Journal of Urban Economics 85 (2015) 16–33

Contents lists available at ScienceDirect

Journal of Urban Economics

www.elsevier.com/locate/jue

Identification in a model of sorting with social externalities and the causes of urban segregation

Maximilian Kasy *

Department of Economics, Harvard University, Littauer Center 200, 1805 Cambridge Street, Cambridge, MA 02138, United States

ARTICLE INFO

Article history:
Received 4 August 2013
Revised 9 October 2014
Available online 18 October 2014

JEL classification:
C31
D62
R21
R23

Keywords:
Identification
Sorting
Social externalities

ABSTRACT

This paper discusses nonparametric identification in a model of sorting in which location choices depend on the location choices of other agents as well as prices and exogenous location characteristics. In this model, demand slopes and hence preferences are not identifiable without further restrictions because of the absence of independent variation of endogenous composition and exogenous location characteristics. Several solutions of this problem are presented and applied to data on neighborhoods in US cities. These solutions use exclusion restrictions, based on either subgroup demand shifters, the spatial structure of externalities, or the dynamics of prices and composition in response to an amenity shock. The empirical results consistently suggest the presence of strong social externalities, that is a dependence of location choices on neighborhood composition.

1. Introduction

Urban areas in the United States and across the world show large degrees of social segregation across neighborhoods. A large and rising degree of segregation of immigrant groups across neighborhoods in US cities since 1920 has been documented, for instance, by Cutler et al. (2008). This is of concern if the social environment in neighborhoods is an important determinant of life outcomes. There are two polar explanations of segregation. Households might sort across locations because of different willingness to pay for exogenous location characteristics, which may be due to differences in income or differences in preferences. This is the explanation emphasized by accounts of sorting such as the classic Tiebout (1956) and Rosen (1974) models. Alternatively, households might care about who their neighbors are, and hence choose their neighborhood based on demographic composition. This possibility was discussed by Schelling (1971) and Becker and Murphy (2000).

The present paper discusses identification problems arising in models which allow for both possibilities. In the setup considered in this paper, households have to choose whether or not to locate in a given neighborhood based on exogenous neighborhood characteristics, and based on the endogenous composition of the residents of a neighborhood. We say social externalities are present if demand depends on endogenous composition. The local housing market is in equilibrium if the composition of households that want to locate in a neighborhood equals the composition of those that are in the neighborhood, and if total housing demand equals housing supply. This setup builds on several important recent contributions to the urban economics literature, in particular Bayer et al. (2007) and Caetano (2009). These authors estimate discrete choice models of sorting that recognize the possibility of a preference for neighborhood composition.

The central contribution of the present paper is to provide a discussion of nonparametric identification in this context. The main goal is to empirically distinguish between the two explanations of segregation, and in particular to test whether social externalities are empirically relevant. It is shown that without further restrictions the presence and degree of social externalities are not identified. This is because, in equilibrium, both composition and rental prices are functions of the exogenous neighborhood characteristics. This prevents the separate identification of the effect of either on demand. Identification requires exclusion restrictions that generate independent variation of composition and the exogenous arguments of demand. Several types of such exclusion restrictions are discussed here. In the setup analyzed, no restrictions on

* Address: Department of Economics, Harvard University, Littauer Center 200, 1805 Cambridge Street, Cambridge, MA 02138, United States.
E-mail address: maximiliankasy@fas.harvard.edu

http://dx.doi.org/10.1016/j.jue.2014.10.003
0094-1190/

functional forms or the nature of heterogeneity of households or neighborhoods are imposed. Discussions of nonparametric identification have been fruitful in the development of many applied fields in recent years, see for instance Manski (2003) or the review in Matzkin (2008).

The presence of social externalities in sorting is of relevance for several reasons. First, it poses a methodological problem in the estimation of willingness-to-pay parameters, which in turn are often used for cost-benefit analyses of policies. Second, externalities matter for understanding the causes of social segregation across locations and can amplify the effects of policies on segregation. Third, if externalities are strong, multiple equilibria in population composition at a given location arise. Multiple equilibria in turn can imply discontinuous and large effects of demand shifting policies, as emphasized by Schelling (1971) and Card et al. (2008). Finally, it is interesting to contrast the importance households attach to neighborhood composition in their location choice with the available evidence on the effect of neighborhood environment on observable outcomes. Evidence on the latter is mixed, see for example Katz et al. (2007). The present paper, on the other hand, finds strong effects of composition on location choice. These results are of course consistent with each other, but suggest that households care about neighborhood composition for other reasons than the causal impact of neighborhood composition on observable outcomes.

Three possible solutions to the identification problem are discussed in this paper. The first approach uses exogenous shifters of demand of certain subgroups that are excluded from the demand of other subgroups. Such shifters allow one to construct instruments that affect neighborhood composition, without directly affecting the demand of some subgroups. Using such instruments we can estimate the causal impact of composition on demand of these subgroups. This builds on the idea of randomized subgroup treatment used in the identification of peer effects, as recommended in Moffitt (2004) and applied for instance by Duflo and Saez (2003).

The second approach exploits the spatial structure of cities in an extension of the baseline model, allowing for interactions across adjacent neighborhoods. Identification comes from the assumption that exogenous demand shifters for neighborhoods beyond a certain distance are excluded from local demand. This allows one to use demand shifters for neighborhoods at a certain distance as instruments which affect local composition through their impact on the composition of intermediate neighborhoods, without directly affecting local demand. Using such instruments we can estimate the causal impact of composition on demand of all subgroups, as well as the causal impact on housing prices. The latter measure the marginal willingness to pay for housing in the neighborhood. This idea is analogous to the use of social network structures to identify endogenous versus exogenous peer effects, as in Bramoullé et al. (2009) and De Giorgi et al. (2010).

The third approach is based on a dynamic extension of the baseline model which is discussed in more detail in the supplementary appendix. This dynamic extension assumes search frictions in moving from one neighborhood to another. The third approach uses the finding that, under certain conditions, past amenity shocks are excluded from future price changes because the value of amenities is immediately reflected in rental prices. Composition, however, does adjust with delay due to search frictions, and hence prices adjust to this composition change with the same delay. Past amenity shocks can therefore be used as instruments which affect future composition changes without directly affecting future changes in prices. Using such instruments we can estimate the causal impact of neighborhood composition on housing prices (marginal willingness to pay). The dynamic model considered is similar to search models of the labor market as surveyed in Pissarides (2000). It builds upon search models of the housing market such as Wheaton (1990).

These approaches are applied to data from the Neighborhood Change Database (NCDB), which aggregates US Census data to the level of census tracts. The composition variable considered is Hispanic share. Various instruments for neighborhood composition are constructed that build on the three approaches to identification just discussed. All instruments yield surprisingly consistent estimates. They suggest that a 1% increase in the Hispanic share of neighborhood population results in a 6% to 10% decline in non-Hispanics' demand, and a 3% to 4% rise in Hispanics' demand. Housing prices appear to decline by around 0.5% to 1% for a 1% increase in Hispanic share. These results are also consistent with the conclusions of Cutler et al. (2008), who use variation in segregation across time, city, and immigrant groups in trying to disentangle the causes of segregation. One might wonder why we are focusing our main empirical analysis on Hispanic share, rather than on other dimensions of urban segregation. The main reason is that immigration created a lot of arguably exogenous variation in composition, which we exploit.

The model in this paper is described in terms of households choosing a neighborhood and paying rents. However, most of the insights should apply to other contexts of sorting. Examples include sorting of workers across firms, students across schools, customers across mobile-phone network providers, faculty across universities, or the spatial agglomeration and dispersion of firms. In each of these settings agents might have a (reduced form) preference for peers, which is empirically hard to separate from location heterogeneity, but which has implications for interesting counterfactuals.

Some further relevant contributions in the recent literature have to be mentioned before proceeding. Solutions to the omitted variable problem in hedonic or choice regressions have been proposed by Black (1999), who controls for border fixed effects, and by Chay and Greenstone (2005), who use exogenous variation in amenities. Nesheim (2001) and Graham (2008b) discuss identification issues in specific models of sorting where peer composition enters an educational production function. Heckman et al. (2002) and Ekeland et al. (2004) derive identification of preferences from cross-sectional price data based on functional form restrictions (separability). Chiappori et al. (2009) show the equivalence of hedonic sorting, matching and optimal transport problems and derive existence results for equilibria in these models.

The rest of the paper is structured as follows: Section 2 introduces a model of locational sorting and discusses its assumptions and the fundamental identification problem in this model. Section 3 proposes three solutions to this problem, based on subgroup shifters, the spatial structure of cities, and the dynamic structure of neighborhood composition and prices. Section 4 applies these three solutions to the NCDB data. Section 5 concludes. All proofs are relegated to Appendix A. Additional discussions can be found in a supplementary appendix.

2. Model and identification problem

This section will first state the model assumptions and the basic non-identification result which motivates the present paper. Then the model assumptions will be discussed. A special case of the model will be used to provide some graphic intuition for the comparative statics of the model and for the source of the identification problem.

We will consider the following model of the local housing market in a given neighborhood. There are $\mathscr{C}$ types of households, $c = 1, \ldots, \mathscr{C}$. A neighborhood is characterized by (i) the mass (number) of households of each type, $M = (M^1, \ldots, M^{\mathscr{C}}) \in \mathbb{R}^{\mathscr{C}}$, (ii) a (rental) price P, and (iii) an exogenous vector $X \in \mathbb{R}^{k_X}$ of *all* other location characteristics and factors influencing demand or supply. An example component of the neighborhood characteristics vector

X would be geographic location. An example of the composition vector M would be the share of various ethnic groups living in the neighborhood. Demand D^c for housing in a neighborhood, for each type c of households, is a bounded continuously differentiable function of X, M, P. Let $D = (D^1, \ldots, D^{\mathscr{C}})$, then

$$D = D(X, M, P). \tag{1}$$

Total demand is given by $E := \sum_c D^c$. Housing supply is a bounded continuously differentiable function of P and X,[1]

$$S = S(P, X). \tag{2}$$

This model allows for social externalities, in the sense that demand for housing at a location may depend on the composition of residents at that location. This dependence can reflect a direct preference over neighbors' types. It can also reflect a preference over amenities or production processes affected by neighbors' types, such as peer effects in education, crime etc., as in Nesheim (2001) or Graham (2008b).

In this setup we can define a notion of *partial sorting equilibrium*. The following definition requires that the neighborhood composition is consistent with the demand of each of the different types, and that housing demand equals housing supply. The equilibrium we consider is a "partial" equilibrium in the sense that it does not take into account interactions across neighborhoods, where a change in the set of available outside options might affect local demand. Partial Sorting Equilibrium extends the conventional requirements of partial equilibrium, where housing supply equals housing demand. In addition to such clearing of the housing market we also require that the composition of people who *want* to live in the neighborhood equals the composition of people who *do* live in the neighborhood.

Definition 1 (*Partial Sorting Equilibrium*). A partial sorting equilibrium (M^*, P^*) given X solves the $\mathscr{C} + 1$ equations

$$D(X, M^*, P^*) = M^*, \tag{3}$$

$$S(P^*, X) = \sum_c M^{*c}. \tag{4}$$

Let $(M^*(X), P^*(X))$ denote the correspondence mapping X into the partial sorting equilibria given X.

Equilibrium existence is guaranteed under the assumptions maintained, proofs of existence and of all further results can be found in Appendix A. Our model is stated in terms of demand functions D and the housing supply schedule S. Definition 1 provides a mapping from demand functions to equilibrium schedules (M^*, P^*). These equilibrium schedules give population composition and rental prices in a neighborhood as a function of exogenous neighborhood characteristics and any other determinants of location choices. These determinants might be observable to the econometrician or not.

This paper is interested in using observational data to identify the slopes of the demand functions D^c with respect to X, M, and P. Of particular interest is the question of whether demand exhibits social externalities.

Definition 2 (*Social externalities*). Demand is said to exhibit social externalities if $D_M \neq 0$. [2]

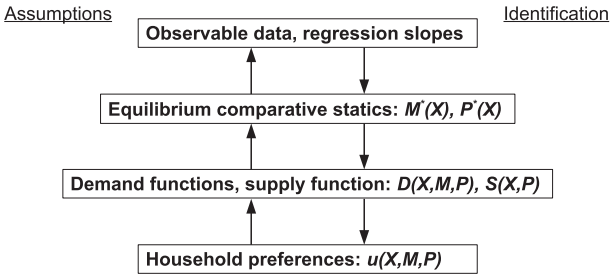

Fig. 1. Assumptions and steps of the identification problem.

2.1. The bigger picture

With the model set up, we can now clarify further what this paper is about. The ultimate goal here, and in many economic equilibrium models, is to learn about economic primitives (preferences, technologies) from observational data. This goal can – and should – be decomposed into several steps, as illustrated in Fig. 1. We can think of identification as essentially inverting a series of mappings from more primitive objects to more "high-level" objects, implied by modeling assumptions. The mapping from preferences to choice functions (demand and supply) follows from an assumption of utility maximization. The mapping from choice functions to equilibrium outcomes follows from the assumption of partial sorting equilibrium in our setting; in other settings this is replaced by Nash equilibrium or Walrasian equilibrium. The mapping from equilibrium comparative statics to observable data distributions follows from assumptions about the distribution of components of X and about observability.

Now for the inversions of these mappings: Learning from observable data about equilibrium comparative statics is an instance of the standard program evaluation problem. We need (quasi-)experimental variation of components of X to learn about the causal effect of these components on M and P. Learning from choice functions about preferences is an instance of discrete choice estimation. This step requires strong restrictions on unobserved heterogeneity and functional forms; we choose to stop at identification of choice functions. The "middle part," mapping from equilibrium comparative statics to choice functions, is what the conceptual half of this paper is about.

We clarify, in particular, the nature of the identification problem, which is related to the problems of simultaneity and the reflection problem.[3] We show that, even with random variation of the components of X, demand functions are not identified due to a fundamental support issue: In equilibrium, there is no variation in neighborhood composition, given the amenities X, so that the joint support of X and (M, P) is degenerate. Therefore we cannot identify demand and supply outside this support, and cannot separate the effects of X and M on demand. We then provide principled solutions to this problem, using exclusions restrictions and various extensions of the baseline model.

Our analysis also illustrates the dangers of parametric identification, which is common in the structural literature. As the proof of our non-identification result reveals, arbitrary conclusions about social externalities can be drawn when imposing certain parametric models, which are always consistent with the data. Section 2.4 below provides some numerical illustrations of this point. There is a subtle but important distinction between the role of parametric

[1] Vertical supply, with S not depending on P is covered as a limiting case.

[2] Throughout this paper, superscripts denote indices (for instance M^c is the cth component of M) and subscripts denote partial derivatives (for instance $D_M = \partial D / \partial M$).

[3] The "reflection problem," as introduced by Manski (1993), is the problem of separately identifying endogenous and exogenous peer effects. Exogenous peer effects are causal effects of predetermined peer characteristics. Endogenous peer effects are causal effects of peer outcomes, which might in turn be affected by their peers, leading to the possibility of feedback or multiplier effects.

assumptions in identification and in estimation. While parametric assumptions at the identification stage lead to wrong conclusions in arbitrarily large samples, parametric models used for estimation might be optimal to deal with the inherent variance-bias tradeoffs of point estimation in finite samples. Such parametric models can be thought of as coming with the implicit promise of being replaced by more flexible models as sample sizes get larger (a point formalized in the theory of series estimation), thus yielding consistent estimates independent of parametric assumptions.

It is finally worth emphasizing the distinctions between our setting and the so-called "reflection problem." Table 1 summarizes these distinctions; we will discuss them further in Section 2.2 below.

2.2. Non-identification and discussion of assumptions

Under the assumption that local housing markets are in equilibrium, observational data will at best reveal the equilibrium correspondence $(M^*(X), P^*(X))$. The fundamental challenge which will be discussed in this paper is to map this equilibrium correspondence back into (the slopes of) demand. The following proposition shows that, without further restrictions, demand slopes are not identified.

Proposition 1 (*(Non) identification*). *Suppose the equilibrium correspondence $(M^*(X), P^*(X))$is known, but no further information about D, S is available. Then $D(X, M, P)$ is identified for $(M, P) \in (M^*(X), P^*(X))$. $D(X, M, P)$ is not identified for $(M, P) \notin (M^*(X), P^*(X))$.*

Assume additionally that partial sorting equilibrium is unique or let (M^*, P^*) be a differentiable selection from the set of partial equilibria. Then linear combinations of the demand slopes are identified as

$$D_X + D_M M^*_X + D_P P^*_X = M^*_X. \tag{5}$$

No other linear combinations of (D_X, D_M, D_P) are identified.

This proposition implies in particular that, without further restrictions, the equilibrium schedule (M^*, P^*) is completely uninformative about the presence of social externalities. We can never reject the hypothesis $D_M = 0$. More formally, Proposition 1 states the non-invertibility of the mapping from demand and supply functions $D(.)$ and $S(.)$ to the equilibrium correspondence $(M^*(.), P^*(.))$ given by Definition 1. In its differentiated form, the proposition states the non-invertibility of the mapping from the demand and supply slopes (D_X, D_M, D_P) and (S_X, S_P) to the slopes of the equilibrium schedule, (M^*_X, P^*_X). The positive identification results discussed in Section 3 will state economically interpretable additional conditions under which we *can* draw conclusions about D_M from knowledge of (M^*_X, P^*_X).

There is a parallel between the identification problem stated in Proposition 1 and other well known identification problems. One is the classic simultaneity problem in identifying price elasticities, the other is the reflection problem (Manski, 1993) in the identification of models with endogenous peer effects. In all these problems, an endogenous equilibrium outcome serves as an argument to some structural relationship. There is no (continuous) variation of the equilibrium outcome conditional on the other arguments of the same relationship, at least not without further exclusion restrictions.

REMARKS REGARDING THE MODEL ASSUMPTIONS:

(1) This is a nonparametric model of the local housing market, as we are not restricting functional forms of demand or utility, the heterogeneity of utility, or the dimensionality of factors X influencing household choices. In particular, the model does *not* restrict the heterogeneity of utility within and across types of households. Composition is restricted, however, to enter household utility only in terms of the number of households of each type present in the neighborhood, which is a strong assumption. This assumption however holds trivially under the null hypothesis that there are no social externalities. The fact that we use a nonparametric model in order to discuss identification is one of the main contributions of the present paper. It stands in contrast to the approach taken in many recent contributions to the urban economics literature, such as Bayer et al. (2007) and Caetano (2009). It is motivated by the argument, made plausible in contributions such as Manski (2003) and Matzkin (2008), that credible identification should rely only on economically interpretable assumptions but not on functional form assumptions – even if we later proceed by estimating the nonparametrically identified model using a parametric specification.
(2) Another important feature of this model is that equilibrium is defined as partial equilibrium of the housing market in a given neighborhood. This is restrictive if we interpret the demand functions as reflecting household utility maximization under constant outside options, as discussed in the supplementary Appendix A. Under this assumption of constant outside options, $-D_X/D_P$ and $-D_M/D_P$ can be interpreted as the average marginal willingness to pay of marginal households for changes in X and M. The partial equilibrium perspective is justified if the given neighborhood is one of many similar ones, so that general equilibrium feedbacks can be ignored to first order. This argument can be made formally rigorous but is beyond the scope of the present paper.
(3) The vector X is defined inclusively as comprising *all* exogenous demand and supply shifters, including random fluctuations. Locations in this model only differ if X is different, all other variables will be endogenously determined given the exogenous X. Note that demand is affected not only by local characteristics. Demand also depends on the composition and size of the population of potential residents, including the distribution of preferences and income. Local housing demand of Hispanics will be affected by a nation-wide increase of the Hispanic population through immigration. The empirical application at the end of the paper exploits this and uses changing demographic composition due to immigration as a demand shifter. The vector X should thus be thought of as including not only local amenities, but also other factors shifting demand including the composition and size of the entire population.
(4) It is important to recognize the differences between the setup developed here and the models of peer effects discussed in the literature, e.g. Manski (1993) and Brock and Durlauf

Table 1
Comparison to models of peer effects.

Sorting with social externalities	Peer effects, as in Manski (1993) or Moffitt (2004)
Endogenous set of agents with fixed characteristics	Fixed set of agents with endogenous outcomes
Simultaneity problem: about identifying whether there are social externalities at all	Reflection problem/simultaneity: distinguishing endogenous from exogenous peer effects
Price mechanism allocating households to neighborhoods	–
Sorting is object of interest	Sorting is cause of identification problems, nuisance

(2001). First, in sorting models a location is matched with an endogenous set of agents with fixed characteristics. In contrast, in models such as those discussed by Manski (1993) or Moffitt (2004), there is a fixed set of agents with endogenous outcomes. Second, the reflection problem in models of peer effects is the problem of distinguishing endogenous from exogenous peer effects, *not* the problem of distinguishing peer effects from non-random matching.[4] In the sorting model developed here the fundamental problem is to identify whether there are social externalities at all. Third, in the setup discussed here, there is a price mechanism allocating households to neighborhoods. Such a mechanism is absent from peer-effects models. Finally, in peer effects models, endogenous sorting might be a cause of identification problems, and as such is a nuisance. Here it is the object of interest.

(5) Proposition 1 assumes that the partial sorting equilibrium is unique or that the observed equilibria correspond to a differentiable selection from the set of partial equilibria. This assumption implies that there is indeed a mapping from X to the realized equilibrium $(M^*(X), P^*(X))$, rather than a correspondence to a set of equilibria. This assumption further implies that empirically estimated slopes for equilibrium comparative statics are driven by the slopes of D and S, rather than by a causal effect of X on equilibrium selection. This assumption finally implies that the conditional support of (M, P) given X is a point, rather than a (finite) set of points. While difficult to leverage in practice, in the latter case we might be willing to impose interpolating assumptions to make predictions about the behavior of D outside the observed support of its arguments.

2.3. Illustration assuming there are only two types of households

To provide some intuition for the implications of this model, let us consider a special case which is easily graphically representable. Suppose that there are only $\mathscr{C} = 2$ types of households. Assume furthermore that the price elasticity of demand of the two types is the same, $D^1_P/D^1 = D^2_P/D^2$, and that both types have the same demand elasticity with respect to the scale of the neighborhood,[5] $(D^1_{M^1}M^1 + D^1_{M^2}M^2)/D^1 = (D^2_{M^1}M^1 + D^2_{M^2}M^2)/D^2$. Define d as the share of type 1 households among those who *want* to live in the neighborhood, i.e., $d = D^1/(D^1 + D^2)$. Similarly, let m be the share of type 1 households among those who *do* live in the neighborhood, $m = M^1/(M^1 + M^2)$. Recall finally that E is the total demand for housing in the neighborhood, $E = D^1 + D^2$. Under the assumption of equal price and scale elasticities, the demand share of type 1, d, can be written as a function of m and X alone, where X is determined outside the model. This implies that partial sorting equilibrium can be defined by the conditions

$$d(m^*, X) = m^*, \tag{6}$$

$$E(P^*, m^*, X) = S(P^*, X), \tag{7}$$

which have a recursive form that we can easily analyze, both graphically and analytically. The share of either type is a solution to the first equation. Given this equilibrium share, the second condition is a conventional partial-equilibrium supply and demand equation. Fig. 2 represents these two equilibrium conditions as well as the comparative statics of the model.

[4] Exogenous peer effects are effects of exogenously determined peer characteristics. Endogenous peer effects are effects of endogenously determined peer characteristics. "Feedback" is only present if there are endogenous peer effects, so that outcomes are determined in equilibrium.

[5] The empirical application in Section 4 will also assume that there are two observable types and that households are indifferent with respect to the scale of the neighborhood. We will *not* use an assumption of equal elasticities w.r.t. prices for identification.

Consider a small change in Z^1, a component of X, that does not affect housing supply, $S_{Z^1} = 0$. Assume that social externalities are not too strong, so that $d_m < 1$. Assume furthermore that partial sorting equilibrium is unique, or let (m^*, P^*) denote a differentiable selection from the set of partial equilibria.

Then

$$m^*_{Z^1} = \frac{d_{Z^1}}{1 - d_m} \tag{8}$$

and

$$P^*_X = \frac{E_{Z^1} + E_m m^*_{Z^1}}{S_P - E_P}. \tag{9}$$

This follows immediately from Eqs. (6) and (7). Eq. (9) gives the response of rents to amenity shifts. It is interesting to compare this to the hedonic slope with inelastic supply in the absence of externalities, $-\frac{E_{Z^1}}{E_P}$. This is the response in prices that would hold housing demand constant if composition m did not change in response to changes in amenities, and corresponds to the slope that hedonic regressions try to estimate. Relative to this hedonic slope, there is an additional term in the numerator of $P^*_{Z^1}$ if there are social externalities, i.e., $E_m \neq 0$, and if equilibrium composition does depend on exogenous location characteristics, i.e., $m^*_{Z^1} \neq 0$.

This reflects the identification problem stated in Proposition 1: Knowledge about the equilibrium schedules M^* and P^* does not allow us to identify the demand functions D^c, nor in particular the slopes D^c_m and D^c_X. The reason is that m^*, in the two type case, is functionally dependent on X. There never is independent variation of the two. Therefore the slopes D^c_X and D^c_m cannot be identified separately. If the partial equilibrium is unique, any equilibrium schedule (M^*, P^*) can be rationalized by a version of the model without social externalities, for instance by setting $D^c(X, m, P) = M^{c*}(X, m^*(X), P^*(X))$.

Eq. (8) implies a "multiplier" effect in the sense that any immediate causal effect of amenities on composition, d_X, is amplified by a factor $\frac{1}{1-d_m}$. This factor is bigger than one iff $d_m > 0$, which holds if households of either type prefer to live with neighbors of the same type. Conversely, if households prefer to live with households of a different type than themselves, so that $d_m < 0$, social externalities have a dampening effect on amenity variation. In this case they lead to a more integrated residential distribution. Finally, if social preferences are strong enough, it is also quite possible that there are unstable equilibria with $d_m > 1$, in which case there must be at least two more stable equilibria. This case is the one emphasized in discussions of "tipping" such as Card et al. (2008). Kasy (2010) proposes a test for such equilibrium multiplicity and applies it to the same data used in the present paper.

2.4. The fallacies of parametric identification

Proposition 1 shows that (absent further restrictions) demand schedules are not identified from equilibrium comparative statics in our setting. Many structural models would be identified nonetheless in this context, by assuming restrictions on functional forms. Such structural models usually impose some kind of latent separability or single index restrictions. In this section we numerically illustrate the dangers of such an approach by demonstrating how minor deviations from the functional form assumed can lead to wildly divergent conclusions about social externalities.

Assume that the data are generated by the following special case of our model:

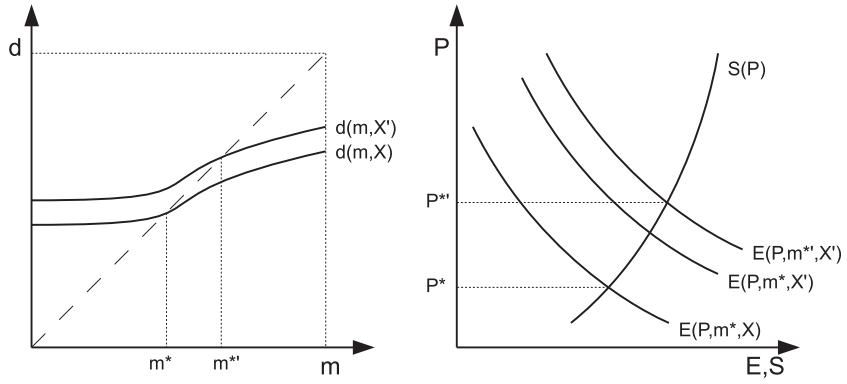

Fig. 2. Comparative statics in the simplified $\mathscr{C} = 2$ model.

- There is one exogenous amenity $X \in [0,1]$, and there are two types of households, $\mathscr{C} = 2$. There are N households of either type in the population who are potential residents of a given neighborhood.
- Housing supply is infinitely elastic, so that prices are fixed exogenously and can be ignored.
- Individuals' utility for living in a given neighborhood, relative to their best outside option, is equal to

$$u_i = V^c(X,m) + \epsilon_i,$$

where ϵ_i has an EV1 distribution, and captures individuals' idiosyncratic preferences for living in the given neighborhood. $V^c(X,m)$ is a utility term common to all households of type c.
- The common utility term $V^c(X,m)$ has a constant elasticity of substitution form,

$$V^c(X,m) = \left(\beta_0^c + \beta_X^c X^\alpha + \beta_m^c m^\alpha\right)^{1/\alpha}.$$

- Households of type 1 are indifferent across values of X and m, so that $\beta^1 = (0,0,0)$.

Under these assumptions demand of type c households is given by

$$D^c(X,m) = N \cdot \frac{1}{1+\exp(-V^c(X,m))} = N \cdot \frac{1}{1+\exp\left(-\left(\beta_0^c + \beta_X^c X^\alpha + \beta_m^c m^\alpha\right)^{1/\alpha}\right)}. \quad (10)$$

This follows immediately from the assumed form of the common utility term, and the fact that ϵ_i is i.i.d. EV1 distributed. For $c = 1$ we get $D^1(X,m) = N \cdot 0.5$, and thus

$$d(X,m) = \frac{D^2}{D^1 + D^2} = \frac{1}{1.5 + 0.5\exp(-V^2(X,m))}.$$

Now suppose that we wish to estimate this model, assuming that α is known. The most common assumption is $\alpha = 1$, in which case X and m are perfect substitutes and D is a logit demand function. The model is (parametrically) identified under the assumptions imposed, so that we can in particular recover social externalities. Assume, however, that we got α slightly wrong. Do we still get the social externalities roughly right? The simulation results in Table 2 show that this is not so.

The numbers in this table are generated assuming $(\beta_0^2, \beta_X^2, \beta_m^2,) = (1,5,0)$ for the top panel and $(\beta_0^2, \beta_X^2, \beta_m^2,) = (1,5,1)$ for the bottom panel. Across rows, we make different assumptions as to what the true parameter α is, across columns different assumptions as to what value $\widehat{\alpha}$ is assumed for α in estimation. The estimates are obtained using a very large sample where X is uniformly distributed on $[0,1]$, so that estimation error can basically be neglected. On the diagonal of either panel the model is correctly specified, and so we do indeed recover the correct value β_m^2, which equals 0 on the top and 1 on the bottom. If the true degree of substitutability is larger than assumed, however (below diagonal estimates), than we end up considerably underestimating social externalities in this model. If the true degree of substitutability is smaller than assumed (above diagonal estimates), we end up considerably overestimating externalities.

Table 2
Bias of misspecified parametric models.

α	$\widehat{\alpha}$				
	0.6	0.7	0.8	0.9	1.0
	$\beta_m^2 = 0$				
0.6	0.00	2.96	6.06	9.40	13.06
0.7	−2.68	0.00	2.73	5.60	8.69
0.8	−5.02	−2.50	0.00	2.56	5.24
0.9	−7.26	−4.80	−2.39	0.00	2.43
1.0	−9.62	−7.16	−4.71	−2.33	0.00
	$\beta_m^2 = 1$				
0.6	1.00	16.32	33.72	54.02	78.10
0.7	−8.24	1.00	11.02	22.29	35.25
0.8	−12.67	−5.97	1.00	8.53	16.92
0.9	−15.68	−10.20	−4.70	1.00	7.10
1.0	−18.46	−13.64	−8.79	−3.95	1.00

Notes: This table shows estimates of the degree of social externalities β_m^2 under various assumptions about α for the true data generating process and for estimation. For details, see discussion in Section 2.4.

3. Identification based on exclusion restrictions

This paper assumes that neighborhood housing markets are in partial sorting equilibrium. This is similar to other models which assume that data are generated as an equilibrium outcome, like standard models of supply and demand or models of static games. These models map economic primitives (preferences and technologies) to an observable data distribution. Such a map can be decomposed into three components, (i) the map from primitives to choice functions, (ii) the map from choice functions to equilibrium schedules, and (iii) the map from equilibrium schedules to the observable data distribution. The problem of identification in such models is the problem of inverting these maps. Proposition 1 states the non-invertibility of (ii) in the present setup, which maps demand and supply functions $D(.)$ and $S(.)$ to the equilibrium correspondence $(M^*(.), P^*(.))$.

Any solution to the identification problem stated in Proposition 1 has to "drive a wedge" between variation in X and in M.

In particular, if we are interested in identifying D^c_M for some subgroup c, then we need to find factors Z which affect M without directly affecting D^c. [6] Section 3.1 proposes to use factors Z which do not affect households of type c, but do affect other types of households, thus moving M. Section 3.2 proposes an extended model with externalities across adjacent neighborhoods. In such a model, local shocks propagate to more remote neighborhoods through changes of composition M in intermediate neighborhoods, thus potentially affecting demand in these remote neighborhoods without affecting the exogenous determinants X of demand in these neighborhoods.[7]

Section 3.3, finally, briefly discusses a dynamic model with search frictions, which is presented in more detail in the supplementary appendix. In this model, composition M adjusts only with delay following local shocks. If prices adjust faster than composition, then past shocks to relative demand of different types of households can serve as instruments for future composition changes in regressions of prices P on composition M.

All these approaches are concerned with inverting the mapping (ii) from demand and supply functions to the equilibrium schedule. In order to draw conclusions from data about household preferences for neighborhood composition, we also need to deal with the problems of inverting (i) and (iii). The application in Section 4 will estimate (components of) (M^*_X, P^*_X) using simple linear (instrumental variable) regressions. These are consistent estimators for weighted averages of (M^*_X, P^*_X). This addresses the problem of inverting (iii), see the supplementary appendix for further discussion. The results presented in this section then allow us to map these estimates into estimates of D_M and D_P. Under the assumption that demand schedules reflect household utility maximization under constant outside options, $-D_M/D_P$ then corresponds to the marginal willingness to pay of marginal households for composition M. This addresses the problem of inverting (i) and is again discussed in the supplementary appendix.

3.1. Subgroup shifters

Proposition 1 showed that D, and in particular the slopes (D_X, D_M, D_P), are unidentifiable due to the functional dependence of (M^*, P^*) and X. Holding X constant, there is no variation in M that allows us to identify the effect of M on D. The following proposition makes the assumption that there is a component of X which is excluded from the demand of a subgroup of households. Under such an exclusion restriction, variation in the component of X that is excluded generates variation in M and P that is not functionally dependent on the relevant arguments of demand.

Proposition 2 (*Subgroup identification*). *Assume that* $\mathscr{C} = 2$, *and that* $D = D(X, m, P)$ *where*[8] $m = M^1/(M^1 + M^2)$. *Assume furthermore that there is a component* Z^1 *of* X *such that* $D^1_{Z^1} = 0$ *but* $D^2_{Z^1} \neq 0$ *for some component* Z^1 *of* X, *so that* $m^*_{Z^1} \neq 0$. *Then*

$$D^1_m = \frac{1}{m^*_{Z^1}} \left(M^{*1}_{Z^1} - D^1_P P^*_{Z^1} \right). \tag{11}$$

Assume additionally $D^1_{Z^2} = D^2_{Z^2} = 0$ *but* $S_{Z^2} \neq 0$, *and that* $d_P = 0$. *Then*

$$D^1_m = \frac{1}{m^*_{Z^1}} \left(M^{*1}_{Z^1} - \frac{M^{*1}_{Z^2}}{P^*_{Z^2}} P^*_{Z^1} \right). \tag{12}$$

Proposition 2 suggests using instruments for neighborhood composition in regressions of subgroup demand on composition. Suppose there is variation in a variable Z^1 which is independent of variation in other components of X, where Z^1 is excluded from demand of group 1. Then correlation between M^1 and Z^1 must reflect the effect of either composition or prices on D^1. In our application, we will use the result stated in Eq. (11), and estimate weighted averages of $M^{*1}_{Z^1}, m^*_{Z^1}$, and $P^*_{Z^1}$ using linear fixed effect regressions. If $P^*_{Z^1}$ is small, wide a priori bounds on D^1_P then map into tight bounds on D^1_m. The second part of Proposition 2 assumes the availability of an additional instrument Z^2 for housing supply, and that both types of households have the same price elasticity of demand. This allows us to point-identify D^1_m.

The most restrictive assumption of Proposition 2 is the assumption that there are only two types of households. Violation of this assumption might potentially lead to biased estimates of social externalities. Note however that under the null of no externalities this assumption is trivially fulfilled. This implies the validity of tests for the presence of social externalities even under violation of the assumption of two types.

Exclusion restrictions of the form assumed in Proposition 2 resemble the use of (randomized) subgroup treatment as a source of identification of peer effects. Compare for instance the general discussion in Moffitt (2004). This idea was used by Duflo and Saez (2003) who provided information about pension plans to a random subset of employees in a random subset of departments of a university, and studied the effect on the behavior of other employees of the same departments.

It is worth highlighting the commonalities and differences between the approach taken here and the one used in Caetano (2009). In both cases, it is assumed that housing demand is a function of both exogenous location amenities and endogenous composition of residents. In both cases, it is also assumed (using my notation) that there is a component of X which is excluded from demand D^c of some subgroup c. The first difference between the two papers is the object of interest: the goal of Caetano (2009) is to identify household preferences for some component of X (school quality) for the group for which this component is *not* excluded (parents). The goal of the approach proposed here is to identify demand slopes with respect to M for the group for which some component of X *is* excluded. The second difference is in the nature of assumptions invoked to achieve identification. Caetano (2009) requires parametric restrictions for identification, but has weaker conditions on the observability of endogenous (compositional) amenities.

3.2. The spatial structure of cities

In the model considered so far, the arguments determining demand and supply and the arguments determining equilibrium outcomes are exactly the same. The identification results in this and the next section are based on model extensions which generate variation in composition conditional on all exogenous arguments of demand and supply. In particular, the model introduced in Section 2 did not allow for the possibility of cross-neighborhood externalities. This section extends this model by adding a spatial structure.

Assume that there are $\mathscr{N}$ neighborhoods, that there are only two types ($\mathscr{C} = 2$), and that the relevant composition variable $\tilde{m}$

[6] For the rest of this paper, we will decompose $X = (Z, W, \epsilon)$. Z are observable components of X, i.e., demand shifters, which are excluded from some demand functions and which will serve as instruments. W comprises other observable components of X, and ϵ is unobservable.

[7] Note that this approach does *not* require that social externalities are more far reaching than the externalities of exogenous amenities. This approach only requires that the *equilibrium effects* of social externalities, mediated by the composition of intermediate neighborhoods, are more far reaching than the *direct* effects of exogenous amenities.

[8] Recall that superscripts for D and M index types of households, superscripts on X or Z index components of exogenous demand shifters, and superscript $*$ marks equilibrium outcomes. Subscripts denote partial derivatives.

affecting demand in each neighborhood is a weighted average of the composition m of adjacent neighborhoods, where $m = M^1/(M^1 + M^2)$; similarly $\widetilde{X}$ is a weighted average of the Xs of adjacent neighborhoods. The weights are given by an $\mathscr{N} \times \mathscr{N}$ matrix $\mathbf{G}$, where the (k,l)th entry of $\mathbf{G}$ describes the strength of externalities from neighborhood l to neighborhood k. Let $\mathbf{m}$ be the $\mathscr{N}$ vector of m for all neighborhoods, $\tilde{\mathbf{m}}$ the vector of $\mathbf{G}$-weighted averages of m, similarly for $\mathbf{X}$ and $\widetilde{\mathbf{X}}$. Then

$$\tilde{\mathbf{m}} = \mathbf{G}\mathbf{m}, \tag{13}$$

$$\widetilde{\mathbf{X}} = \mathbf{G}\mathbf{X}. \tag{14}$$

Assume furthermore that households are indifferent w.r.t. the scale of the neighborhood. We get

$$D^c = D^c(\widetilde{X}, \tilde{m}, P)$$

for $c = 1, 2$. Equilibrium prices P^* and composition M^* in this setup are defined as before, that is for each neighborhood the endogenous outcomes P^* and M^* satisfy the equilibrium conditions

$$D(\widetilde{X}, \tilde{m}^*, P^*) = M^*, \tag{15}$$

$$S(P^*, X) = M^{*1} + M^{*2}. \tag{16}$$

Note that in this model D is constant in components of $(\mathbf{m}, \mathbf{X})$ with a corresponding zero entry in $\mathbf{G}$. D^k does depend on m^l if $G^{kl} \neq 0$.

Proposition 3 (*Spatial identification*). *Maintain the assumptions stated in this subsection, and assume that $S_{Z^1} = 0$ for all neighborhoods and some component Z^1 of X. Fix two neighborhoods k and l. If the k,lth entry of $\mathbf{G}$ equals 0 but $\tilde{m}^k_{Z^{1,l}} \neq 0$ then*

$$D^{c,k}_{\tilde{m}^k} = \frac{1}{\tilde{m}^k_{Z^{1,l}}} \left(M^{*c,k}_{Z^{1,l}} - D^{c,k}_P P^k_{Z^{1,l}} \right). \tag{17}$$

If we assume in addition that $d = d(\tilde{m}, \widetilde{X})$, then

$$d^k_{\tilde{m}^k} = \frac{m^k_{Z^{1,l}}}{\tilde{m}^k_{Z^{1,l}}}. \tag{18}$$

The condition $\tilde{m}^k_{Z^{1,l}} \neq 0$ is guaranteed, in particular, if $d = d(\tilde{m}, \widetilde{X})$, $0 < d_{\tilde{m}} < 1$ and $d_{\widetilde{X^1}} \neq 0$, and there exists a power $j > 1$ of $\mathbf{G}$, such that the k,lth entry of $\mathbf{G}^j$ is not equal to zero.

Proposition 3 again suggests to construct instruments for neighborhood composition in regressions of demand on composition. Suppose there is variation in a variable Z^1 in a neighborhood l which is independent of variation of X in neighborhood k, and which does not directly affect demand in neighborhood k. Then correlation between M^k and $Z^{1,l}$ must reflect the effect of composition on demand, where variation in composition in the vicinity of k is induced by composition changes in neighborhoods between k and l.

It is important to note that this proposition does *not* rely on an assumption of differential ranges of externalities for endogenous composition M and exogenous composition X – in fact, the result could be easily modified to accommodate a longer range of externalities for X than for M. Validity of instruments based on this idea does require, however, that Z^1 is excluded from demand in locations sufficiently far away. Relevance of such instruments requires that externalities of composition do propagate through intermediate neighborhoods. Also, as before, we have assumed that there are only $\mathscr{C} = 2$ types of households. This is restrictive in general, but not under the null hypothesis of no social externalities. Finally, note that this is still a partial equilibrium model which does not consider the effect on outside options of changes in remote locations, even though it does consider the externalities of nearby locations.

The idea of using the spatial structure of cities to identify social externalities in a model of sorting is formally analogous to the use of social network structures to identify endogenous versus exogenous peer effects, as in Bramoullé et al. (2009) and De Giorgi et al. (2010). In the context of social networks, exogenous changes of some sort affecting a person might directly affect her friends, but only indirectly affect her friends' friends, through the change of endogenous outcomes for her friends.

3.3. The dynamics of composition and prices

The models discussed so far are static. We can think of them as describing an economy with negligible search frictions in which equilibrium is instantaneously achieved. Alternatively, they could be considered as describing the long run steady state of an economy with frictions. However, explicitly considering dynamics and frictions reveals additional sources of identification.

Considering models with search frictions is useful, in particular, because they imply delayed adjustment of composition M following changes in exogenous characteristics X. This generates independent variation between X and M, contrary to the static case where M is essentially a function of X. This independent variation allows us, under certain conditions, to separately identify household willingness to pay for X and for M. In particular, under assumptions detailed in the supplementary appendix, the dynamics of composition over time in a given neighborhood can be approximated by the difference equation

$$\Delta M := M^1 - M^0 = \kappa \cdot (D(M^s, Z^s, \overline{P}^s) - M^s), \tag{19}$$

where M^t, Z^t are the composition and exogenous amenities of a neighborhood at time t, and s is an intermediate time between 0 and 1. The average rental price (actual or imputed) for housing in the neighborhood at time t is denoted by $\overline{P}^t = E^t[P]$. $\kappa \in (0,1)$ is a parameter reflecting search frictions. Following a shock to X, the time path of M converges towards a new steady state value M^* at a rate of κ. In the special case where $d = d(m, X)$, the steady state composition m^* satisfies $d(m^*, X) = m^*$.

Furthermore, average prices in the dynamic model reflect the average willingness to pay for housing in the neighborhood. Thus

$$\Delta P := \overline{P}^1 - \overline{P}^0 = E^s\left[-\frac{u_X}{u_P}\right]\Delta X + E^s\left[-\frac{u_m}{u_P}\right]\Delta m, \tag{20}$$

where $u = u(X, m, P)$ is the flow utility for households living in the neighborhood (which might be different for every household and change over time), and s is again an intermediate time between 0 and 1. E^s is an average over households living in the neighborhood at time s. Combining these two results yields

Proposition 4 (*Dynamic identification*). *Assume that $D_{Z^1} \neq 0$. If Eqs. (19) and (20) hold then*

$$E^s\left[-\frac{u_m}{u_P}\right] = \frac{\partial \Delta^2 P / \partial \Delta^1 Z^1}{\partial \Delta^2 m / \partial \Delta^1 Z^1}, \tag{21}$$

for some s in the interval $(1,2)$, where $\Delta^2 P = \overline{P}^2 - \overline{P}^1$, $\Delta^2 m = m^2 - m^1$, and $\Delta^1 Z^1 = Z^{1,1} - Z^{1,0}$. The partial derivatives are understood as derivatives holding constant all other components of $\Delta^1 X$.

Eq. (19) implies that past shocks to X, say between times 0 and 1, affect future changes in composition m, say between times 1 and 2. Eq. (20) implies that shocks to X before time 1 affect price changes after time 1 only through their effect on changes in composition after time 1. Past shocks to Z^1, if they are uncorrelated with future shocks to X, can thus be used as instruments for changes in composition m in hedonic regressions of changes in rental prices P on changes in composition m. Furthermore, if we

do believe that prices adjust quickly, then the short run response in prices to shocks in X can be used to estimate the willingness to pay for X. Past shocks to Z^1 cannot be used as instruments for future composition changes in demand regressions, since changes in demand will be affected by Z^1 due to frictions, even in the absence of social externalities – the necessary exclusion restriction does not hold.

To give some idea of the conditions necessary for Eqs. (19) and (20) to hold, the dynamic model of the local housing market discussed in the supplementary appendix can be summarized as follows. In this model, there is an explicit, continuous time dimension, and exogenous location characteristics X can change over time. Households that would like to move to a different neighborhood are subject to search frictions. If they decide to search for a new home, offers arrive at Poisson rate λ. Similarly, owners of vacant units have to search for tenants and find them at rate μ. Households are maximizing expected discounted utility, and make their search decisions in a forward looking way. Due to search frictions, composition M changes continuously over time and only reacts with delay to shocks in X. Finally, once a match is formed between homeowner and household, they are in a situation of bilateral monopoly: By breaking the match they both would have to search again, and thereby incur a loss of utility. Therefore, they have to negotiate over the division of the surplus, and rents are match-specific. This model builds on a well established literature in labor economics which discusses the dynamics and comparative statics of unemployment and wages in models with search frictions. The central presumption of this literature is that finding a job or an employee takes time and unemployment is due to this search time. Pissarides (2000) provides an extensive overview of this literature. Wheaton (1990) applies the insights of this literature to the housing market. The focus of either of these is the relationship between vacancies (unemployment) and prices. Wheaton (1990) in particular models housing vacancies as corresponding to the search time of households who decide to move due to lifecycle events (shocks), find another place and then attempt to sell their old home.

4. Application to data on cities in the United States

In this section we use our identification results to shed some light on the causes of urban segregation in the United States. We first focus on the impact of Hispanic share in a neighborhood on housing demand of both Hispanics and non-Hispanics in that neighborhood. Various instruments based on the ideas developed in Section 3 allow us to credibly identify this impact. We then generalize to investigate the impact of neighborhood composition in terms of other dimensions of ethnicity, education and income. This generalization provides a richer picture, but depends on somewhat less credible sources of variation.

The next section provides a description of the dataset used, and discusses sample selection as well as variable construction. Section 4.2 proposes various instruments which are motivated by the theoretical considerations of Section 3. Section 4.3 provides estimates of demand slopes and hedonic slopes using these instruments. These estimates consistently suggest a large positive dependence of demand of Hispanics on Hispanic share and a large positive dependence of demand of non-Hispanics on non-Hispanic share. The estimates of price slopes with respect to Hispanic share imply a moderately negative average willingness to pay for Hispanic share. Section 4.4 checks the robustness of these results by applying the estimators to various subsamples and different housing cost variables. Section 4.5 considers other dimensions of neighborhood composition, including the share of other ethnic groups, education, and incomes. Consistently we find economically and statistically significant own-group preferences, which appear to be strongest for Asian/Pacific Islanders.

Table 3
Summary statistics.

	1980	1990	2000
Population	3413	3899	4416
	(1822)	(1802)	(2205)
Hispanic	.08	.10	.14
	(.15)	(.18)	(.21)
Black	.13	.15	.17
	(.25)	(.26)	(.27)
Asian/Pacific Islander	.02	.04	.06
	(.06)	(.08)	(.10)
Non-Hispanic white	.77	.71	.63
	(.29)	(.31)	(.32)
Some college	.35	.42	.47
	(.18)	(.18)	(.19)
Average family income (in 1000$)	24.2	47.2	68.0
	(10.0)	(24.1)	(36.3)
Median rent	274	532	711
	(95)	(190)	(285)
Mean imputed rent	276	528	695
	(34)	(163)	(207)
Median value		119,240	159,513
		(89,999)	(124,759)

Notes: This table shows averages and standard deviations over all neighborhoods in the sample used for our analysis. Note that the means shown are unweighted averages across neighborhoods, so that for instance average Hispanic share is not equal to the population share of Hispanics.

4.1. The data

The data set used is an extract from the Neighborhood Change Database (NCDB) which aggregates US census variables to the level of census tracts. Tract definitions are changing between census waves but the NCDB matches observations from the same geographic area over time, thus allowing us to observe the development over several decades of the universe of US neighborhoods.

4.1.1. Sample construction

The sample is selected as follows.[9] All rural tracts are dropped, and so are all metropolitan statistical areas (MSAs) with fewer than 100 tracts, all tracts with population below 200 and tracts that grew by more than 5 standard deviations above and beyond the MSA mean. This leaves neighborhoods from the 114 largest MSAs in the sample. The definition of MSA used is the MSAPMA from the NCDB, which is equal to "Primary Metropolitan Statistical Area" (PMA) if the tract lies in one of those, and equal to the MSA it lies in otherwise. Our sample, constructed in this way, contains 40,030 tracts.

Table 3 shows some summary statistics for each of the three waves in this sample. The average tract (neighborhood) in this sample has a population of about 4000 individuals, and a Hispanic share of about 10 percent. Both average neighborhood size and Hispanic share are increasing over time. Note that average Hispanic share is not equal to Hispanic share in the sample population, since the average is not weighted by neighborhood size, and similarly for all other variables.

4.1.2. Imputation of rents

Three measures of housing prices are used, median reported rents, median reported values, and an "imputed rent" variable created by myself. Imputed rents are calculated as a share-weighted average of rents imputed from housing values and reported rents. Rents are imputed from housing values as the predicted rents from cross-sectional OLS regressions of housing values on rents in each decade. This imputation method can be justified as follows. Let r be

[9] This replicates the sample construction of Card et al. (2008), who use the same dataset.

the relevant discount rate for a household, P the rental price for a given type of housing, and $\mathbf{P}$ the market value for ownership of the same type of housing. Then households are indifferent between renting and ownership if

$$P = r\mathbf{P} - \dot{\mathbf{P}},$$

where $\dot{\mathbf{P}}$ denotes the expected value appreciation of a housing unit. Assume that the expected value appreciation $\dot{\mathbf{P}}$ is uncorrelated with baseline value $\mathbf{P}$, and that both ownership and renting are observed. Then the slope of an OLS regression of P on $\mathbf{P}$ provides an estimator of the average discount rate of households which are indifferent between ownership and renting. The predicted value of such a regression provides an estimator of the average rental value of owner occupied housing in a neighborhood.

Table 3 shows that the nominal increases of reported rents, median reported values, and imputed rents over the two decades under consideration are roughly similar. Median house values are missing from the 1980 census data, and are thus not available for part of our analysis. The dispersion across neighborhoods is higher for median observed rents than for imputed mean rents.

4.2. Instruments

We will use three instruments, motivated by the theoretical considerations of Section 3. We first define the three instruments. We then discuss their relationship to our theoretical results, as well as potential threats to their validity. We will focus on estimating the dependence of the local housing demand of Hispanics and non-Hispanics on Hispanic share, as well as the (weighted average marginal) willingness to pay for composition. Other dimensions of neighborhood composition will be considered below in Section 4.5. Hispanics are denoted by $c = 1$ and non-Hispanics by $c = 2$.

4.2.1. Subgroup instrument

Let $\tilde{c}$ denote the country of origin for Hispanic migrants, where we only consider migrants from Mexico, Puerto Rico and Cuba. $M^{\tilde{c}}$ is the population of type $\tilde{c}$ in a neighborhood in the first year of a given decade. Denote by $M^{\tilde{c},tot}$ the total initial population of type $\tilde{c}$ summed over all neighborhoods, and let $\Delta M^{\tilde{c},tot}$ be the decadal change of $M^{\tilde{c},tot}$. The instrument dZ^1 is defined as

$$dZ^1 = \frac{1}{M^1 + M^2} \sum_{\tilde{c}} M^{\tilde{c}} \cdot \frac{\Delta M^{\tilde{c},tot}}{M^{\tilde{c},tot}}. \tag{22}$$

This is the change in Hispanic share that a neighborhood would experience if growth of each group $\tilde{c}$ in a given neighborhood was equal to the national average.

4.2.2. Spatial instrument

Denote the average predicted shift dZ^1 of Hispanic demand in neighborhoods that are at least 3 km away, but among the 15 closest neighborhoods, by dZ^2,

$$dZ^{2,k} = \frac{1}{n^{k,l}} \sum_{l} dZ^{1,l}. \tag{23}$$

This equation defines the value of the instrument for neighborhood k, and averages over the appropriate neighborhoods l. We furthermore denote by $\tilde{m}$ the average Hispanic share in the given neighborhood and its 4 closest adjacent neighborhoods (tracts).[10]

[10] To calculate dZ^2 and $\tilde{m}$ we need to construct a measure of distance between neighborhoods as follows. For each census tract, the Neighborhood Change Database reports latitude α and longitude β of an interior point. Distance between neighborhoods is defined, based upon these coordinates, as the Euclidean distance between the corresponding coordinates in $\mathbb{R}^3$, which are given by $6371 \cdot (\cos(\alpha)\cdot\cos(\beta), \cos(\alpha)\cdot\sin(\beta), \sin(\alpha))$. Here 6371 is taken to be the radius of earth in kilometers.

4.2.3. Dynamic instrument

Let dZ^3 be the decadal change in m (Hispanic share in a neighborhood), lagged by a decade,

$$dZ^3 = m^{t-1} - m^{t-2}. \tag{24}$$

The variable dZ^3 is used as an instrument for $\Delta m = m^t - m^{t-1}$ in regressions of $\Delta P = P^t - P^{t-1}$ on Δm, controlling for m^t.

4.2.4. Discussion

The *first instrument* (subgroup instrument) for change in composition m which we propose, dZ^1, is motivated by Proposition 2. The idea of this instrument is to construct a local predictor of the change in housing demand of Hispanics induced by immigration. This is a synthetic instrument (or "Bartik-instrument") similar to the one used by Card (2001) (and others) on the MSA level as a predictor of changes in labor supply. It is predictive for local changes in Hispanic share if new immigrants from the same source countries have a similar distribution of preferences as prior migrants, whether the preference is for exogenous location characteristics or for the presence of their compatriots.

Formally, in order to use dZ^1 as an instrument for composition m in estimation of D^2_m, we need dZ^1 to satisfy the conditions (i) $D^2_{Z^1} = 0$ (exclusion), (ii) $m_{Z^1} \neq 0$ (relevance) and (iii) dZ^1 has to be uncorrelated with counterfactual changes in M^2 (exogeneity/randomness). The instrument dZ^1 is excluded from the demand of type $c = 2$ (condition (i) holds) if there is no causal effect of total Hispanic immigration (across all MSAs) on demand of type 2, i.e., non-Hispanics. This might not hold if the set of outside options is affected by immigration, i.e., if there are general equilibrium effects. All our regressions control for MSA $\times$ time fixed effects, however, absorbing any city-wide shifts in outside options. They furthermore control flexibly for initial Hispanic share, so that identification is driven by variation in the composition of initial Hispanic population of a neighborhood in terms of country of origin, conditional on Hispanic share.

We also need the instrument dZ^1 to be independent of changes in unobserved factors affecting demand of the non-Hispanics (so that condition (iii) holds). A potential threat to validity here would be some delayed adjustment due to frictions, which could imply a correlation between current composition and future adjustment of non-Hispanic population. Instrument relevance (condition (ii)) can be easily checked empirically and is not an issue with our data.

The *second instrument* (spatial instrument) for change in composition m which we propose, dZ^2, is motivated by Proposition 3. The idea of this instrument is to construct a predictor of local composition based on predicted changes in composition in neighborhoods more than 3 km away. These predicted changes should affect local demand only indirectly, through their effect on composition in intermediate neighborhoods.

In order to use dZ^2 as an instrument for composition $\tilde{m}$ in estimation of $D^2_{\tilde{m}}$, in the context of the spatial model of Section 3.2, we need dZ^2 to satisfy the conditions (i) $D_{Z^2} = 0$ (exclusion), (ii) $\tilde{m}_{Z^2} \neq 0$ (relevance), and (iii) dZ^2 has to be uncorrelated with counterfactual changes in M (exogeneity/randomness). Furthermore, we need to assume that the composition variable which *does* matter for households' location choices is $\tilde{m}$. The regressions using dZ^2 as an instrument for $\tilde{m}$ will control for dZ^1, and thus use variation in composition orthogonal to the one used in the subgroup approach. It seems plausible that the exclusion restrictions are satisfied (condition (i) holds) in the case of predicted immigration for neighborhoods at a certain distance conditional on local predicted immigration. The assumption that $\tilde{m}$, the average of m for 5 adjacent neighborhoods, is the relevant composition variable is somewhat arbitrary. The results are robust to different specifications of $\tilde{m}$, however, as we will show.

Our *third instrument* (dynamic instrument) for changes in composition, dZ^3, is motivated by the arguments of Section 3.3. The idea is to use past composition changes in a neighborhood as instruments for future composition changes in hedonic regressions of rental prices on composition. This is justified if intertemporal correlation in composition changes reflects incomplete adjustments due to search frictions.

In the context of the dynamic model, past changes in m are predictive of future changes if they reflect incomplete adjustments to past shocks in X. As we argued, under certain conditions any shocks to X are quickly incorporated into prices P according to household willingness to pay. Due to search frictions, however, composition m only adjusts with delay, with prices following accordingly. Past changes in m are a valid instrument for future changes in m in hedonic regressions if and only if they are uncorrelated with future changes in X (exogeneity). We shall make the strong identifying assumption that this holds true, conditional on current Hispanic share m. The main threat to the validity of this assumption would be anticipated changes in amenities X that are reflected in past composition changes. Exclusion of dZ^3 conditional on m is immediate in our model, and instrument relevance is again easily checked empirically.

A strong assumption necessary for a structural interpretation of the slopes estimated in the next section is that $\mathscr{C} = 2$ and that the relevant type variable is Hispanic origin. The assumption of two types is trivially fulfilled under the null hypothesis of no social externalities, so all tests for the presence of social externalities remain valid even if $\mathscr{C} \neq 2$. Put differently, if we assume that demand does not depend on neighborhood composition, then it is not restrictive to additionally assume that demand depends on neighborhood composition only through Hispanic share.[11] This assumption will be relaxed in Section 4.5 below.

Figs. 3 and 4 illustrate the geographic variation both in our dependent variables and in the instruments used; we chose Suffolk county (which includes Boston) and San Francisco for illustration. Visual inspection suggests a positive correlation of both the subgroup-instrument and the spatial instrument with the realized change in Hispanic share. Visual inspection suggests furthermore that these two instruments seem fairly uncorrelated and predict variation in Hispanic shares for different sets of neighborhoods.

4.3. Results

The dependent variables we consider throughout this section are the log populations (interpreted as demand) of non-Hispanics and Hispanics in a neighborhood, as well as log housing costs. For the first stage regressions, the dependent variable is Hispanic share in a neighborhood.

Table 4 shows a number of "naive" hedonic and demand regressions. These regressions would give consistent estimates of D_m and $-E_m/E_P$ only (i) absent omitted variables and (ii) absent the endogeneity of composition in the presence of social externalities motivating the present paper. Clearly, problems of omitted variable bias are severe as demand is increasing in prices for all the demand regressions shown, suggesting price variation is driven by fluctuations in demand due to variation in omitted factors X. Taken at face value, these regressions would furthermore suggest a negative preference of non-Hispanics for Hispanic share, and a strong positive preference of Hispanics for Hispanic share, as well as a small (positive or negative) willingness to pay for Hispanic share.[12]

[11] Restrictive models are often valid under the null hypothesis of interest; Graham (2008a), for instance, relies on the fact that linear-in-means models of social interactions are valid if there are no social interactions.

[12] An estimated coefficient of -0.5 would imply that an increase of Hispanic share by 1 percentage point would lead to a decrease of housing prices by 0.005 percent.

4.3.1. Main results using our preferred specifications

We now turn to our main empirical results, involving instrumental variable regressions of demand and of prices on composition. We then discuss these results in the context of the theoretical arguments made in Section 3. All our regressions are run in decadal differences, pooling changes over the 1980s and 1990s. They control for MSA $\times$ time fixed effects. They furthermore control for neighborhood and decade specific initial conditions – the subgroup and dynamic instrument regressions control for initial Hispanic share and its square, the spatial instrument regressions control for predicted immigration.[13] The results of our preferred empirical specifications are shown in Table 5. The theoretical interpretations of the entries of Table 5 are summarized in Table 6. Instrumental variable regressions which are not theoretically meaningful are omitted from these tables.

As can be seen in the first column of Table 5, the instrument is a highly significant predictor of the change in local composition for all three specifications. The t-statistics for first stage significance for the three instruments are equal to about 9, 17, and 18 respectively, corresponding to F-statistics (given by the square of the t-statistic) of 80 and higher. Strength of the instruments is therefore not an issue. The instrumental variables regressions consistently suggest that (i) there is a strong negative dependence of non-Hispanic housing demand on Hispanic share, (ii) there is a positive dependence of Hispanic housing demand on Hispanic share, and (iii) there is a negative but small average marginal willingness to pay for Hispanic share.

4.3.2. Correcting for the endogeneity of prices

In interpreting the slopes of demand on Hispanic share, we have to take care to correct for the price effect of changing Hispanic share in order to obtain structural slopes of demand. This is reflected in the bias terms of the form $D_P^2 \frac{P_Z^*}{m_Z^*}$ in Table 6. Given the price slope of demand D_P^2, the size of the bias is proportional to $\frac{P_Z^*}{m_Z}$ as estimated by the IV regressions in the last column of Table 5. Luckily $\frac{P_Z^*}{m_Z}$ appears to be fairly small for our instruments, implying that the size of the bias does not depend very much on the exact value of D_P^2. If we assume that the elasticity of non-Hispanic demand with respect to rents is between 0 and 2,[14] and taking into account that the IV regressions of P on m yield coefficients of around -0.5, this implies a bias of around 0 to 1. Subtracting this bias yields estimates of D_m^2 of -6.3 to -9.4. For Hispanics, the estimate based on the spatial instrument implies a positive dependence of demand on Hispanic share. Correcting again for the rent-bias, we get an estimate of D_m^1 of around 2.4 to 3.4. Given that our estimated price slopes are small, the results are fortunately not very sensitive with respect to price elasticities. The IV regressions of prices on Hispanic share, using the spatial and dynamic instruments, yield moderately negative estimates of $-E_m/E_P$ of -0.75 and -0.52. This implies a moderately negative average marginal willingness to pay for Hispanic share.

4.3.3. Comparison to naive regressions

These results are remarkably consistent across instruments. While we might have doubts about the validity of each of the

[13] We report results for linear estimators only. This might seem to stand in contrast to the nonparametric nature of the identification results discussed in Section 3. Our view is that we should make sure that identification does not rely on functional form assumptions. Once that is guaranteed, it is however useful to use simple and easily interpretable estimators such as linear OLS, if they correspond to nonparametrically meaningful objects.

[14] There is a large older literature estimating the price elasticity of housing demand, see for instance Polinsky (1977); Hanushek and Quigley (1980); and Ermisch et al. (1996). It is not clear that the estimates from this literature extrapolate to the present setting, given the differences in geographic unit, time horizon, and historical period considered. However, it is still reassuring that all the estimates from this literature comfortably fit into the range assumed here.

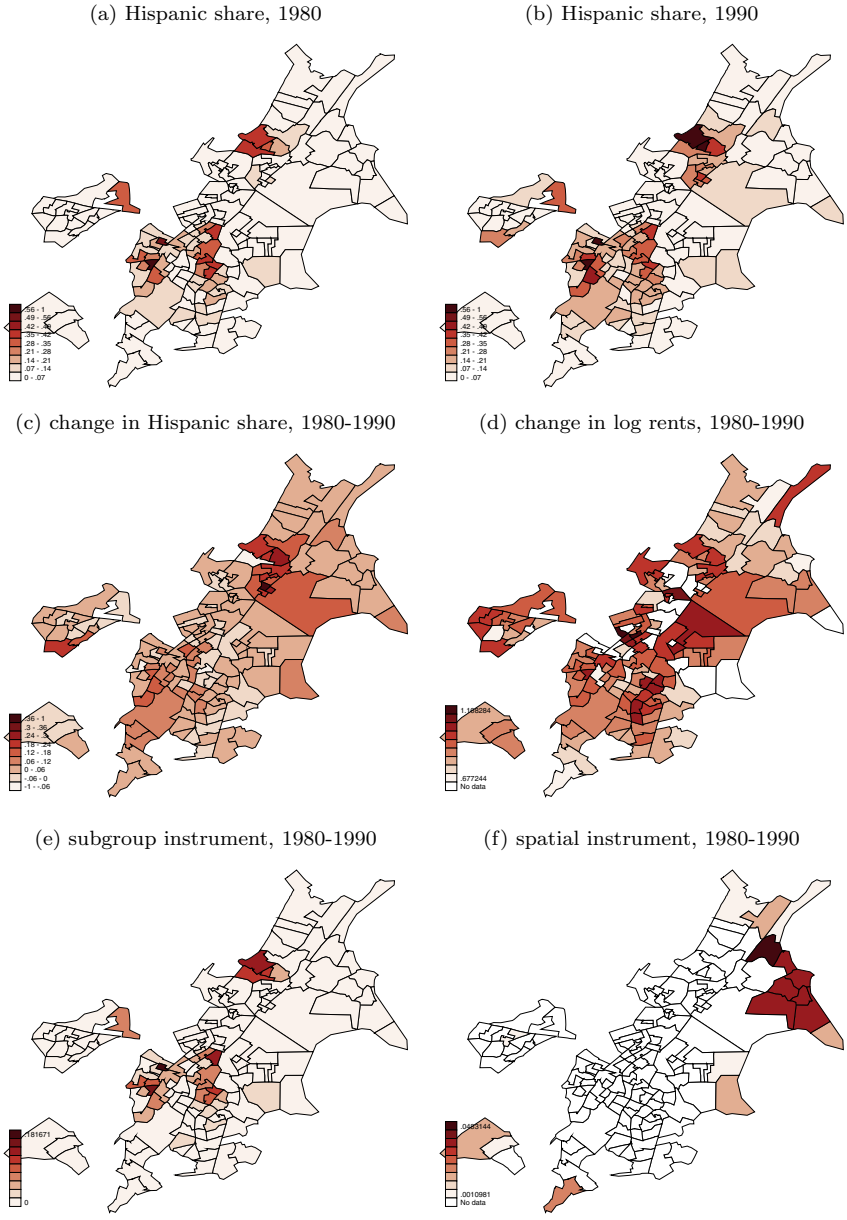

Fig. 3. Suffolk county (Boston) 1980–1990.

instruments, they do rely on different assumptions and use orthogonal variation in the data, so that this consistency might add to the credibility of the results. Finally, let us compare these results to those using "naive" regressions, as shown in Table 4. Consistently across specifications, it seems that the naive estimates of D_m^1, D_m^2 are strongly upward biased, and the estimates of $-E_m/E_P$ are moderately upward biased. This also holds true for the specifications in differences controlling for initial Hispanic share and MSA $\times$ time fixed effects. One interpretation of this result might be that Hispanic location decisions were more "pro-cyclical" relative to non-Hispanics, i.e., Hispanic demand reacted more strongly to unobserved shocks in X.

4.4. Robustness checks

4.4.1. Replication for subsamples

The regressions of the previous section used the full sample of the 114 largest MSAs in the United States, pooling the data for changes in the 80s and in the 90s. We can check the robustness of the results by replicating the regressions on subsamples. In particular, Table 7 presents estimates for the subset of MSAs with Hispanic shares larger than 8% in 2000. This corresponds to roughly 50% of the sample. Furthermore, there might be concerns about the effect of rent controls. Table 8 replicates the regressions on the sample of MSAs excluding California and the state of New York, where rent controls might play some role. Table 9 shows estimates for the 80s and for the 90s separately. This table also uses median rents and median reported house values as alternative housing cost variables.

The results are largely consistent with those obtained previously, with a few exceptions. First, in the sample of MSAs with large Hispanic shares, Hispanics seem less responsive in their location decision to the Hispanic share of a neighborhood. Second, in the sample excluding California and New York, price responses seem somewhat stronger. This might indicate a certain role of rent controls. Finally, in this sample the subgroup instrument is quite

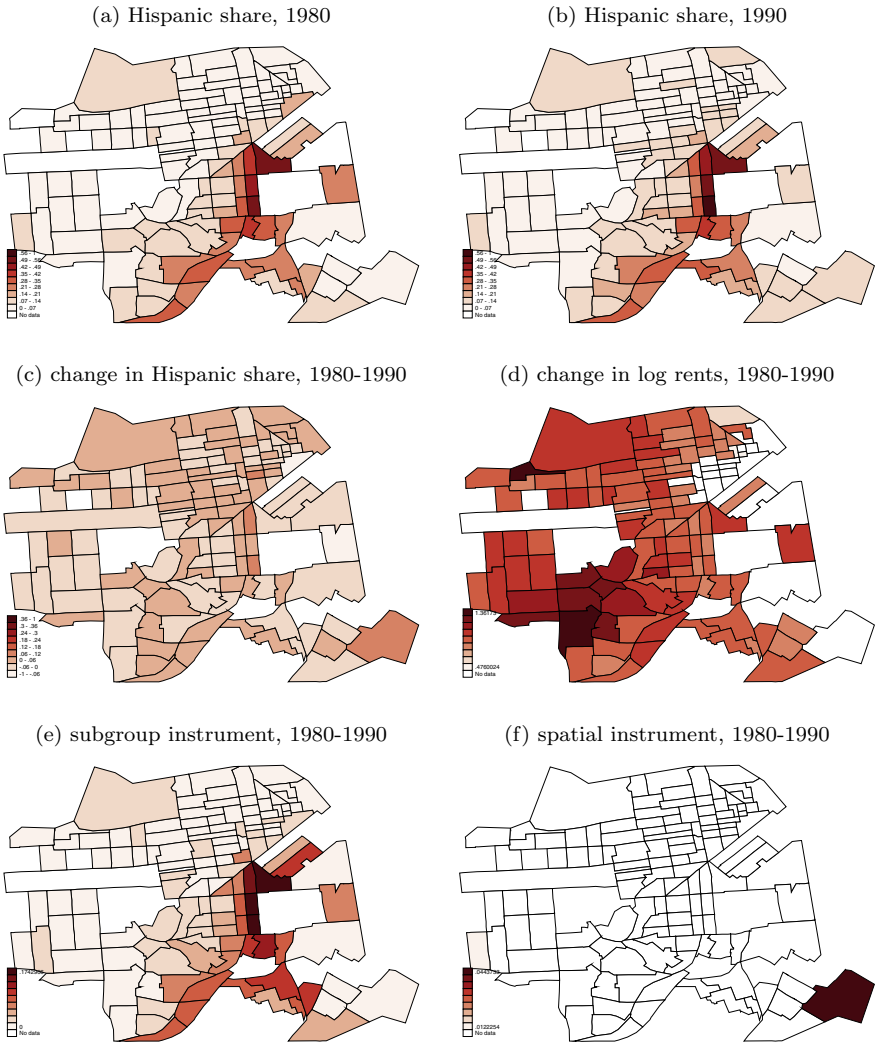

Fig. 4. San Francisco 1980–1990.

Table 4
Naive hedonic and demand regressions.

	(1) Log non-Hisp pop	(2) Log Hisp pop	(3) Log mean imputed rent
Cross-section			
Hisp shr	−1.815	5.616	−0.476
	(0.023)	(0.039)	(0.005)
Log mean imputed rent	0.117	0.198	
	(0.014)	(0.031)	
Differences			
Hisp shr	−1.674	5.946	−0.321
	(0.025)	(0.064)	(0.008)
Log mean imputed rent	0.398	−0.278	
	(0.014)	(0.015)	
Differences with controls			
Hisp shr	−1.681	7.433	0.293
	(0.027)	(0.076)	(0.009)
Log mean imputed rent	0.378	0.555	
	(0.014)	(0.037)	

Notes: This table shows demand regressions of log non-Hispanic population and log Hispanic population on Hispanic share and log mean imputed rents, as well as hedonic regressions of log mean imputed rents on Hispanic share. The first specification is a pooled cross-sectional regression using data from 1980, 1990 and 2000, the second and third are regressions in decadal differences for the 1980s and 1990s. All regressions control for MSA × time fixed effects, the third specification additionally for initial Hispanic share and its square.

weak, and the corresponding estimate of D_m^2 very high with a very large standard error. The different housing cost variables behave in a roughly similar way.

4.4.2. Different spatial cutoffs

Table 10 replicates the results from the regressions using the spatial instrument in Table 5. This table shows results using different specifications of the relevant composition variable $\tilde{m}$, where $\tilde{m}$ is taken alternatively to be the average composition of the given neighborhood and its five closest neighboring tracts, the average composition of the given neighborhood and its three closest neighboring tracts, or a weighted average of the composition of the given neighborhood and its five closest neighboring tracts, where the latter gives weight 1 to the neighborhood itself and weight .25 to its neighboring tracts. This table also shows results using an alternative instrument, using the average of dZ^1 for neighborhoods that are 2 km or more away. As we can see from this table, the results are quite robust to the specific choices of $\tilde{m}$ and the instrument. This is comforting given the somewhat arbitrary nature of the spatial cutoffs we had to choose.

4.5. Other dimensions of neighborhood composition

So far we considered the impact of Hispanic share on the demand of both Hispanics and non-Hispanics. We used three sources of variation, which arguably satisfy the relevant exclusion

Table 5
Instrumental variable estimates, decadal changes in the 1980s and 1990s.

(1) First stage	IV regressions		
	(2) Log non-Hisp pop	(3) Log Hisp pop	(4) Log mean imputed rent
Subgroup instrument			
0.146	−8.360	–	–
(0.016)	(0.740)		
Spatial instrument			
0.119	−6.251	3.437	−0.758
(0.007)	(0.620)	(0.733)	(0.119)
Dynamic instrument			
0.198	–	–	−0.516
(0.011)			(0.049)

Notes: This table shows instrumental variables regressions of the change in log non-Hispanic population, log Hispanic population, and mean imputed rent on the change in Hispanic share using the instruments discussed in the text. All regressions pool data for the 1980s and the 1990s and control for time × MSA fixed effects. The subgroup and dynamic instrument regressions control for initial Hispanic share and its square, the spatial instrument regressions control for predicted immigration.

Table 6
Theoretical interpretation of the entries of Table 5.

(1) First stage	IV regressions		
	(2) Log non-Hisp pop	(3) Log Hisp pop	(4) Log mean imputed rent
Subgroup instrument			
$m^*_{z^1}$	$\frac{M^{*2}_{z^1}}{m^*_{z^1}} = \mathbf{D^2_m} + D^2_P \frac{P^*_{z^1}}{m^*_{z^1}}$	–	–
Spatial instrument			
$\tilde{m}_{z^2}$	$\frac{M^{*2}_{z^2}}{\tilde{m}_{z^2}} = \mathbf{D^2_m} + D^2_P \frac{P^*_{z^2}}{\tilde{m}_{z^2}}$	$\frac{M^{*1}_{z^2}}{\tilde{m}_{z^2}} = \mathbf{D^1_m} + D^1_P \frac{P^*_{z^2}}{\tilde{m}_{z^2}}$	$\frac{P_{z^2}}{\tilde{m}_{z^2}} = -\mathbf{E_m}/\mathbf{E_P}$
Dynamic instrument			
m_{z^3}	–	–	$\frac{P_{z^3}}{m_{z^3}} = E\left[-\frac{\mathbf{u_m}}{\mathbf{u_P}}\right]$

Notes: This table shows the theoretical interpretations of the first stage and instrumental variable coefficients displayed in Table 5. The regression coefficients estimate weighted averages of the slopes shown here.

Table 7
Subsample of MSAs with large Hispanic share – instrumental variable estimates.

(1) First stage	IV regressions		
	(2) Log non-Hisp pop	(3) Log Hisp pop	(4) Log mean imputed rent
Subgroup instrument			
0.146	−8.262	–	–
(0.016)	(0.742)		
Spatial instrument			
0.114	−6.419	0.651	−0.760
(0.007)	(0.677)	(0.762)	(0.128)
Dynamic instrument			
0.210	–	–	0.210
(0.011)			(0.011)

Notes: This table replicates Table 5 for the subset of cities with Hispanic shares larger than 8% in 2000, which corresponds to roughly 50% of the neighborhoods in the full sample.

and randomization assumptions. We further made the assumption that the relevant dimension of neighborhood composition is

Table 8
Subsample excluding California and New York – instrumental variable estimates.

(1) First stage	IV regressions		
	(2) Log non-Hisp pop	(3) Log Hisp pop	(4) Log mean imputed rent
Subgroup instrument			
0.043	−33.575	–	–
(0.024)	(17.116)		
Spatial instrument			
0.122	−8.257	5.513	−1.021
(0.010)	(0.891)	(1.046)	(0.163)
Dynamic instrument			
0.171	–	–	−0.981
(0.016)			(0.092)

Notes: This table replicates Table 5 for the subset of cities outside the states of California and New York.

Hispanic share. This assumption is without loss of generality under the null-hypothesis of no social effects, so that our estimates provide a valid test even in the case of preferences depending on additional features of neighborhood composition. This assumption is restrictive under the alternative of social externalities. We need to interpret our estimated slopes as "reduced form" effects of the induced changes in neighborhood composition on demand in that case, if our instruments do indeed change the composition *within* the groups of Hispanics and non-Hispanics.

4.5.1. Additional dependent variables

In this section we explore our data further, studying other dimensions of neighborhood composition. We start by replicating our baseline specifications, considering additional dependent variables. Table 11 is based on the same specifications as Table 5, using our subgroup instrument as well as the spatial instrument. The table shows the estimated slopes of demand of non-Hispanic, Hispanic, Black, and Asian/Pacific Islander residents, as well as the impact of Hispanic share on the average income of residents as well as on the share with a college degree.[15] As our subgroup instrument is not excluded from Hispanic demand, as well as from average income and share with a college degree, the corresponding entries are omitted from Table 11.

The results based on the subgroup instrument do suggest some heterogeneity in preferences among non-Hispanics for Hispanic share, with the estimated slope of non-Hispanic Whites being the largest. The spatial instrument yields results which are similar across the groups considered – on a similar order of magnitude and not statistically significantly different across non-Hispanic Whites, Blacks, and Asian/Pacific Islanders. Education level and incomes are declining in Hispanic share, which is unsurprising given the differences in education levels between Hispanics and non-Hispanics. Taken together these results suggest that induced changes in the relative composition of non-Hispanics might play a limited role in determining our estimated coefficients, but that our estimates are probably fairly close to true structural slopes of demand with respect to Hispanic share.

4.5.2. Subgroup instruments based on MSA-level variation

One of the reasons our main empirical analysis has focused on Hispanic shares is that there is significant variation in nation-wide immigration levels across different countries of origin for Hispanics, and countries of origin are available for them in the NCDB. Using this nation-wide variation allowed us to run IV regressions controlling for MSA and time fixed effects.

[15] These slopes need again to be corrected for price effects, as before.

Table 9
Decades separately, different housing price variables - Instrumental Variable estimates.

Sample	(1) First stage	IV regressions, dependent variable is log-				
		(2) Non-Hisp pop	(3) Hisp pop	(4) Mean imputed rent	(5) Median rent	(6) Median house value
1980s	*Subgroup instrument*					
	0.146	−9.515	–	–	–	–
	(0.027)	(1.171)				
	Spatial instrument					
	0.093	−8.673	3.181	−1.395	0.293	NA
	(0.008)	(1.147)	(1.218)	(0.221)	(0.331)	
	Dynamic instrument					
	0.181	–	–	−0.092	0.008	NA
	(0.015)			(0.085)	(0.130)	
1990s	*Subgroup instrument*					
	0.285	−4.67055	–	–	–	–
	(0.024)	(0.475)				
	Spatial instrument					
	0.16	−4.053	3.665	−0.134	−0.429	−0.757
	(0.012)	(0.639)	(0.853)	(0.108)	(0.236)	(0.207)
	Dynamic instrument					
	0.254	–	–	−0.343	−0.606	−0.575
	(0.016)			(0.049)	(0.098)	(0.147)

Notes: This table replicates Table 5 for the 1980s and 1990s separately, and includes alternative measures of housing costs as dependent variables. For 1980, median house value is not available.

Table 10
Spatial instrument, different spatial cutoffs - Instrumental Variable estimates.

Composition of	(1) First stage	IV regressions		
		(2) Log non-Hisp pop	(3) Log Hisp pop	(4) Log mean imputed rent
Instrument using tracts more than 3 km away				
5 closest tracts	0.096	−6.251	3.437	−0.758
	(0.006)	(0.620)	(0.733)	(0.119)
3 closest tracts	0.088	−6.579	3.617	−0.801
	(0.007)	(0.653)	(0.781)	(0.127)
5 tracts, weighted	0.092	−6.329	3.479	−0.776
	(0.007)	(0.598)	(0.741)	(0.120)
Instrument using tracts more than 2 km away				
5 closest tracts	0.098	−6.126	3.490	−0.770
	(0.006)	(0.461)	(0.534)	(0.091)
3 closest tracts	0.094	−6.223	3.545	−0.782
	(0.006)	(0.475)	(0.538)	(0.092)
5 tracts, weighted	0.099	−5.947	3.388	−0.751
	(0.006)	(0.433)	(0.505)	(0.087)

Notes: This table replicates the spatial instrument regressions of Table 5, using different definitions of the regressor $\bar{m}$ and the instrument dZ^2.

In order to get similar sources of variation for other subgroups, we can employ the same construction using MSA-level rather than national variation in the growth of these other groups. For each of the groups of Hispanics, non-Hispanic Whites, Blacks, and Asian/Pacific Islanders we construct an MSA-based subgroup instrument interacting the MSA level growth of this subgroup with the share of this subgroup in a given neighborhood.

Table 12 shows the results of instrumental variable regressions of each group's demand on own-group share, instrumenting with the three MSA-based subgroup instruments for the *other* three groups (which are arguably excluded from a given group's demand), and controlling for the shares of all four groups in the given neighborhood, as well as for MSA-level population growth. In this construction, we cannot control for MSA × time fixed effects, since our identifying source of variation (we do control for initial group shares), comes precisely from variation across MSAs and time in the relative growth of different groups. Not being able to control for such fixed effects opens up the possibility that the estimated slopes might partly be driven by variation in outside options; on the other hand this approach allows us to consider richer variation in neighborhood composition beyond Hispanic share. The variation in outside option is likely to bias our results toward 0, leading us to underestimate social preferences.

This not-withstanding, the estimates in Table 12 consistently suggest a strong own-group preferences with estimates close to those obtained using the MSA-based subgroup instruments, which seem largest for Asian/Pacific Islanders, and smallest for Blacks.

Table 11
Baseline specifications, additional dependent variables.

	(1) Log non-Hisp pop	(2) Log Hisp pop	(3) Log non-Hisp white pop	(4) Log black pop	(5) Log Asian and Pacific Isl pop	(6) Log some college pop	(7) Average income
Subgroup instrument							
Reduced form	−1.218	–	−1.253	−0.201	−0.503	–	–
	(0.099)	–	(0.108)	(0.183)	(0.203)	–	–
IV	−8.360	–	−8.597	−1.381	−3.452	–	–
	(0.740)	–	(0.901)	(1.241)	(1.390)	–	–
Spatial instrument							
Reduced form	−0.745	0.410	−0.701	−1.042	−0.803	−0.557	−14,060
	(0.071)	(0.088)	(0.084)	(0.142)	(0.156)	(0.057)	(1962)
IV	−6.251	3.437	−5.881	−8.738	−6.731	−4.704	−117,916
	(0.620)	(0.733)	(0.708)	(1.257)	(1.338)	(0.514)	(16,347)

Notes: This table replicates the regressions of Table 5, considering additional left hand side variables. Coefficients which are not meaningful in the context of our theoretical model are omitted.

Table 12
MSA-based subgroup instruments.

(1) Log Hisp pop	(2) Log non-Hisp white pop	(3) Log black pop	(4) Log Asian and Pacific Isl pop
OLS			
7.278	2.219	6.147	17.000
(0.084)	(0.033)	(0.068)	(0.291)
Subgroup instrument			
10.165	5.976	0.927	28.795
(0.412)	(0.463)	(0.774)	(1.706)

Notes: This table shows OLS and IV regressions of the change in housing demand for four groups on the change of own-group share, controlling for the shares of all four groups in the given neighborhood, as well as for MSA-level population growth. The instrumental variable regressions instrument own-group share, with the three MSA-based subgroup instruments for the other three groups, as described in the text.

4.5.3. Spatial instruments based on MSA-level variation

We next employ the same construction as before to get an instrument justified based on spatial exclusion restrictions rather than based on subgroup-exclusion restrictions: The spatial instrument for each subgroup is the average predicted shift (using MSA-level variation) of this group's demand in neighborhoods that are at least 3 km away, but among the 15 closest neighborhoods. The relevant composition variable in this setting is assumed to be the share of a given group in the given neighborhood and its 4 closest adjacent neighborhoods (tracts).

Table 13 shows the resulting estimates. The top part of this table displays reduced form coefficients regressing changes in demand for each group on the excluded instruments for all groups, controlling for predicted changes in the given neighborhood for all groups. The bottom part of Table 13 shows the corresponding instrumental variables estimates, instrumenting own group share with the spatially excluded instruments for all four groups, and controlling again for predicted changes in the given neighborhood for all groups.

Both the reduced form and the instrumental variables estimates once again consistently suggest a strong own-group preference. And again this preference seems strongest for Asian/Pacific Islanders.

4.5.4. Dynamic instruments

We can lastly extend our dynamic analysis to the changing share of subgroups other than Hispanics in a neighborhood's

Table 13
MSA-based spatial instruments.

	(1) Log Hisp pop	(2) Log non-Hisp white pop	(3) Log black pop	(4) Log Asian and Pacific Isl pop
Reduced form				
Hispanic	0.917	−0.203	−0.315	−0.421
	(0.089)	(0.033)	(0.098)	(0.105)
Non Hispanic white	−2.760	1.120	−1.219	−3.506
	(0.277)	(0.101)	(0.305)	(0.325)
Black	0.115	−0.894	1.062	−0.503
	(0.102)	(0.037)	(0.112)	(0.119)
Asian and Pacific Islander	0.424	−0.242	−0.251	5.831
	(0.202)	(0.074)	(0.222)	(0.237)
Instrumental variables				
Own group share	8.678	4.269	5.435	27.093
	(0.560)	(0.153)	(0.473)	(1.022)

Notes: This table shows reduced form and instrumental variables regressions of changes in group shares in a given neighborhood, controlling for MSA-level population growth and for the MSA-based subgroup instruments for all subgroups. The top part of this table shows reduced form coefficients regressing on the spatially excluded, MSA-based instruments as described in the text. The bottom part shows instrumental variable regressions on own group share, instrumented by the spatially excluded, MSA-based instruments.

Table 14
Dynamic instruments, other demographic groups.

	(1) First stage	IV regressions		
		(2) Log mean imputed rent	(3) Log median rent	(4) Log median house value
80s				
Hispanic	0.181	−0.092	0.008	–
	(0.015)	(0.085)	(0.130)	
Black	0.201	−0.389	−0.457	–
	(0.007)	(0.030)	(0.054)	
90s				
Hispanic	0.254	−0.343	−0.606	−0.575
	(0.016)	(0.049)	(0.098)	(0.147)
Non Hispanic white	0.121	0.512	0.550	1.185
	(0.004)	(0.022)	(0.050)	(0.062)
Black	0.324	−0.359	−0.360	−0.866
	(0.013)	(0.020)	(0.049)	(0.060)
Asian and Pacific Islander	0.470	0.000	0.000	0.000
	(0.044)	(0.000)	(0.000)	(0.000)
Pooled				
Hispanic	0.198	−0.516	−0.240	–
	(0.011)	(0.049)	(0.082)	
Black	0.240	−0.536	−0.373	–
	(0.006)	(0.019)	(0.036)	

Notes: This table shows first stage and IV coefficients of instrumental variable regressions, instrument with lagged changes in any given group's share in a neighborhood and controlling for this group's share and share squared.

population. The argument justifying our dynamic analysis was that, in the presence of search frictions in the housing market, composition should adjust with delay to any amenity shocks, while rental prices and housing values should adjust quickly. By this argument, which holds for a general class of search models, past composition changes are predictive of future composition changes, but excluded from future price changes.

Table 14 applies this idea to our four subgroups. Changes in each group's share are instrumented with past changes in the same group's share, controlling again for this group's share, as well as share squared. Table 14 shows first stage and hedonic IV regressions for decades separately and pooled, and for each of our three housing cost variables. Estimates for non-Hispanic Whites and for Asian/Pacific Islanders are only shown for the 1990s, since membership in either of these groups is not available for the 1970 census.

The resulting estimates suggest a moderately negative average marginal willingness to pay for the share of Hispanics and Blacks, and a moderately positive average marginal willingness to pay for the share of non-Hispanic Whites. It needs to be emphasized again, however, that these estimated values are not very large in an economic sense. A coefficient on Black share of -0.5, for instance, implies that a 1% increase in the share of Blacks in a given neighborhood would lead to a decline of housing prices by 0.5%.

5. Summary and conclusion

This paper presented models of sorting in which location choices depend on the location choices of other agents, as well as exogenous location characteristics. In such a setup, the composition of agents at a location is an endogenous equilibrium outcome with generically degenerate support given exogenous location characteristics. This leads to an identification problem similar to the "simultaneity problem" and the "reflection problem" discussed in the literature: the effects of endogenous composition and exogenous characteristics on agents' location choices and prices are not separately identified.

A series of approaches to overcome this problem was proposed here. The first is based on assuming that some exogenous, location specific demand shifters are excluded from the choices of a subgroup of agents. If that is the case, random variation in such exogenous characteristics can serve as an instrument for endogenous composition. The second is based on assuming a spatial structure with externalities across adjacent locations. Given such a spatial structure, variation in exogenous characteristics at a location generates variation in composition propagating across adjacent neighborhoods, and can serve as an instrument for composition in neighborhoods not immediately adjacent. The third is based on a dynamic search-model extension. In this extension prices adjust quickly but location composition reacts only with delay to changes in exogenous characteristics, because of search frictions. Past shocks in exogenous characteristics can therefore serve as instruments for future composition changes.

In an application of these approaches, the impact of the share of Hispanics in neighborhoods in the United States on housing demand of Hispanics and non-Hispanics as well as rental prices was studied. The results consistently suggest a strong impact of composition on location choices, in the form of an own-group preference. This contrasts with the rather weak evidence on the impact of neighborhood composition on observable outcomes of residents, as in Katz et al. (2007). It remains a task for future research to further disentangle the nature of the social externalities that were found here. For instance, we could think of the reduced form demand functions $D(X,M,P)$ as reflecting preferences over endogenous amenities $W(X,M)$, $D(X,M,P)=D(X,W(X,M),P)$, where $\dim(W)=\dim(M)$. Under this assumption, $D_M = D_W W_M$. Given identification of D_M, one could attempt to identify either D_W or W_M, and then invert to get for instance $D_W = D_M W_M^{-1}$. Identification of W_M could come from shocks to X which are excluded from W but do affect composition M. This approach would require full observability of W.

Application of the methods developed here to a number of different problems seems interesting. For instance, in the field of economic geography firm location choices are studied, where location choices depend on exogenously given geographic factors and the location choices of other firms (and households). One central question of this field is to understand the mechanisms determining the agglomeration or dispersion of economic activity, see for instance Krugman (1991) and Ellison and Glaeser (1999). It seems that the problem of firm location choice has a very similar structure to the problem of household neighborhood choice within a city, which motivated this paper. Another interesting application might be the academic job market: In choosing among job-offers, academics will generally make their decision based not only on exogenous characteristics (location, facilities, ...) and pay, but based also on who else is working at a given university.

Acknowledgments

I thank seminar participants at UC Berkeley, UCLA, USC, Brown, NYU, UPenn, LSE, UCL, Sciences Po, TSE, Mannheim and IHS Vienna, as well as several anonymous referees, for their helpful comments and suggestions. I particularly thank David Card, Kiril Datchev, Ellora Derenoncourt, Gilles Duranton, Michael Jansson, Bryan Graham, Jinyong Hahn, Susanne Kimm, Patrick Kline, Rosa Matzkin, Enrico Moretti, Denis Nekipelov, James Powell, Alexander Rothenberg, Jesse Rothstein, Elie Tamer, and Mark van der Laan for many valuable discussions. This work was supported by a DOC fellowship from the Austrian Academy ofSciences.

Appendix A. Proofs

A.1. Proof of equilibrium existence:

Under the assumptions maintained in Section 2, this follows from applying Brouwer's fixed point theorem to the following bounded continuous mapping with convex domain:

$$(M,P) \to \left(D(X,M,P), P - \left(S(P,X) - \sum_c M^c \right) \right). \tag{A.1}$$

The fixed points of this mapping are exactly the partial sorting equilibria. □

Proof of Proposition 1:

Identification of $D(X,M^*(X),P^*(X))$ follows immediately from the equilibrium condition $D(X,M^*,P^*)=M^*$. Differentiating with respect to (the components of) X yields $D_X + D_M M_X^* + D_P P_X^* = M_X^*$, showing the identification of linear combinations of the demand slopes.

To show non-identification, we have to construct demand functions D such that the equilibrium condition (3) is fulfilled for the known equilibrium correspondence $(M^*(X),P^*(X))$, but D takes arbitrary values off the support of $(X,M^*(X),P^*(X))$. For simplicity, assume that the partial sorting equilibrium is unique and that the mapping from X to M^*,P^* is into (the general case follows in a similar manner). Set $D(X,M,P)=(1-A)M^*(X)-BP^*(X)+AM+BP$ for arbitrary matrices A,B of appropriate dimension. The claim follows. In particular, for this choice of D we get $D_M=A$ and $D_P=B$ for arbitrary matrices A,B, showing non-identification of social externalities and price elasticities. □

Proof of Proposition 2:

Eq. (11) is immediate from $M_X^{*1} = D_X^1 + D_m^1 m_X^* + D_P^1 P_X^*$, since we have $m_{Z^1}^* \neq 0$. Eq. (12) follows from (11) if we can show

$$\frac{M^{*1}_{Z^2}}{P^{*}_{Z^2}} = D^1_P.$$

Under the assumption that, $d_P = 0$, and by assumption of this lemma $d_{Z^2} = 0$, hence $m^{*}_{Z^2} = 0$. It follows that $M^{*1}_{Z^2} = D^1_P P^{*}_{Z^2}$. Finally,

$$P^{*}_{Z^2} = \frac{S_{Z^2}}{E_P - S_P} \neq 0$$

again by assumption. □

Proof of Proposition 3:

For ease of notation, we will drop the superscript 1 for X in this proof. Totally differentiating Eq. (15) yields

$$M^{*c,k}_{Z^l} = D^{c,k}_{\tilde{m}} \tilde{m}^k_{Z^l} + D^{c,k}_P P^k_{Z^l},$$

since $\widetilde{X}^k_{Z^l} = 0$ by the assumption that the k,lth entry of $\mathbf{G}$ equals 0. Similarly

$$m^k_{Z^l} = d^k_{\tilde{m}^k} \tilde{m}^k_{Z^l}.$$

To prove the claim, it remains to show that the denominator $\tilde{m}^k_{Z^l}$ does not equal zero under the conditions stated. Differentiating the equilibrium condition $d = m$ in its vector form, i.e., stacking up the equations for all neighborhoods, gives

$$\mathbf{d}_{\tilde{\mathbf{m}}}\mathbf{G}\mathbf{m}_{\mathbf{X}} + \mathbf{d}_{\widetilde{\mathbf{X}}}\mathbf{G} = \mathbf{m}_{\mathbf{X}}$$

and hence

$$\mathbf{m}_{\mathbf{X}} = (I - \mathbf{d}_{\tilde{\mathbf{m}}}\mathbf{G})^{-1}\mathbf{d}_{\widetilde{\mathbf{X}}}\mathbf{G}, \tag{A.2}$$

where I is the $\mathcal{N} \times \mathcal{N}$ identity matrix and $\mathbf{d}_{\tilde{\mathbf{m}}}$ is a diagonal matrix with positive diagonal entries, by assumption. Invertibility of $(I - \mathbf{d}_{\tilde{\mathbf{m}}}\mathbf{G})$ follows from the normalization of rows of $\mathbf{G}$ to sum to one, and $d_{\tilde{m}} < 1$. We can expand Eq. (A.2) as a geometric series,

$$\mathbf{m}_{\mathbf{X}} = \left(\sum_{j \geqslant 0} (\mathbf{d}_{\tilde{\mathbf{m}}}\mathbf{G})^j\right)\mathbf{d}_{\widetilde{\mathbf{X}}}\mathbf{G}. \tag{A.3}$$

All of the terms in the series have non-negative entries, the k,lth entry of $\mathbf{G}^j$ is not equal to 0 for some power j by assumption, the same holds for $(\mathbf{d}_{\tilde{\mathbf{m}}}\mathbf{G})^j$ by $\mathbf{d}_{\tilde{\mathbf{m}}}$ being a diagonal matrix with positive diagonal entries, and finally $\mathbf{d}_{\widetilde{\mathbf{X}}}\mathbf{G}$ has non-zero diagonal entries. □

Proof of Proposition 4:

Eq. (20) immediately implies

$$\partial\Delta^2 P / \partial\Delta^1 Z^1 = E^s\left[-\frac{u_m}{u_P}\right] \cdot \partial\Delta^2 m / \partial\Delta^1 Z^1.$$

Eq. (19) implies that $\partial\Delta^2 m / \partial\Delta^1 Z^1 \neq 0$ if $D_{Z^1} \neq 0$. The claim follows □

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jue.2014.10.003.

References

Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. J. Polit. Econ. 115 (4), 588–638.

Becker, G., Murphy, K., 2000. Social Economics: Market Behavior in a Social Environment. Harvard University Press.

Black, S., 1999. Do better schools matter? Parental valuation of elementary education. Quart. J. Econ. 114 (2), 577–599.

Bramoullé, Y., Djebbari, H., Fortin, B., 2009. Identification of peer effects through social networks. J. Econometr. 150 (1), 41–55.

Brock, W.A., Durlauf, S.N., 2001. Chapter 54 interactions-based models. Handbook of Econometrics, vol. 5. Elsevier, pp. 3297–3380.

Caetano, G., 2009. Estimation of parental valuation of school quality in the U.S. Dissertation, UC Berkeley.

Card, D., 2001. Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. J. Labor Econ. 19 (1), 22–64.

Card, D., Mas, A., Rothstein, J., 2008. Tipping and the dynamics of segregation. Quart. J. Econ. 123 (1), 177–218.

Chay, K., Greenstone, M., 2005. Does air quality matter? Evidence from the housing market. J. Polit. Econ. 113 (2), 376–424.

Chiappori, P., McCann, R., Nesheim, L., 2009. Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. Econ. Theory, 1–38.

Cutler, D., Glaeser, E., Vigdor, J., 2008. Is the melting pot still hot? Explaining the resurgence of immigrant segregation. Rev. Econ. Stat. 90 (3), 478–497.

De Giorgi, G., Pellizzari, M., Redaelli, S., 2010. Identification of social interactions through partially overlapping peer groups. Am. Econ. J.: Appl. Econ. 2 (2), 241–275.

Duflo, E., Saez, E., 2003. The role of information and social interactions in retirement plan decisions: evidence from a randomized experiment. Quart. J. Econ. 118 (3), 815–842.

Ekeland, I., Heckman, J., Nesheim, L., 2004. Identification and estimation of hedonic models. J. Polit. Econ. 112 (S1), 60–109.

Ellison, G., Glaeser, E., 1999. The geographic concentration of industry: does natural advantage explain agglomeration? Am. Econ. Rev. 89 (2), 311–316.

Ermisch, J., Findlay, J., Gibb, K., 1996. The price elasticity of housing demand in britain: issues of sample selection. J. Housing Econ. 5 (1), 64–86.

Graham, B., 2008a. Identifying social interactions through conditional variance restrictions. Econometrica 76 (3), 643–660.

Graham, B., 2008b. On the identification of neighborhood externalities in the presence of endogenous neighborhood selection. Working paper, UC Berkeley.

Hanushek, E., Quigley, J., 1980. What is the price elasticity of housing demand? Rev. Econ. Stat. 62 (3), 449–454.

Heckman, J., Matzkin, R., Nesheim, L., 2002. Nonparametric estimation of nonadditive hedonic models. Manuscript, University of Chicago.

Kasy, M., 2010. Nonparametric inference on the number of equilibria. Working paper.

Katz, L., Kling, J., Liebman, J., 2007. Experimental analysis of neighborhood effects. Econometrica 75 (1), 83–119.

Krugman, P., 1991. Increasing returns and economic geography. J. Polit. Econ. 99 (3), 483–499.

Manski, C., 1993. Identification of endogenous social effects: the reflection problem. Rev. Econ. Stud. 60 (3), 531–542.

Manski, C., 2003. Partial Identification of Probability Distributions. Springer Verlag.

Matzkin, R.L., 2008. Nonparametric structural models. In: Durlauf, S.N., Blume, L.E. (Eds.), The New Palgrave Dictionary of Economics. Palgrave Macmillan, Basingstoke.

Moffitt, R., 2004. Policy interventions, low-level equilibria, and social interactions. In: Durlauf, S., Young, H. (Eds.), Social Dynamics. MIT Press, pp. 45–82.

Nesheim, L., 2001. Equilibrium Sorting of Heterogeneous Consumers Across Locations: Theory and Empirical Implications. Dissertation, University of Chicago.

Pissarides, C., 2000. Equilibrium Unemployment Theory. The MIT press.

Polinsky, A., 1977. The demand for housing: a study in specification and grouping. Econometr.: J. Econometr. Soc., 447–461.

Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. J. Polit. Econ. 82 (1), 34–55.

Schelling, T., 1971. Dynamic models of segregation. J. Math. Sociol. 1, 143–186.

Tiebout, C., 1956. A pure theory of local expenditures. J. Polit. Econ. 64 (5), 416–424.

Wheaton, W., 1990. Vacancy, search, and prices in a housing market matching model. J. Polit. Econ. 98 (6), 1270–1292.