of the ideal variety always decreases with the market size. The result for the price behavior follows from the discussion above.

To understand the behavior of the intensity of competition λ notice that $\Pi(p, \lambda) = 0$ together with $\Pi_p = 0$ imply that

$$\frac{d\lambda}{dL} = -\frac{\Pi_L}{\Pi_\lambda} = -\frac{F(\mu)/L}{-\frac{1}{\lambda}F(\mu) + (p-m)\frac{2L}{\lambda t}[-pD(\lambda p)]} > 0$$

Therefore, the intensity of competition increases with the market size regardless of the nature of preferences. In addition,

$$\frac{d\lambda}{dt} = -\frac{\Pi_t}{\Pi_\lambda} = -\frac{-F(\mu)/t}{-\frac{1}{\lambda}F(\mu) + (p-m)\frac{2L}{\lambda t}[-pD(\lambda p)]} < 0$$

and,

$$\varepsilon_t^\lambda = -\frac{t}{\lambda}\frac{d\lambda}{dt} = \frac{F(\mu)}{F(\mu) + (p-m)\frac{2L}{t}[pD(\lambda p)]} = 1 - \varepsilon_u(q_0) < 1/2$$

Thus, the intensity of competition increases when transportation costs decrease. However, it does not increase too fast: $\varepsilon_t^\lambda < 1/2$ implies that both λt and $\lambda^2 t$ decrease when transportation costs decrease.

Now, we focus on the cases we find plausible, i.e. on the decreasing or slowly increasing elasticity of utility. In these cases q_0 decreases, thus, $\lambda p = u'(q_0)$ increases. Therefore, the radius of service $\hat{\theta} = \frac{u_0' - \lambda p}{\lambda t}$ decreases as market size increases. The case of decreasing transportation costs is less clear. On the one hand, equilibrium forces, as before, push the radius of service down, yet at the same time, the mechanical effect of cheaper transportation leads to the expansion of coverage. The direction of the aggregate affect, which is the sum of these two (direct and equilibrium) effects, is unclear. The last equilibrium variable of interest is the number of firms μ. To understand its behavior, we return to the consumer's budget constraint:

$$\frac{1}{2\mu}=\int_0^{\hat{\theta}}(p+t\theta)D(\lambda p+\lambda t\theta)d\theta=\frac{1}{\lambda t}\int_{D(\lambda p)}^{0}q\frac{u'(q)}{\lambda}du'(q)=\frac{1}{2t\lambda^2}\int_0^{D(\lambda p)}qd(-u'(q)^2)$$

The integrand does not depend on any equilibrium variables, and at the same time the upper limit of integration q_0 decreases with the market size. Therefore, the entire integral decreases. In addition, the intensity of competition increases, thus, as expected, the expanding market size leads to more entry and an increase in the density of firms. However, as shown before, the intensity of competition does not increase sufficiently fast with the decrease in the transportation cost, and $\lambda^2 t$ is decreasing. This generates an ambiguous effect on the density of firms when transportation costs decrease. **Q.E.D.**

Thus, we have classified markets according to the ε_u into two categories: those with DEU react to relative market size pro-competitively (decreasing prices under higher competition), and those with IEU behave anti-competitively. The open question is what case is more realistic? For instance, under the widely used linear demand, and CARA and HARA utility functions, elasticity of utility is decreasing. All these preferences generate similar pro-competitive effects in usual spaceless monopolistic competition as well but for *a different reason*: they generate increasingly-elastic demand (IED); see ZKPT. In principle, a combination of properties IED+DEU of demand/utility is widely used and considered natural but not guaranteed.

These comparative static results generally look intuitive. Indeed, more dense consumer population should entoce more firms to each location. This shift increases local competition and pushes consumption of each individual variety down, because more varieties become closer and available to the consumer. As a consequence, one would expect decreasing prices. Indeed, this is really the case under the natural and widely used implicitly DEU assumption. Thus, here increasing or decreasing elasticity of *utility* governs prices, unlike increasing or decreasing *demand* elasticity in ZKPT.

The difference stems from the fact that now gross demand is the aggregation of the local demands u'^{-1} of various consumers (different in distance from the producer). Integrating u'^{-1} can be looked upon as integrating u', which is why maximizing profit reminds us of maximizing utility u. Put differently, what is important for price behavior is the elasticity of the aggregate demand a firm faces. At the same time, the aggregation of heterogeneous demands does not directly inherit properties of individual demands. In other respects, general explanation of price behavior is the same: a sufficiently flat gross demand curves generate natural effects, more convex demands enable paradoxical price behavior in response to growing competition.

Complete coverage

In the case of complete coverage, every firm sells its product to every consumer but, unlike the ZKPT spaceless model, in different quantities. This case is more difficult to analyze because it is a mixture of two very different market operating modes. Intuitively, assume first that the transportation cost is very small. Then, the model converges to the spaceless case, and the comparative statics is governed by the behavior of the elasticity of individual demands, i.e. marginal utility, as in the ZKPT model. On the other hand, if the space is just covered, i.e. consumption of the most remote varieties is very small, the model is basically the same as the model with partial coverage, and comparative statics is governed by the elasticity of the elementary utility function. Therefore, the comparative statics in any case in between these will depend on both elasticity of utility and elasticity of marginal utility. For this reason, here we provide only partial characterization of this case, focusing on the most popular and arguably natural case: that of not very convex demand.

To study this case, in addition to the quantity of an ideal variety $q_0 = D(\lambda p)$ we introduce the quantity of the least preferred variety $q_1 \equiv D(\lambda p + \lambda t 0.5)$. This allows us to express a firm's gross demand Q in a similar fashion as before, i.e. as a combination of consumer surpluses:

$$\begin{aligned} Q(p,\lambda) = 2L\int_0^{0.5} D(\lambda p + \lambda t\theta)d\theta \ &= \ \frac{2L}{\lambda t}\int_{q_0}^{q_1} q du'(q) = \\ = \frac{2L}{\lambda t}\left[q_1 u'(q_1) - q_0 u'(q_0) - \int_{q_0}^{q_1} u'(q)dq\right] \ &= \ \frac{2L}{\lambda t}[u(q_0) - \lambda p q_0 - u(q_1) + (\lambda p + \lambda t 0.5)q_1] \end{aligned}$$

In words, the total demand is proportional to the difference in consumer surpluses between the closest and the furthest consumer. Again, this relatively straightforward representation relies on the linear distance cost. Although this assumption is very restrictive, it is not uncommon in the literature. Now, as in the case of partial coverage, firms' optimal behavior and free entry condition define the equilibrium in (p, λ) variables. All other equilibrium quantities of interest can be recovered from them. Indeed, the free entry condition requires that in equilibrium:

$$\Pi(p, \lambda) = (p - m)Q(p, \lambda) - F(\mu) = 0$$

and firm's optimal behavior is characterized by the first order condition:

$$\Pi_p = Q(p, \lambda) + (p - m)\frac{2L}{\lambda t}[-\lambda D(\lambda p) + \lambda D(\lambda p + \lambda t 0.5)] = 0$$

Similarly to the case studied before, consumption of varieties q_0 and q_1 and density of firms μ can be derived from (p, λ) using a consumer's optimality condition and budget constraint. We now formulate our comparative statics result. As we have mentioned, the analytical complexity of the case precludes complete characterization of the comparative statics effects. Nevertheless, we show that in the most popular cases, when individual demand is relatively flat, the market behaves pro-competitively.

Proposition 12. *In the version of the model with monetary cost of transportation and with complete market coverage, let the elasticity of marginal utility* $\varepsilon_{u'} = -\frac{qu''(q)}{u'(q)} < 1$*. Then an increase in the market size* L *leads to an increase in the intensity of competition* λ*. Moreover, if the demand is not very convex, i.e. if the ratio* $-u''(q)/u'(q)$ *is an increasing function, then increasing market size leads to decreasing prices* p*, and the market is pro-competitive.*

Proof. First, totaly differentiating free entry condition and using the fact that $\Pi_p = 0$ in equilibrium, we get:

$$\frac{d\lambda}{dL} = -\frac{\Pi_L}{\Pi_\lambda} = \frac{F(\mu)/L}{F(\mu)/\lambda + (p - m)\frac{2L}{\lambda t}[pD(\lambda p) - (p + t/2)D(\lambda p + \lambda t/2)]} > 0$$

where the inequality follows from the fact that $pD(p)$ is a decreasing function whenever the elasticity of demand is greater than one (i.e. the elasticity of marginal utility is less then one). Second, totally differentiating the free entry condition, we get: $\Pi_{pp}\frac{dp}{dL} + \Pi_{p\lambda}\frac{d\lambda}{dL} + \Pi_{pL} = 0$, notice further that $\Pi_{pL} = L\Pi_p = 0$, therefore:

$$\frac{dp}{d\lambda} = -\frac{\Pi_{p\lambda}}{\Pi_{pp}}\frac{d\lambda}{dL}$$

Since $\Pi_{pp} < 0$ in equilibrium and $\frac{d\lambda}{dL} > 0$ as established, the sign of the comparative statics of the price with respect to the market size is determined by the cross derivative of the profit function. The last step is to characterize that sign:

$$\Pi_{p\lambda} = \frac{2L}{\lambda t}[-pD(\lambda p) + (p + t/2)D(\lambda p + \lambda t/2)] +$$

$$+(p - m)\frac{2L}{\lambda t}[-D(\lambda p) + D(\lambda p + \lambda t 0.5) - \lambda p D'(\lambda p) + (\lambda p + \lambda t/2)D'(\lambda p + \lambda t 0.5)]$$

We now rewrite it in terms of variables $q_0 = D(\lambda p)$ and $q_1 = D(\lambda p + \lambda t/2)$ using the fact that $D'(p) = \frac{1}{u''(D(p))}$:

$$\Pi_{p\lambda} \propto [-q_0 u'(q_0) + q_1 u'(q_1)]/\lambda + (p - m)[-q_0 + q_1 - \frac{u'(q_0)}{u''(q_0)} + \frac{u'(q_1)}{u''(q_1)}]$$

Since naturally $q_0 > q_1$because demand is a decreasing function, and $q_0 u'(q_0) > q_1 u'(q_1)$ since the elasticity of marginal utility is less than one, the first two terms in the cross-derivative are negative. The question of the sign of the cross-derivative comes down to understanding the nature of the ratio of the first and second derivatives of the utility function. Whenever $\frac{u'(q)}{u''(q)}$ is an increasing function, i.e. the elementary utility function exhibits increasing absolute risk aversion, the last term is also negative and the comparative statics are pro-competitive, i.e. prices decrease with increasing market size and competition. It is worth noting that increasing absolute risk aversion corresponds to low convexity demands, i.e. less convex than demand generated by CARA utility function. Furthermore, this condition is only sufficient and not necessary for pro-competitive effects. Indeed, even if the last term is positive it is not guaranteed to dominate the two other terms. **Q.E.D.**

3.2.3 Model with Disutility of Distance

Consumers and varieties. We now study an alternative formulation of the model. Instead of bearing monetary cost for transporting varieties produced elsewhere to their consumption point, consumers experience disutility from consumption of varieties different from their “ideal variety". In other words, transportation costs now enter the utility function rather then the budget constraint. In all other respects the setup is the same as before. In this setup the consumer problem becomes

$$\max_{q_\theta > 0} 2 \int_0^{\hat{\theta}} \mu(u(q_\theta) - q_\theta t\theta) d\theta \tag{3.4}$$

$$\text{s.t.} \quad 2\int_0^{\hat{\theta}} \mu p q_\theta d\theta = 1.$$

Here again $\hat{\theta} \in (0, 1/2]$ is her range of consumption, with $\hat{\theta} = 1/2$ representing the case of consumption of all present varieties (full coverage). For a symmetric model, studying location $x \equiv 0$ or any other makes no difference, and instead of absolute location we focus on the distance between consumer and producer θ. As before, denoting the demand function $D(\cdot) \equiv {u'}^{-1}(\cdot)$, solution to the consumer problem implies that demand for a variety from θ, whenever positive, is given by $q_\theta = D(\lambda p_\theta + t\theta)$, where λ is the Lagrange multiplier of the budget constraint, i.e. marginal utility of money and, at the same time, intensity of competition. Now we can observe the main difference between the two setups: since the cost of the mismatch between consumer and producer locations are now non-monetary, they are not multiplied by the marginal utility of money in the demand function. In other words, there is no need for the additional step of "translating" monetary cost into utility units.

Producers. As before, there is a continuum of producers, and each producer takes the intensity of competition λ and the demand schedule as given when maximizing her profit in price

$$\max_{p\geq 0} \Pi(p, \lambda) = \max_{p \geq 0} 2(p-m)L\int_0^{\hat{\theta}(p)} D(\lambda p + t\theta)d\theta - F(\mu).$$

The producer's problem is similar to the monetary transportation cost case. We simplify the objective function using the change of integration axes: instead of integrating over locations, we integrate consumption over prices. That gives rise to a relatively simple representation of aggregate demand. Then, in the case of partial coverage $\hat{\theta} < 1/2$, any producer's profit can be rewritten as

$$\Pi(p, \lambda) = 2(p-m)\frac{L}{t}[u(D(\lambda p)) - \lambda p D(\lambda p)] - F(\mu)$$

Analogously for the case of full coverage $\hat{\theta} = 1/2$:

$$\Pi(p, \lambda) = 2(p-m)\frac{L}{t}\left[u(D(\lambda p)) - \lambda p D(\lambda p) - [u(D(\lambda p + t/2)) - (\lambda p + t/2)D(\lambda p + t/2)]\right] \\ -F(\mu)$$

Equilibrium. We allow firms to relocate in space and enter/exit the market. Thereby, in equilibrium, profit must vanish at each location: $\Pi(p,\lambda)=0$. The free entry condition along with the firm's optimality condition defines equilibrium in (p,λ) variables. All other equilibrium variables can be derived from these two variables using the consumer's optimality condition and budget constraint.

Symmetric equilibrium is a bundle $(p,\mu,\lambda,\hat{\theta})\in R_{+}^{4}$ including the price, mass of firms, marginal utility of income, and radius of service that satisfy consumer and producer optimization conditions, free-entry, and budget constraint.

Characterizing the equilibrium. We start by characterizing the equilibrium in the case of only partial coverage $\hat{\theta}<1/2$. First, observe that the first and second order condition for profit maximization essentially do not differ between the two versions of the model. Indeed, the only difference between a firm's objective function in the two cases is the multiplier λ in the variable part, which is treated by the producers as exogenous. This observation allows for the straightforward characterization of equilibrium in variables (p,λ) through the profit maximization and free entry conditions:

$$\frac{u(D(\lambda p))}{\lambda D(\lambda p)}=2p-m, \qquad 2(p-m)\left[u(D(\lambda p))-\lambda pD(\lambda p)\right]=\frac{tF(\mu)}{L}$$

The only difference from the previous case is an absence of multiplier λ in the free entry condition. This relatively simple characterization of the equilibrium allows us to study the comparative statics with respect to the market size and disutility cost.

Proposition 13. *Consider the version of the model with disutility cost and with partial market coverage. Then an increase in the market size L or a decrease in the transportation cost t leads to: (i) an increase in the intensity of competition λ; (ii) a decrease in the purchase of the ideal variety q_0; (iii) a decrease (an increase) in the prices when $\varepsilon_u(q)$ is a decreasing (increasing) function. In addition, expanding market size L leads to the competition being more localized viewed as the decrease in the coverage $\hat{\theta}$.*

Proof. We start again by noticing that market size L and disutility cost t enter the free entry condition as a ratio. Therefore, the comparative statics effects on prices and intensity competition are symmetric. We use the firm's second order condition, which is the same as in the case of the monetary cost of distance, and which can be expressed as

$$2+\frac{p-m}{p}\frac{u'(q_0)}{q_0u''(q_0)}>0$$

We again study the equilibrium through quantities, and make use of the fact that consumption of a variety produced by the closest firm is $q_0 = D(\lambda p)$, and $\lambda p = u'(q_0)$. The zero profit and free-entry conditions become

$$\frac{u(q_0)}{q_0 u'(q_0)} = 2 - m/p, \qquad (p-m)\left[u(q_0) - q_0 u'(q_0)\right] L = tF(\mu)/2$$

Totally differentiating them, we obtain

$$-\frac{\varepsilon_u'(q_0)}{\varepsilon_u^2(q_0)}\frac{dq_0}{dL} = \frac{m}{p^2}\frac{dp}{dL}$$

and

$$\left[u(q_0) - q_0 u'(q_0)\right] L\frac{dp}{dL} - q_0 u''(q_0)(p-m)L\frac{dq_0}{dL} + \left[u(q_0) - q_0 u'(q_0)\right](p-m) = 0$$

Again, price and quantity co-move when the elasticity of utility is decreasing, and move oppositely when the elasticity of utility is increasing. Combining the two we get:

$$\left[-\frac{\varepsilon_u'(q_0)}{\varepsilon_u^2(q_0)}\frac{p^2}{m}\left[u(q_0) - q_0 u'(q_0)\right] - q_0 u''(q_0)(p-m)\right]\frac{dq_0}{dL} = -\frac{\left[u(q_0) - q_0 u'(q_0)\right](p-m)}{L}$$

The right hand side of it is clearly negative. We now study the the bracketed term on the left hand side, using the fact that from the firm's first order condition, it follows that $u(q_0) - q_0 u'(q_0) = \frac{p-m}{p} q_0 u'(q_0)$:

$$-\frac{\varepsilon_u'(q_0)}{\varepsilon_u^2(q_0)}\frac{p^2}{m}\left[u(q_0) - q_0 u'(q_0)\right] - q_0 u''(q_0)(p-m) =$$

$$= -q_0 u''(q_0)\left[p - m + \frac{u'(q_0)u(q_0) + q_0 u''(q_0)u(q_0) - q_0 u'^2(q_0)}{q_0^2 u'^2(q_0)}\frac{p^2}{m}\frac{u(q_0) - q_0 u'(q_0)}{q_0 u''(q_0)}\right] =$$

$$= -q_0 u''(q_0)\left[p - m + \left(\frac{1}{q_0 \varepsilon_u(q_0)} + \frac{u''(q_0)}{u'(q_0)\varepsilon_u(q_0)} - \frac{1}{q_0}\right)\frac{p^2}{m}\frac{p-m}{p}\frac{q_0 u'(q_0)}{q_0 u''(q_0)}\right] =$$

$$= -q_0 u''(q_0)(p-m)\left[1 + \left(\frac{1}{q_0 \varepsilon_u(q_0)} + \frac{u''(q_0)}{u'(q_0)\varepsilon_u(q_0)} - \frac{1}{q_0}\right)\frac{p}{m}\frac{u'(q_0)}{u''(q_0)}\right] =$$

$$= -q_0 u''(q_0)(p-m)\left[1 + \frac{p}{m}\frac{1}{\varepsilon_u(q_0)} + \frac{1}{q_0}\left(\frac{1}{\varepsilon_u(q_0)} - 1\right)\frac{p}{m}\frac{u'(q_0)}{u''(q_0)}\right] =$$

$$= -q_0 u''(q_0)(p-m)\left[1 + \frac{p}{m}(2 - \frac{m}{p}) + (1 - \frac{m}{p})\frac{p}{m}\frac{u'(q_0)}{q_0 u''(q_0)}\right] =$$

$$= -q_0 u''(q_0)(p-m)\frac{p}{m}\left[2 + \frac{p-m}{p}\frac{u'(q_0)}{q_0 u''(q_0)}\right] =$$

In the last expression the outer term is clearly positive since $u''(\cdot) < 0$, and the bracketed term is positive because of the firm's second order condition. Altogether, this implies that $\frac{dq_0}{dL} < 0$, i.e. consumption of an ideal variety decreases with increasing market size. The result for the prices follows from the discussion above.

The next parameter of interest is the range of service $\hat{\theta}$. To understand its behavior consider the demand there: $D(\lambda p + t\hat{\theta}) = 0$, or alternatively $\lambda p + t\hat{\theta} = u'(0) = u_0$. The last step is to note that $\lambda p = u'(q_0)$, thus,

$$\hat{\theta} = \frac{1}{t}(u_0 - u'(q_0))$$

It immediately follows that, following the change in the market size L, behavior of the range of service replicates that of the consumption of ideal variety q_0. Hence, the range of service decreases with market size in the case of the increasing elasticity of utility. Put differently, when elasticity of utility is an increasing function, an increase in the market size leads to more localized competition. Finally, to show that intensity of competition increases, we totally differentiate free entry condition and use $\Pi_p = 0$. Then

$$\frac{d\lambda}{dL} = -\frac{\Pi_L}{\Pi_\lambda} = \frac{F(\mu)/L}{2(p-m)\frac{L}{t}pD(\lambda p)} > 0$$

and intensity of competition increases. **Q.E.D.**

Thus, we have shown that both versions of the model exhibit similar comparative statics: the market is pro-competitive whenever the elasticity of utility is a decreasing function. The other variables behave naturally: an increase in the market size intensifies competition, leads to smaller consumption of each variety, and to more localized competition.

The intuition behind the result remains the same independently of the model formulation. The pro- or anti-competitive behavior of the market is determined by the increasing or decreasing elasticity of the aggregate demand. However, the aggregate demand of the heterogeneous consumers does not inherit the properties of their individual demand, rather the behavior of the elasticity of aggregate demand depends on higher order properties of individual demands.

Characterizing the equilibrium with full coverage. Now we consider properties and comparative statics of equilibrium when coverage is full. As in the version of the model with monetary transportation cost, this case is substantially less analytically tractable. Start with the case when the reason for a consumer to buy from each and every firm is sufficiently low transportation cost. Then, using the first order Taylor approximation, we can write

$$u(D(\lambda p+t/2))-(\lambda p+t/2)D(\lambda p+t/2)=u(D(\lambda p))-\lambda pD(\lambda p)+-D(\lambda p)\frac{t}{2}+o(t)$$

Substituting it into profit definition for the case of full coverage and using only the first order approximation, we effectively obtain an approximation of the profit function:

$$\Pi(p,\lambda)=(p-m)LD(\lambda p)-F(\mu).$$

In other words, the model collapses to a case with no distance, as studied in ZKPT. As they show, in this case, the behavior of the elasticity of marginal utility (rather than utility itself) defines the direction of comparative statics effects with respect to market size. This observation sheds light on the model behavior between the two extreme cases, i.e. when firms serve all consumers but the disutility from shopping far away is not sufficiently small.

Now we turn to formal analysis of the comparative statics under full coverage. Denote $q_0 = D(\lambda p)$ the consumption of the ideal variety and $q_1 = D(\lambda p + t/2)$ the consumption of the least liked variety, i.e. one produced at the opposite point of the circumference. Differentiating profit with respect to price, we obtain the firm's first order condition:

$$\Pi_p(p,\lambda)=2\frac{L}{t}\left[u(q_0)-\lambda pq_0-[u(q_1)-(\lambda p+t/2)q_1]\right]-2(p-m)\frac{L}{t}\lambda(q_0-q_1)=0$$

The transparent complexity of the model in the case of full coverage precludes full characterization of the comparative statics. Nevertheless, it is still possible to guarantee pro-competitive behavior of the market when demand is not too convex.

Proposition 14. *Consider the version of the model with monetary cost of transportation and with complete market coverage. Then an increase in market size L leads to an increase in the intensity of competition λ. Moreover, if the demand function $D(\cdot)$ is concave, or marginal cost of production $m = 0$, then increasing market size generates leads to decreasing prices p, and the market is pro-competitive.*

Proof. As before, we begin comparative statics analysis with respect to the market size assuming that congestion is sufficiently small, so that we can disregard the indirect effect stemming from entry or exit of firms. In this case, one can write down the free entry condition as:

$$\Pi(p, \lambda, L) = \frac{2L}{t}(p-m)\left[u(D(\lambda p)) - \lambda p D(\lambda p) - \left[u(D(\lambda p + \frac{t}{2})) - (\lambda p + \frac{t}{2})D(\lambda p + \frac{t}{2})\right]\right]$$
$$-F(\mu) = 0.$$

Totally differentiating it with respect to L we arrive at:

$$\Pi_p \frac{dp}{dL} + \Pi_\lambda \frac{d\lambda}{dL} + \Pi_L = 0$$

The first term is zero because of the profit maximization. Hence

$$\frac{d\lambda}{dL} = -\frac{\Pi_L}{\Pi_\lambda} = \frac{tF(\mu)/L}{2(p-m)Lp(q_0 - q_1)} > 0$$

In other words, as intuitively expected, an increase in the market size leads to an increase in the intensity of competition measured as the marginal utility of money.

To understand the effect of increasing market size on the pricing behavior of firms, consider now the firm's first order condition written as:

$$\Pi_p(p, \lambda, L) = 2\frac{L}{t}\left[u(D(\lambda p)) - \lambda p D(\lambda p) - \left[u(D(\lambda p + t/2)) - (\lambda p + t/2)D(\lambda p + t/2)\right]\right] -$$

$$-2(p-m)\frac{L}{t}\lambda(D(\lambda p) - D(\lambda p + t/2)) = 0$$

Firstly, note that the second order condition requires:

$$\Pi_{pp} = -4\frac{L}{t}\lambda(q_0 - q_1) - 2(p-m)\frac{L}{t}\lambda^2(D'(\lambda p) - D'(\lambda p + t/2)) < 0$$

Now, we totally differentiate the firm's first order condition with respect to market size to characterize the comparative statics:

$$\Pi_{pp}\frac{dp}{dL} + \Pi_{p\lambda}\frac{d\lambda}{dL} + \Pi_{pL} = 0$$

As $\Pi_{pp} < 0$, $\frac{d\lambda}{dL} > 0$ as established, and linearity in L implies that $\Pi_{pL} = L\Pi_p = 0$, the direction of the comparative statics is determined by the reaction of marginal profit to changes in the intensity of competition $\Pi_{p\lambda}$:

$$\frac{dp}{dL} \propto \Pi_{p\lambda} = -2(2p-m)\frac{L}{t}\left[q_0 - q_1\right] - 2(p-m)\frac{L}{t}\lambda p(D'(\lambda p) - D'(\lambda p + t/2))$$

The first term here is clearly negative. The sign of the second term depends on the shape of demand function. In case of concave or linear demand $D''(\cdot) \leq 0$ and the second term is negative as well, implying unambiguously that in response to an increase in the market size, prices go down. Alternatively, if demand is convex, the second term is positive and the total effect is ambiguous. Whether the second term outweighs the first depends on how strong the convexity of demand is. However, the general conclusion of the literature on the role of demand shapes persists; concave and not too convex demands generate pro-competitive effects — prices go down with the increase in market size and competition — whereas very convex demands generate anticompetitive market outcomes.

Alternatively, if the marginal cost $m = 0$, then $\Pi_p(\lambda, p)$ is a function of the product λp only. Therefore, in equilibrium the product is constant independently of the market size L, and the behavior of the price is exactly opposite to the behavior of the intensity of competition, and the market is always pro-competitive. **Q.E.D.**

The conditions elicited in Proposition 14 are strong. However, they are only sufficient but not necessary. In fact, from the continuity argument we expect the market also to behave pro-competitively in the presence of small marginal cost and low convexity demands.

3.3 Conclusion

This paper attempts to bridge two traditions in modeling markets with horizontal product differentiation. We develop a model that features both product space characteristics in the spirit of Hotelling (1929) and monopolistic competition as introduced by Chamberlin (1933). The preference structure we employ allows consumers in the model to have an ideal product and love for variety at the same time, which leads to the consumption of a wider range of varieties of products but in different quantities. In doing so we aimed to capture the idea that in real life consumers shop in a limited number of shops and consume a particular type of a given product most of the time, but occasionally deviate. In addition, the model formalizes the idea that love for variety in aggregate stems not only from personal preference for variety, but also from heterogeneity of preferences, and therefore might seem stronger on the aggregate than the individual level.

Our main contribution is to show that despite its cumbersome structure, this approach can still be tractable in a number of important cases. We characterize a uniform equilibrium when product space is a circumference, i.e. symmetric, consumers are uniformly distributed across it, and the cost of a mismatch between the location of a consumer and product are linear, either in monetary or in utility terms. We show that in all versions of the model under the most natural and widely used preference shapes — when demand is not too convex — the market behaves pro-competitively: in response to an increasing market size, prices decrease. At the same time convex demands can generate anti-competitive market effects. In other words, our work reinforces the conclusion of spaceless monopolistic competition theory on the connection between comparative statics effects and the shape of consumer preferences.

The other important question that remains outside the scope of this paper is what spacial distribution of firms may arise under the market structure we study. Throughout the paper we have focused on the uniform distribution of firms which can be intuitively understood as the maximal differentiation equilibrium. However, can it be the case that free entry of firms can lead to standardization of products in the characteristic space and minimal differentiation as was believed by Hotelling? Can competition of firms per se lead to the agglomeration of firms in space? More formally, this is the question of multiplicity of equilibriua and stability of the uniform equilibrium. Our preliminary inquiry shows that indeed, under very flat demands, maximum differentiation is unstable and standardization occurs as an equilibrium outcome, but the general result is yet to come. We believe that further clarification of this question is an important issue and leave it for future work.

Bibliography

Anderson, S. P., De Palma, A., Thisse, J. F. (1992). *Discrete choice theory of product differentiation*. MIT press.

Chamberlin, E. H. (1933). *The theory of monopolistic competition*. Cambridge, MA: Harvard University Press.

Dixit, A. K., Stiglitz, J. E. (1977). Monopolistic competition and optimum product diversity. *The American Economic Review*, 67(3), 297-308.

Hotelling, H. (1929. Stability in Competition. *Economic Journal* 39 (153): 41–57

Lancaster, K. J. (1966). A new approach to consumer theory. *The Journal of Political Economy*, 132-157.

Osharin, A., Thisse, J. F., Ushchev, P., Verbus, V. (2014). Monopolistic competition and income dispersion. *Economics Letters*, 122(2), 348-352.

Salop, S. C. (1979). Monopolistic competition with outside goods. *The Bell Journal of Economics*, 10(1), 141-156.

Vogel, J. (2008). Spatial competition with heterogeneous firms. *Journal of Political Economy*, 116(3), 423-466.

Zhelobodko, E., Kokovin, S., Parenti, M., Thisse, J. F. (2012). Monopolistic competition: Beyond the constant elasticity of substitution. *Econometrica*, 80(6), 2765-2784.

3.A Appendix 3

Here we quantify the notion of small enough congestion force of firm agglomeration in the space, for the case of disutility of distance and incomplete coverage. The other cases can be studied analogously.

First, denote elasticity of utility function as $\frac{qu'(q)}{u(q)} = \varepsilon_u(q)$. Then the firm's first order condition can be rewritten as $\frac{1}{\varepsilon_u(q_0)} = 2 - \frac{m}{p}$. Concavity of $u(\cdot)$ together with $u(0) = 0$ imply that $\varepsilon_u(q) < 1$. In addition, a positive solution to the price maximization problem implies $\varepsilon_u(q_0) > 1/2$ in equilibrium. We start with the budget constraint combined with optimal pricing by the firm:

$$1 = 2\int_0^{\hat{\theta}} \mu p q_\theta d\theta = 2\mu p \int_0^{\hat{\theta}} D(\lambda p + t\theta) d\theta = 2\mu \frac{m\varepsilon_u(q_0)}{2\varepsilon_u(q_0) - 1}\frac{1}{t}(u(q_0) - u'(q_0)q_0)$$

Putting together the firm's first order condition, free entry condition and budget constraint leads to an equation with only one variable q_0:

$$2m\frac{1-\varepsilon_u(q_0)}{2\varepsilon_u(q_0)-1}\left[u(q_0) - u'(q_0)q_0\right] = \frac{t}{L}F\left(\left[2\frac{m\varepsilon_u(q_0)}{2\varepsilon_u(q_0)-1}\frac{1}{t}(u(q_0) - u'(q_0)q_0)\right]^{-1}\right)$$

$$2m\frac{1-\varepsilon_u(q_0)}{2\varepsilon_u(q_0)-1}\left[u(q_0) - u'(q_0)q_0\right] = \frac{t}{L}F\left(\left[2\frac{m\varepsilon_u(q_0)}{2\varepsilon_u(q_0)-1}\frac{1}{t}(u(q_0) - u'(q_0)q_0)\right]^{-1}\right)$$

Using this equation as an implicit function $q_0(L)$ and taking derivatives with respect to L we obtain the comparative statics characterization:

$$2m\frac{-1}{(2\varepsilon_u(q_0)-1)^2}\left[u(q_0) - u'(q_0)q_0\right]\varepsilon_u'(q_0)\frac{\partial q_0}{\partial L} + 2m\frac{1-\varepsilon_u(q_0)}{2\varepsilon_u(q_0)-1}\left[-u''(q_0)q_0\right]\frac{\partial q_0}{\partial L} =$$

$$= -\frac{t}{L^2}F\left(\mu\right) + \frac{t}{L}F'(\mu)\frac{t}{2m}\left[\frac{\varepsilon_u'(q_0)}{\varepsilon_u^2(q_0)(u(q_0) - u'(q_0)q_0)} + \frac{u''(q_0)q_0(2\varepsilon_u(q_0)-1)}{\varepsilon_u(q_0)(u(q_0) - u'(q_0)q_0)^2}\right]\frac{\partial q_0}{\partial L}$$

The last term in this equation is the indirect effect stemming from congestion force. First, note that in the case of the decreasing elasticity of utility, the indirect effect reinforces the direct effect of the market size through increasing competition. Therefore, independently of the size of the $F'(\cdot)$, the consumption of ideal variety is always decreasing with the market size. However, in the case of increasing elasticity of utility, an indirect effect can potentially have a different sign, therefore, the requirement that congestion is not too strong, i.e. fixed costs do not rise too fast, can be formally expressed as:

$$\frac{2m}{2\varepsilon_u(q_0)-1}Abs\left\{\frac{\left[u(q_0)-u'(q_0)q_0\right]\varepsilon_u'(q_0)}{(2\varepsilon_u(q_0)-1)^2}+u''(q_0)q_0(1-\varepsilon_u(q_0))\right\} >$$

$$> \frac{t^2}{2mL}F'(\mu)\frac{1}{\varepsilon_u(q_0)(u(q_0)-u'(q_0)q_0)}Abs\left\{\frac{\varepsilon_u'(q_0)}{\varepsilon_u(q_0)}+\frac{u''(q_0)q_0(2\varepsilon_u(q_0)-1)}{(u(q_0)-u'(q_0)q_0)}\right\}$$