THE CAUSES AND CONSEQUENCES OF RESIDENTIAL SEGREGATION: AN EQUILIBRIUM ANALYSIS OF NEIGHBORHOOD SORTING[*]

Patrick Bayer
Department of Economics
Yale University

Robert McMillan
Department of Economics
University of Toronto

Kim Rueben
Public Policy Institute of California

All Comments Welcome

June 2001

[*]We would like to thank Fernando Ferreira (University of California – Berkeley) for outstanding research assistance. Financial support from the Public Policy Institute of California is gratefully acknowledged. We would also like to thank the California Census Data Research Center for providing access to the data, and Ritch Milby in particular. Please send correspondence to any of the authors - patrick.bayer@yale.edu, mcmillan@chass.utoronto.ca, or rueben@ppic.org.

1 INTRODUCTION

Residential segregation on the basis of race and socioeconomic status is both a highly visible phenomenon in the United States and one perceived to have important social implications. Where segregation is extreme, as in the case of urban ghettos, there is a sense that the combination of poverty, adverse neighborhood spillovers, and isolation from mainstream society all make it difficult for an individual to perform well – in school, the labor market, and in non-criminal activities generally.[1] And cognizant of problems associated with residential segregation, policy makers have long been interested in measures to counter it, leading in the past to policies such as the busing of students between school districts to achieve racial balance. Yet our understanding of the causes and consequences of segregation remains incomplete, and this hinders the formation of sound policy.[2]

Theory work in economics and elsewhere has helped inform thinking about the forces underlying observed segregation patterns. Schelling's models of social interactions, for instance, emphasize the role of preferences for neighborhood racial composition, showing how even small differences in such preferences can give rise to high levels of racial segregation and produce important dynamic phenomena such as 'neighborhood tipping.'[3] In Tiebout's theory, the emphasis is on preferences for local public goods, with households sorting across communities offering different public goods packages that are excludable on the basis of location;[4] residential stratification on the basis of race or income is likely to the extent that household preferences for local public goods vary with these characteristics.

Of necessity, such models abstract from other important aspects of the underlying sorting process that cannot be ignored in empirical work. In terms of preferences, households care about

[1] Segregation may have both positive and negative impacts on households from various backgrounds, as Cutler and Glaeser (1998) note. To the extent that recent immigrants are able to take advantage of family and ethnic networks to ease their transition into American society, for instance, ethnic segregation may be extremely valuable. However, if racial segregation limits housing opportunities for certain groups of households to neighborhoods that are poorly situated in relation to employment opportunities - the spatial-mismatch hypothesis - such segregation can have serious detrimental impacts on the educational attainment, employment, and welfare participation of households living in those neighborhoods, likely to be stronger if peer or neighborhood effects are important. In their work, Cutler and Glaeser find the overall impact of greater segregation to be negative, especially for black households.

[2] Because the potential causes of observed segregation patterns are so diverse and because households may react to reform policies, measures designed to decrease segregation may have unintended consequences. School desegregation policies implemented jurisdiction-by-jurisdiction, for example, may lead to increases in inter-jurisdictional residential segregation. See Inman and Rubinfeld (1979) and Rivkin (1994).

[3] See Schelling (1969, 1978). Moreover, Anas (1980) demonstrates that 'neighborhood tipping' on the basis of race is further accentuated by differences in the financial assets of households of different races.

[4] Epple, Filimon and Romer (1984, 1993), for example, develop equilibrium models of community sorting when households are differentiated on the basis of their demand for local public goods and communities are distinguished by the level of local public good that they provide.

more than just the race of their neighbors or the level of local public goods provision when making their location decisions; they make tradeoffs among the wide variety of housing and neighborhood attributes associated with the available choices, and their demands for a given attribute vary with household characteristics. A married couple with a 5 year-old child, for example, may place a high value on the quality of the neighborhood school, while a duel-earner couple may place a premium on the accessibility of the home to both jobs. Learning more about the heterogeneity in individual household preferences is an important step toward gaining a better understanding of the forces driving residential segregation in the aggregate. It is worth emphasizing, though, that the distribution of households across neighborhoods within a metropolitan area arises through a complex sorting process. While racial segregation may be attributable in part to households' preferences over the race of their neighbors, the correlation of race/ethnicity with other household characteristics makes it likely that many other factors contribute to the observed segregation patterns. As an example, if households sort across school districts according to their willingness to pay for school quality, school districts will be stratified on the basis of income and, because race is correlated with income, this stratification may lead to increased residential segregation by race.

The central goal of this paper is to gain a fuller understanding of the underlying causes and consequences of residential segregation. To that end, we make use of newly available restricted-access Census data. Unlike the publicly available Census data, which match each household with a PUMA (a Census area of at least 100,000 residents), the restricted-access data provide a household's residential and employment locations at the level of a Census block (a Census area with approximately 100 residents). These unprecendently rich data make it possible to characterize each household's actual neighborhood much more accurately than has been possible in past studies.[5] Using these new Census data as a centerpiece, we have assembled a extensive data set characterizing the housing market in the San Francisco Bay Area. In addition to the housing and neighborhood sociodemographic data drawn from the Census, we have collected neighborhood-level data on schools, air quality, climate, crime, drug use, topology, geology, land use, and urban density. Moreover, the Census data provide detailed information on the households in the sample, including each household member's race, education, income, age, immigration status, employment status and job location.

[5] One exception is the study by Borjas (1998), which uses a restricted version of the NLSY, combining individual data with limited information about the ZIP code – in particular, for the 1979 wave of the survey, whether other individuals in the sample reside in the same ZIP code. This allows a rough picture of the characteristics of other ZIP code residents to be built up for that year on condition that they fall in the NLSY sample, even though the exact location of each ZIP code within the metropolitan area cannot be established.

In Bayer, McMillan, and Rueben (2001), we use this new data set to examine the extent to which the correlation of race with other household characteristics such as education, income, language, and immigration status explains the observed racial segregation patterns in the San Francisco Bay Area. Specifically, using a reduced-form approach, we first measure the average racial composition of neighborhoods in which Asian, black, Hispanic, and white households live.[6] We then explore whether the marked differences in the typical neighborhood in which households of different races live are in part explained by variation in other household attributes. This approach allows us to examine, for example, whether the differences in income and education across races are partly responsible for the some of the observed segregation patterns.[7] While the approach begins to inform our understanding of the forces that drive the observed segregation patterns, it is difficult to distinguish the mechanisms through which other household characteristics, such as income and education, influence the residential location decision. For instance, do differences in education levels lead to greater racial segregation because they cause households to value school quality differently or because jobs that employ workers with different levels of human capital are distributed differently geographically within the metropolitan area? The reduced-form approach also makes it difficult to consider the consequences of segregation. To what extent do direct preferences for racial sorting, for example, affect the quality of schools or housing prices that households from different backgrounds receive?

In the current paper, we develop a methodology that allows us to examine the causes and consequences of racial segregation in a general equilibrium framework. Specifically, we set out an equilibrium model of the housing market for the San Francisco Bay Area centered on a discrete-choice model of the residential location decision. In the model, households have preferences defined over housing, schooling, and other neighborhood attributes, preferences being allowed to vary with own-household characteristics. Parameters of the utility function are recovered by appealing to the revealed preference principle implicit in a Nash equilibrium assumption – i.e., each household is assumed to have made its optimal location decision given the set of alternatives and the location decisions of other households. The resulting estimates provide a detailed picture of household preferences over the many characteristics of the choices in the choice set, including neighborhood attributes (location, schools, crime, environmental amenities, sociodemographic composition, housing characteristics and price), showing also how these

[6] Here, we find evidence of a striking tendency of black households to live in communities with a high proportion of other black households. The typical black household lives in a neighborhood that is 40 percent black, five times higher than the proportion of black households in the sample as a whole. A similar, though less striking pattern of segregation is apparent for other racial groups.

[7] Note summarizing results of Bayer, McMillan, and Rueben (2001).

preferences vary with a household's own characteristics - race, education, income, and place(s) of work.

We confront an important endogeneity problem in estimating the model, recovering parameters of the household utility function consistent with the fact that neighborhood sociodemographic compositions and housing prices are determined as part of a sorting equilibrium. Households are allowed to have preferences over many attributes that depend directly on the way that other households sort across neighborhoods. To the extent that households sort taking into account unobserved house and neighborhood characteristics, so the correlation between observed sociodemographics and unobservables is likely to confound conventional estimates of preferences. As an example, the presence of unobserved attractive features of a neighborhood should raise the average income of that neighborhood *ceteris paribus*; any estimation procedure that does not account for this correlation will tend to attribute what are really preferences for these attractive unobserved neighborhood characteristics to preferences over observable characteristics, such as the average income of one's neighbors.

In light of this difficulty, we follow the methodology developed in Bayer (2001) for identifying preferences over variables determined through the sorting process in the presence of unobserved fixed house and neighborhood attributes. In essence, identification requires variables that shift the sociodemographic composition of a particular neighborhood but are uncorrelated with the unobserved characteristics of that same neighborhood. We develop variables that satisfy these conditions by drawing on the underlying geographic distribution of neighborhoods and their physical features within the study area. Because each household chooses its neighborhood from a large set of potential alternatives, each household's choice, and, consequently, the resulting sociodemographic composition of each neighborhood, is shaped by the spatial distribution of neighborhoods and their characteristics – not solely the characteristics of the chosen neighborhood. The particular identification strategy that we develop uses variables that characterize the underlying spatial distribution of the *exogenous* features of neighborhoods as a source of variation that affects the sociodemographic composition of a neighborhood, but which is uncorrelated with the unobserved characteristics of that neighborhood. In essence, this strategy exploits information contained in the discrete distribution of available choices to identify social interactions and is applicable for identifying social interactions in a wide class of discrete choice models.[8]

[8] For a full discussion of the identification of social interactions in endogenous sorting models see Bayer and Timmins (2001).

The resulting parameter estimates describe how household preferences for a wide set of neighborhood, schooling, and housing characteristics vary with a wide set of household characteristics.[9] The main economic analysis of this paper is conducted by using counterfactual general equilibrium simulations of the estimated model. These simulations solve for the new sorting equilibrium that arises as a result of changing the primitives of the model. In each simulation, a new equilibrium is calculated that is characterized by an updated vector of housing prices and an updated set of household residential location decisions. These revised location decisions impact the sociodemographic composition of each neighborhood and the schools within each neighborhood, which in turn alter the levels of crime and school quality in each neighborhood. The simulations account for the full impact of changing neighborhood sociodemographics on household preferences, solving for a new equilibrium in which the vector of housing prices clears the market and each household makes its optimal location decision given the location decisions of all other households.

Each simulation produces a new distribution of households across neighborhoods as well as an updated set of housing prices, neighborhood crime rates and school quality levels, tenure decisions, and commuting patterns. Using the outcomes of the simulation, we first examine the impact of the experiment in question on the observed patterns of racial segregation. By restricting all households to have identical preferences for the race of their neighbors, for example, we solve for a new equilibrium that is informative about the relative importance of this factor in explaining the observed segregation patterns. We then explore how the experiment changes the ways that households with different characteristics are matched with housing prices, school quality, crime, ownership rates, and commutes. Because we back out measures of both observed and unobserved components of housing and neighborhood quality, we can consider, for example, how the restriction of identical preferences for neighborhood racial composition affects the average quality-adjusted housing price for households of different races.

The remainder of this paper is organized as follows. In Section 2, we discuss the methodological contributions of this paper relative to the previous empirical literature that has examined residential location decisions, and in Section 3, we introduce the modeling framework and describe the equilibrium properties of the model. We describe the extensive new data set that we have assembled for the analysis in Section 4 and discuss the estimation of the model in

[9] It is important to note that our modeling framework is based on the assumption that each household chooses its optimal location from the full set of housing units in the study area. To the extent that other factors, such as discrimination, informational problems, and moving frictions, have a large role in the sorting process our model will interpret these as preferences.

Section 5. In Section 6, we present the parameter estimates of the model and set out the general equilibrium simulations that constitute the heart of our analysis. Section 7 concludes.

2 PREVIOUS LITERATURE AND THE CONTRIBUTIONS OF THIS PAPER

Following the path-breaking work by McFadden (1973, 1975, 1978), a number of authors have used discrete choice models to estimate the factors that drive the residential location decision. In essence, the discrete choice approach uses the revealed information in a household's observed location choice given the range of potential locations to infer the household's preferences for the characteristics of these locations. In principle, this approach can be used to estimate household preferences for a wide variety of housing, neighborhood, and community characteristics, even though the majority of previous studies in the literature have had a much narrower focus, restricting their attention to estimating how households trade off between housing characteristics, commuting modes and times.[10] Many features of a house or neighborhood are likely to be unobserved in any dataset, a fact that past studies have routinely ignored. Because such unobserved features influence the attractiveness of a choice, they are certainly correlated with the housing prices, community sociodemographic characteristics, and any other features (such as public safety or school quality) that depend in turn on these sociodemographic characteristics; the failure to account for such unobserved features and thus to deal with this important endogeneity problem represents a serious shortcoming in the literature.

Epple and Sieg (1999) offer a new approach to estimating preferences for the factors that drive household sorting across communities by estimating an equilibrium model of community sorting based on the underlying theoretical framework developed in Epple, Filimon, and Romer (1984, 1993). In essence, this underlying framework is a vertical model of differentiation, with communities delineated by a single index of quality, called a public goods index. In equilibrium, the model predicts the stratification of households across communities based on income and taste for the public good index, estimation proceeding by attempting to match the predicted equilibrium distribution of households across communities to the observed patterns in the data. Explicitly accounting for the fact that local public goods are determined in part by household sorting, Epple and Sieg make a significant advance over prior work in the area.

[10] Many researchers in the urban economics literature, for example, have used this framework to estimate the importance of housing characteristics and commuting costs in the residential location decision. Ellickson (1981), Quigley (1985), Blackley and Ondrich (1988) focus on estimating the influence of housing characteristics on the location decision, while Anas (1982), Pollakowski (1982), Anas and Chu (1984) focus on estimating the influence of commuting costs. Quigley (1985) and Nechyba and Strauss (1998) include measures of public expenditures on education, policing in their analyses.

The current paper builds on the approach set out in Bayer (2001), differing form both the previous discrete choice literature and the work of Epple and Sieg. Like the discrete choice literature, the central feature of this framework is an indirect utility function that specifies household preferences for a wide variety of community, school, and housing characteristics. By specifying this utility function in a flexible way, a household's valuation of *each* community characteristic can vary with household characteristics such as race, education, income, and family structure.

The key contribution of this approach relative to the previous discrete choice literature is that we formally consider the sorting equilibrium implied by the underlying model of household location choice. When estimating the model, this requires addressing the fact that many community characteristics including sociodemographic characteristics, housing prices, and local public goods are correlated, in equilibrium, with any unobserved features that shape household location decisions. In order to identify household preferences for these sorting-dependent characteristics, it is necessary to find instruments that are correlated with these variables but uncorrelated with the unobservable characteristics of a community; and in Section 5 below, we develop a set of instruments based on the geographic distribution of communities and their exogenous features. These instruments rise naturally out of the choice process itself as they influence how household sort across communities but can reasonably be assumed to be uncorrelated with unobserved community characteristics.

The key contribution of our approach relative to Epple and Sieg lies in the flexible specification we adopt for the utility function, allowing household preferences to vary with a wide variety of community characteristics as well as characteristics of households themselves. This allows for the horizontal differentiation of communities as opposed to the pure vertical model of differentiation used in their work.[11] In practice, the vertical model constrains households with different characteristics and income to make the same trade-offs between community characteristics, so that, for instance, workers employed in the suburbs are restricted to have the same preferences for central residential locations relative to other community characteristics as workers employed in the central city. Combining the flexible preference

[11] It is also worth noting that the empirical analysis of Epple and Sieg (1999) does not include measures of community sociodemographic characteristics. As discussed in the introduction, this may seriously bias estimates of preferences for local public goods, as these sociodemographic characteristics (unobserved in the Epple and Sieg framework) are certainly highly correlated with the observed local public goods. The problem of considering preferences for sociodemographics is complicated within the Epple and Sieg framework by the fact that preferences for these characteristics may differ quite non-monotonically across households of different races and ethnicities. Simply including them as part of the public goods index would place undue constraints on these preferences.

specifications of the discrete choice literature with the consideration of important general equilibrium properties of community sorting, we seek to provide a general framework for analyzing a wide range of economic and policy questions in urban economics and local public finance

3 THE EMPIRICAL FRAMEWORK

The central component of the equilibrium model of the housing market that we take to the data is a discrete choice model that specifies how each household makes its location decision. Given the underlying distribution of households and their preferences, we define a housing market equilibrium to be a distribution of households across houses and a set of housing prices such that the housing market clears and each household's decision is optimal given the decisions of all other households. While it is generally possible to estimate a more flexible housing supply functions, we assume in this paper that the supply of houses and their characteristics (e.g., age, size) is fixed. In conducting the general equilibrium simulations that constitute the core of our analysis, therefore, we do not account for any effects that would occur through changes in the housing stock. In this section, we first introduce the discrete choice model of the residential location decision and then describe the equilibrium properties of the model.

Modeling the Residential Location Decision

The discrete-choice model specified here describes the residential location decisions of all households in the San Francisco Bay Area. The particular model is based on the random utility model developed in McFadden (1978) and the specification of Berry, Levinsohn, and Pakes (1995), which includes choice-specific unobservable characteristics. Each household chooses its residence h to maximize its utility, which depends on the observable and unobservable characteristics of its choice. Let X_h represent the observable characteristics that vary with its housing choice and let p_h denote the price of house h. In our framework, the observable characteristics of a housing choice include characteristics of the house itself (e.g., size, age, and type), its tenure status (rented vs. owned), and the characteristics of its neighborhood (e.g., sociodemographic composition, school, crime, topology, and air quality). Because of the difficulties comparing prices between housing units that are rented versus owned, we estimate separate price elasticities for housing units depending on their tenure status. This additional flexibility permits a household's willingness to pay for housing and neighborhood characteristics

to vary with its tenure status,[12] and does away with the need to make arbitrary conversions between housing values and rents.

The household's optimization problem is given by:

$$(3.1) \qquad \underset{(h)}{Max} \quad V_h^i = \alpha_X^i X_h - \alpha_D^i D_h^i - \alpha_p^i p_h + x_h + e_h^i$$

where x_h are unobserved characteristics associated with each housing unit including unobserved attributes of the corresponding neighborhood. The $\alpha^i_D D^i_h$ term in the utility function captures the disutility of commuting – the negative impact of the distance between household i's workplace and house h. The final term of the utility function, e^i_h, is an idiosyncratic error term that captures unobserved variation in household i's preference for a particular housing choice.

Each household's valuation of choice characteristics is allowed to vary with its own characteristics, Z_i, including education, income, race, employment status, and household composition. We also assume that each working household is initially endowed with a primary employment location, l_i. We treat employment status and employment location as exogenous variables throughout this paper. Each parameter associated with housing characteristics, distance to work, and price, α^i_j, j *in* $\{X, D, p\}$, is allowed to vary with a household's own characteristics,

$$(3.2) \qquad \alpha_j^i = \alpha_{0j} + \sum_{r=1}^{R} \alpha_{rj} Z_r^i \, .$$

In this way, equation (3.2) describes household i's preference for choice characteristic j. The first term captures the taste for the choice characteristic that is common to all households and the other terms captures observable variation in the valuation of these choice characteristics across households due to socioeconomic characteristics. This heterogeneous coefficients specification allows for great variation in the preferences of households with different characteristics. While it would also be possible to include stochastic terms in (3.2), our current analysis does not include such random coefficients.

The specification of utility given in equations (3.1)-(3.2) contains two stochastic components, which allow the model flexibility in explaining the observed data. The first component is the idiosyncratic term e^i_h, which is assumed to be additively separable from the rest of the utility function. Different assumptions about the distribution of e give rise to different

[12] Note that we also include tenure status directly in the utility function.

substitution patterns. While we intend to estimate more flexible stochastic structures in the future, we currently assume that e is distributed according to the Type II extreme value distribution, giving rise to the multinomial logit model.

The second stochastic component of the utility function is the house-specific unobservable, x_h. This term is captures those aspects of a particular housing choice (including unobserved neighborhood and house attributes) that are observable to the households in the sample but are unobservable to the econometrician. Because many housing and neighborhood attributes are likely to be unobserved in any data set, simpler specifications of the utility function that do not include such unobservables are likely to lead to biased parameter estimates. The houses in neighborhoods with high level of unobserved quality, for example, will certainly command higher prices and attract higher income households, *ceteris paribus*. In this way, the failure to account for unobservable choice characteristics in the error structure can produce undesirable results such as upward-sloping demand curves.

Equilibrium: Definition and Existence

While the random utility specification in equations (3.1)-(3.2) is extremely flexible from an empirical point of view, it also has a convenient theoretical interpretation. Without the idiosyncratic error component, e^i_h, this specification would suggest that two individual households with identical characteristics and employment locations would make identical choices. Since this is unlikely to be true in the data, a useful interpretation of e^i_h is that it captures unobserved heterogeneity in preferences across otherwise identical individuals. In this way, for any set of otherwise identical households, the model predicts not a single choice but a probability distribution over the set of housing choices. By working with this probability distribution over the set of choices rather than a single discrete decision for each household in the sample, it is straightforward to define and explore the properties of the sorting equilibrium for the class of models depicted in equations (3.1)-(3.2). We discuss the implications of this assumption below.

For a general distribution of e, we can write the probability P^i_h that household i chooses house h as:

$$(3.3) \qquad P^i_h = f(Z^i, \hat{X}, \hat{p}, \hat{x})$$

where "^" over a variable refers to a vector over houses and we have included the household's employment location in Z^i and the location of the house in X_h for simplicity of exposition.

Aggregating the probabilities in equation (3.3) over all households yields the predicted number of households that chooses each house, N_h:

$$(3.4) \qquad N_h = \sum_i P_h^i$$

In order for the housing market to clear, the number of households choosing each house must equal one:

$$(3.5) \qquad N_h = 1, \quad \forall h$$

It is a straightforward extension of Berry (1994) to show the existence of a unique vector of housing prices (up to a scaleable constant) that solves the system of equations in (3.5) for utility specifications in which price enters linearly and e is drawn from a continuous distribution. In essence, Berry shows that it is possible to write the solution to (3.5) as a contraction mapping in p_h. Thus, starting from any vector $\hat{p}$, an iterative process that increases the prices of houses with excess demand and decreases the prices of houses with excess supply eventually results in an even spread of households across houses. Writing this market-clearing vector of prices as $\hat{p}^*$, the probability that household i chooses house h can be written:

$$(3.6) \qquad P_h^i = f(Z^i, \hat{X}, \hat{p}^*, \hat{x})$$

If the entire set of house and neighborhood characteristics X that households value were exogenous, an equilibrium could simply be defined as the set of choice probabilities in equation (3.6) along with the vector of market clearing prices, $\hat{p}^*$.

For the analysis undertaken in this paper, however, we allow households to have preferences for the sociodemographic characteristics of their neighbors. Such preferences may arise through multiple channels as households may value the characteristics of their neighbors directly and may also value other neighborhood attributes such as public safety and school quality that are influenced by neighborhood sociodemographic characteristics. In this case, we can rewrite the set of probabilities defined in equation (3.6) to explicitly depend on a set of average neighborhood sociodemographic characteristics $\overline{Z}_h$:

(3.7) $$P_h^i = f(Z^i, \hat{\bar{Z}}, \hat{X}, \hat{p}^*, \hat{x})$$

In writing equation (3.7) in this way, we make two notational simplifications that do not affect the derivation of the equilibrium properties of the model. First, the role of $\bar{Z}$ in equation (3.7) captures the full impact of neighborhood sociodemographics on utility (both the direct and indirect channels mentioned above). Second, while equation (3.7) is specified in terms of the average neighborhood sociodemographic characteristics, any continuous function of neighborhood sociodemographic characteristics could be used.

By aggregating P^i_h over all of the houses in a particular neighborhood we can solve for the predicted sociodemographic characteristics of that neighborhood. Specifically, averaging the product of the probabilities in equation (3.7) and household characteristics Z_i over all houses h in neighborhood n yields the average sociodemographic characteristics for neighborhood n, $\bar{Z}_n$:

(3.8) $$\bar{Z}_{h \in n} = \bar{Z}_n = \tfrac{1}{N_n} \sum_{h \in n} \sum_i Z^i P_h^i = \tfrac{1}{N_n} \sum_{h \in n} \sum_i Z^i f(Z^i, \hat{\bar{Z}}, \hat{X}, \hat{p}^*, \hat{x})$$

Here N_n equals the number of houses in neighborhood n and $\bar{Z}_{h \in n}$ indicates that house h is assigned the average sociodemographic characteristics for its neighborhood n. Rewriting equation (3.8) in vector notation yields:

(3.9) $$\hat{\bar{Z}} = \tfrac{1}{N_n} \sum_{h \in n} \sum_i Z^i f(Z^i, \hat{\bar{Z}}, \hat{X}, \hat{p}^*, \hat{x})$$

This system of equations represents a mapping of the vector $\hat{\bar{Z}}$ into itself and we define an equilibrium as a fixed point of this mapping. Specifically, we define a sorting equilibrium to be a set of choice probabilities for each household as defined in equation (3.7), along with the corresponding market clearing set of prices $\hat{p}^*$, such that the system of equations defined in equation (3.9) holds.

Notice that this equilibrium concept relies on aggregating choice probabilities rather than actual discrete choices. Defining the equilibrium in this way is convenient because it ensures that an equilibrium always exists under a set of simple and reasonable assumptions. More specifically, the existence of an equilibrium is guaranteed as long as utility is continuous in $\bar{Z}_h$

and the distribution of the idiosyncratic component of preferences e^i_h is continuous.[13] The assumption about the continuity of the utility function ensures that the conditional probabilities defined in equations (3.7) are continuous functions of $\overline{Z}_h$. The assumption about the error distribution in turn implies that equation (3.9) represents a continuous mapping of the vector $\overline{Z}_h$ into itself and that the vector $\overline{Z}_h$ is confined to a closed and bounded set. The existence of a vector $\overline{Z}_h{}^*$ that satisfies equation (3.9) then follows directly from Brouwer's Theorem. Under a simple set of continuity assumptions, an equilibrium always exists for this class of models.

In defining both the conditions for the existence of a market-clearing set of prices and a full sorting equilibrium, we use choice probabilities for each household defined over the full set of houses rather than a specific housing choice. As the analysis above indicates, this way of simplifying the equilibrium concept is extremely valuable because it smoothes the discrete decision problem, transforming it into a set of continuous probabilities. Without this smoothing, it would not be possible to establish that a unique vector of housing prices clears the market, as there would be a range of prices between the price that would cause another household to move in and the price that would cause the current occupant to move out that could hold for each house. The use of choice probabilities for each household defined over the full set of houses rather than a specific housing choice also plays a crucial role in establishing a full sorting equilibrium. In this case, this assumption smoothes the response of households to changes in the set of housing alternatives. Consider, for example, a household that initially strongly prefers one of two possible housing choices to the other. In this case, an increase in the attractiveness of the second house would have no effect on the household's actual choice as long as it was small, but would cause the household to switch choices as soon as it reached the level of utility that the initial choice provided. It is precisely these large changes in probability (0 to 1), which can in turn affect prices and neighborhood sociodemographics enough to lead other households to change their decisions, that can cause equilibrium problems when discrete decisions are used. By avoiding these discrete jumps, the use of choice probabilities smoothes the decision-making process, thereby ensuring that an equilibrium always exists.

Using choice probabilities rather than discrete decisions is essentially equivalent to assuming that each household in the data represents a continuum of similar households. This is

[13] The assumption that households value the average characteristics of their neighbors is not necessary for existence. Existence is guaranteed as long as the utility function is continuous in any continuous function of neighborhood sociodemographic characteristics. When neighborhood sociodemographic characteristics work through indirect channels, such as affecting the crime rate, this continuity assumption also implies that neighborhood sociodemographics must affect these other choice characteristics in a continuous way.

precisely the type of assumption that is typically made in theoretical analyses of sorting models.[14] In an empirical context, we believe that this assumption is also reasonable, as it implies only modest changes to the behavior predicted by the model relative to using discrete decisions. First, in the case of the market clearing set of prices, the main difference between these equilibrium concepts is that using choice probabilities abstracts away from the fact that using discrete decisions produces a range of prices that would be high enough to keep others from moving in and low enough to keep the current occupant from moving out. The implicit assumption in our framework, that the total demand for each house is the same, seems to capture the essential role that prices play in setting the metropolitan area sorting equilibrium. In the case of the full sorting equilibrium, the use of choice probabilities is even less worrisome, especially if neighborhoods have a large enough number of houses. In this case, the predicted average sociodemographic composition of the neighborhood is likely to be almost identical to that predicted using discrete decisions.

Equilibrium: Uniqueness

While it is straightforward to establish the existence of an equilibrium using the equilibrium concept defined above, a unique equilibrium need not arise in this framework. Consider, for example, an extreme case in which two types of households that have strong preferences for living with only their own type must choose between two otherwise identical neighborhoods. In this case, it is easy to see that multiple equilibria could result. Households could, for example, sort by type with all households of one type choosing either one of the neighborhoods. Thus, uniqueness is not a generic property of these models.

This extreme example, however, overstates the likelihood that multiple equilibria arise in this model. Consider an example at the opposite extreme, in which neighborhood sociodemographic characteristics do not enter the utility function In this case, because a unique set of prices clears the market, a unique equilibrium arises. In general, a unique equilibrium will arise as long as the meaningful variation in the exogenous attributes of neighborhoods and houses is sufficiently rich relative to the role that preferences for neighborhood demographics play in the location decision.[15]

[14] See, for example, Nechyba (1997, 1999), Epple, Filimon, and Romer (1984, 1993), and Benabou (1996).
[15] See Bayer and Timmins (2001b) for a formal analysis of the conditions under which unique equilibria arise in these models.

4 Data

At the heart of our analysis lies an extensive data set built around restricted Census data for 1990. The unique feature of these restricted Census data is that they provide information about the location of individual people's residence and place of work, down to a very disaggregated level. For the purposes of the current analysis, this allows us to identify far more accurately than has been possible with such a large-scale data set the local neighborhood each individual inhabits as well as each working individual's place of work. The restricted Census data also provide the standard individual, household and housing variables found in the public-use version of the Census.

The study area for our analysis consists of six contiguous Bay Area counties: Alameda, Contra Costa, Marin, San Mateo, San Francisco, and Santa Clara. Examination of Bay Area commuting patterns reveals that a very small proportion of commutes originating within these six counties end up at work locations outside the area. Similarly, a relatively small number of commutes originate outside the six counties and involve travel to work within the six counties. Thus, our study area is reasonably self-contained. These six counties include over 1,100 Census tracts, and almost 39,500 Census blocks, the smallest unit of aggregation in our data. For our sample, the average Census block contains 17 households, a Census block group contains an average of 92 households, a tract contains 255 households, while a county contains on average 49,488 households. Overall, the sample includes over 650,000 people in 244,629 households.

Our data include a wealth of data that characterize the households in the sample. Household attributes include the age, race, level of education attainment, employment status and location of each household member as well as characteristics of the household such as size, marital status, presence of school-aged children, etc. To characterize the choice set, we begin by extracting information about each house in the sample – tenure status, number of rooms, number of bedrooms, house value, age of the unit, occupancy status, and more – from the Census. Using the same individual person and household data, we construct a series of variables characterizing the neighborhoods a given person or household lives in. We are able to define a variety of neighborhoods based on conventional boundaries – the block, block group, tract, and county. In addition, as we know the latitude and longitude of the area center of each Census block, we define a succession of neighborhoods of a given block according to whether other blocks lie within a given radius. Using this approach, we can construct the racial, education and income distributions for a given neighborhood drawn around a given block.

To provide a detailed picture of the local environment, we have also merged (using Census block identifiers) a great deal of additional data for California's schools, school districts,

and neighborhoods with each household record. The school data include achievement scores, teachers' characteristics and salaries, and many other school inputs (such as average class size) for each school district and for each school. We have also collected community-level data on air quality, climate, crime rates, topology, geology, urban density, and congestion, among other location-specific data.[16] Merging the separate components together, the resulting data set includes detailed information on each household, the house they inhabit, their neighborhood, schools, topography, and other local conditions.

5 Estimation

We begin the description of the estimation procedure by introducing some notation that simplifies the exposition. Dividing the terms of the utility function into a *house fixed effect*, δ_h, an *interaction component*, μ^i_h, which includes any parts of the utility function that interact household and choice characteristics, and the *idiosyncratic error term*, ε^i_h, the utility function can be rewritten as:

$$(5.1) \qquad V^i_h = \delta_h + \mu^i_h + \varepsilon^i_h.$$

where:

$$(5.2) \qquad \delta_h = \alpha_{0X} X_h - \alpha_{0p} p_h + \xi_h$$

$$(5.3) \qquad \mu^i_h = \left(\sum_{r=1}^{R} \alpha_{rX} Z^i_r \right) X_h - \left(\alpha_{0D} + \sum_{r=1}^{R} \alpha_{rD} Z^i_r \right) D^i_h - \left(\sum_{r=1}^{R} \alpha_{rp} Z^i_r \right) p_h$$

Here, the δ_h term in these equations captures those components of the utility function (including the unobservable characteristics term ξ_h) that are common to all households.

Because uniqueness is not a general feature of the model, it is clearly not possible to estimate the parameters of the model using Maximum Likelihood, which requires a one-to-one mapping of the parameters and unobservable components of the model to the endogenously determined outcomes. Consequently, we develop a strategy for estimating the parameters using

[16] In generating the data at the Census block level, we make use of locally weighted regression techniques to assign data on climate stations and air quality monitoring stations to a lower level of aggregation (in this case, a Census block) – there are far fewer climate stations in the state than there are Census blocks. Such assignment procedures are likely to induce less measurement error than assigning climate readings of the nearest station or a simple weighted average of the closest stations.

the Generalized Method of Moments (GMM). In this case, the underlying theoretical model need not have a unique equilibrium.[17] Instead, we base the estimation of many of the model's parameters on the assumption that the observed decisions are individually optimal given the collective choices made by other individuals.[18] In this way, our estimation strategy relies on the assumption that an equilibrium exists and is observed, but not on the fact that this equilibrium is unique.

By forming moment conditions based on maximizing the probability that the model correctly predicts each household's choice given the location decisions of all other households, this GMM procedure corresponds to estimating the model using a naïve-ML procedure, (i.e., a ML procedure that ignores the fact that the housing prices and neighborhood sociodemographic characteristics are determined as part of a sorting equilibrium). This assumption is reasonable as long as each household's location decision has a small effect on the decisions (utility) of the remaining population. Notice, however, that forming moment conditions from matching the observed choices made by the households in our sample only permits the estimation of the parameters of the interaction component m^i_h and the vector of house fixed-effects, d_h. Importantly, the observed choices provide no information about how to distinguish the elements of the house fixed effect (i.e., the parameters and unobservables in equation (5.2)). In order to estimate these parameters, it will be necessary to bring additional econometric information to bear on the problem – and it is in this portion of the estimation procedure that the fact that housing prices and neighborhood sociodemographics are determined as part of a sorting equilibrium will be important.

Estimating 10,000 Fixed Effects

For our sample of 10,000 households (for the results presented in this paper) we must estimate 10,000 house fixed effects. As mentioned above, forming moment conditions based on maximizing the probability that the model correctly predicts each household's choice given the location decisions of all other households allows the estimation of the parameters of the interaction component m^i_h and the vector of house fixed-effects, d_h. Estimating such a large number of fixed effects in this way is computationally impossible. Consequently, we turn instead

[17] Even with the application of GMM, in order to perform counterfactual general equilibrium simulations, it is still desirable for the model to have a unique equilibrium for the estimated set of parameters and unobservables.

[18] This corresponds to the use of first-order conditions as moment conditions in the GMM estimation of preferences or technologies in continuous settings.

to a two-step procedure. For each potential set of interaction parameters, we first choose the house fixed effects such that the housing market clears (i.e., a total of one household chooses each house). Using these fixed effects, we then form a GMM objective function based on maximizing the probability that each household make the correct choice. As we explain below, for each set of interaction parameters, a quick contraction mapping backs out the set of house fixed effects, and consequently, by searching over the interaction parameters we are able to estimate the model in a computationally feasible way.

The estimation of the utility parameters begins by setting all of the interaction parameters to a set of starting values, which temporarily fixes m to be a fixed matrix of known values, $\hat{m}$. For any combination of interaction parameters and house fixed effects, d, the model predicts the probability that each household chooses house h:

$$(5.4) \qquad P_h^i = \exp(d_h + \hat{m}_h^i) \Big/ \sum_k \exp(d_k + \hat{m}_k^i)$$

Summing this probability over all households yields the predicted number of households that choose each house N_h:

$$(5.5) \qquad N_h = \sum_i P_h^i$$

The same application of Berry (1994) that ensured that a unique vector of housing prices clears the market in the discussion of the equilibrium in Section 3, also ensures that a unique vector of house fixed effects (up to scaleable constant) guarantees that a total of one household chooses each house here (i.e., $N_h=1$). Berry (1994) not only proves that a unique solution exists to this system of equations but also provides a quick contraction mapping that solves for these house fixed effects. For this application, the contraction mapping is simply:

$$(5.6) \qquad d_h^{t+1} = d_h^t - \ln(N_h)$$

Given any set of interaction parameters, then, the first step of this part of the estimation procedure solves for the unique vector of house fixed effects that exactly clears the housing market.

Having calculated the set of house fixed effects, all of the elements of equation (5.4) are known and we can now calculate the probability that each household i chooses each house h, P^i_h. Following a typical ML approach, we then form the GMM objective function based on maximizing the probability that each household makes its correct housing choice.

$$(5.7) \qquad O = \sum_i \sum_h I^i_h \ln(P^i_h)$$

We then repeat this full process, searching over the interaction parameters, until this objective function is maximized.

Identifying the Social Interactions

By restricting a total of one household to choose each house and matching the individual housing choices as closely as possible, we are able to estimate the majority of the model's parameters. The remaining parameters to be estimated are the components of the house fixed effect shown in equation (5.2). With estimates of the house fixed effects in hand, this equation is simply a regression equation. Consequently, the most obvious approach to identifying the parameters of this equation would be to estimate it using OLS. It is immediately obvious, however, that forming covariance restrictions between x and housing prices or any choice characteristic that depends on neighborhood sociodemographics is not consistent with the logic of the choice model, as changes in the unobserved quality of a house or neighborhood certainly raises demand and alters the composition of households who live in the neighborhood. It is necessary, therefore, to find additional instruments (not included in X_j) to take the place of these endogenous variables in the regression.

We divide the full set of housing choice characteristics into two sets: *endogenous* variables that depend on how households sort across neighborhoods (e.g., school quality, crime, neighborhood sociodemographics, and price), and *exogenous* variables that are relatively fixed (topology, housing stock, air quality, climate, seismology, land use, etc.). In this paper, we assume that the second set of characteristics is uncorrelated with the unobservable, x_h.[19] While

[19] While this may seem reasonable because features such as topology, location, climate, and geology are relatively fixed, these characteristics will certainly be correlated with the unobservable if the unobservable itself depends on the way that household sort across neighborhoods. If, for example, households care about the average income of their neighbors and this is not included in the model, the unobservable is likely to be correlated with every included choice characteristic. In light of this problem, the best we can do is to include as many variables that describe the sociodemographic composition of the community as possible in order to limit the size of this potential bias.

we use similar instruments for price and other neighborhood attributes that depend on sociodemographic characteristics, it is useful to consider these two types of endogenous variables separately.

In constructing an appropriate instrument for housing prices, what is required is a variable that is correlated with the price of a house, but not directly correlated with its unobservable quality. Because the demand for a house is driven by both the specific characteristics of the house as well as the relative scarcity of these characteristics in the market, we develop a set of instruments that characterize how common the exogenous characteristics of the house are relative to other houses not in the same neighborhood but in other reasonably close neighborhoods. Houses that possess scarce characteristics will naturally command higher prices than otherwise similar house positioned in a place in the metropolitan area where such characteristics are not as rare. We form similar instruments for neighborhood sociodemographic characteristics and other attributes that depend on these characteristics. Because any household's demand for a house is also affected by the availability of potential substitutes in the market, the relative scarcity of a house will also affect the sociodemographic characteristics of the households that choose it.

We also form additional instruments for price based on the fact the demand for a house is driven by the preferences of households in the full market, while the actual utility that any particular household gains from a house is driven by the preferences of that household. Specifically, we include in the utility that household i receives from choosing house h a measure of the accessibility of the house to jobs employing individuals with the same education level as household i. In addition to the distance to work measure in the utility function, this term captures the fact that households may place a premium on houses that are accessible to the types of jobs for which they are generally qualified. Thus, while the utility that a particular household receives from choosing a house depends on its own characteristics, the overall demand for the house will be driven by its accessibility more generally. Consequently, we form additional instruments based on measures of the general accessibility of a house.

Summary of Estimation Procedure

1. For a given set of interaction parameters (those in m^i_h), solve for the vector of fixed effects d_h, that implies that one and only one household chooses each house.
2. Using the vector of fixed effects d_h and mi_h, calculate objective function based on the product of the probability that each household makes its correct choice.

---

3. Search over the interaction parameters until objective function in (2) is maximized. The estimated fixed effects are those calculated in (1) on the final iteration.
4. Using the estimated fixed effects from above, estimate equation (5.2) via IV regression.

Comparing This Empirical Framework with a Hedonic Price Regression

In order to see the advantages of the empirical framework that we have just described, it is helpful to make a simple comparison with a standard hedonic price regression. Specifically, consider a specification of the utility function that includes no heterogeneity in household tastes other than the idiosyncratic error term, e^i_h. In other words consider a specification with only a house fixed effect and the idiosyncratic error term:

(5.8) $U^i_h = \delta_h + e^i_h$, where

(5.9) $\delta_h = \alpha_{0X} X_h - \alpha_{0p} p_h + \xi_h$

Given the estimation procedure outlined above, the first step is to solve for the vector of fixed effects that clears the market. In this case, without any additional household heterogeneity, it is immediately obvious that each element of this vector of fixed effect must be identical. Consequently, we can rearrange equation (5.9) as:

(5.10) $p_h = {}^{\alpha_{0X}}\!/_{\alpha_{0p}} X_h + {}^{1}\!/_{\alpha_{0p}} \xi_h$

Equation (5.10) is a standard hedonic price regression. In this way, there is an equivalence between a standard hedonic price regression and a discrete choice framework with no household heterogeneity, a linear utility function, and a basic multinomial logit error structure.[20]

When additional heterogeneity is included in the model, it is easy to view the estimated fixed effects as a correction to the standard hedonic regression that accounts for household heterogeneity and the spatial distribution of employment. In this case, the adjusted price regression is given by:

[20] It is important to note that our estimation procedure would not estimate the hedonic price regression in equation (5.10) using OLS. Instead we use the instrumental variables approach outlined in the previous subsection.

$$(5.11) \quad (p_h + \tfrac{1}{\alpha_{0p}} d_h) = \tfrac{\alpha_{0x}}{\alpha_{0p}} X_h + \tfrac{1}{\alpha_{0p}} x_h$$

To see how this adjustment works, consider a simple example with two houses located in two towns separated by a distance of 10 miles. If both households work in one of the towns and commuting distance enters utility negatively, the estimation procedure will need to assign a greater fixed effect to the house in the non-work town in order to explain why one household chooses each house. In this way, the adjustment shown in equation (5.11) would raise the price of the house non-work town and lower the price of the house in the town where the households work. In the adjusted hedonic price regression, then, this change amounts to controlling for that part of the differences in prices across locations due to variation in employment access.

6 Results (Preliminary)

The results presented in this version of the paper are preliminary and meant primarily to illustrate the general nature of the analysis that we are currently undertaking. Two important restrictions to the general model and dataset described above apply to these results. First, we have estimated the model for a sample of 10,000 households drawn at random from San Francisco County. While we could have drawn the sample from 6-county region of the Bay Area that makes up our full sample, we chose a smaller geographic area in order to be sure to have enough households in each neighborhood when forming estimates of neighborhood sociodemographic compositions in the simulations. The second main restriction in the analysis that follows is that we have not yet estimated separate crime and education production functions. Consequently, for the simulations that we conduct, the only choice characteristics that are affected by the way households sort are housing prices and neighborhood sociodemographic characteristics. We begin this section by presenting the parameter estimates for the most comprehensive specification that we have estimated to date. After describing the general procedure for the general equilibrium simulations that we conduct, we then present the results of these simulations.

Parameter Estimates

The estimated parameters of the model are shown in Table 6.1. As the number of parameters in the table indicates, the heterogeneous coefficients model of equations (3.1)-(3.2) allows for great variation in the preferences of households for the characteristics of their housing choice. The main portion of the table displays the interaction terms from the utility function, which describe a specific taste parameter associated with the household characteristic shown in

the row for the housing choice characteristic shown in the column. The first entry in the table, for example, shows the interaction between households with education *HS Degree or Less* and *Average Test Score*. The fact that this term is negative implies that, controlling for all of the other factors included in the model, households with a high school degree or less have weaker preferences for the achievement of the neighborhood school than the baseline group, households with some college but without a college degree.

The final row of Table 6.1 shows the IV estimates from the regression of the house fixed effects on the housing choice characteristics. These terms measure baseline preferences, i.e., the preferences of a household with values of zero for all of the included household characteristics. The interaction terms just described, then, increment these baseline parameters if the household has the characteristic in question. For example, the preference parameter on *Average Test Score* for a household with baseline characteristics but with an education level of *College Degree or More* would be 0.038 = 0.025 (baseline) + 0.013 (interaction). The preference parameter for an identical household with the baseline education level (*Some College*), on the other hand, would be simply the baseline: 0.025.

Scanning the housing choice characteristics included in the model, notice that we have included a tenure dummy variable (1 if house is a rental unit) and have also provided separate terms for house characteristics (including price) for rental units and owner-occupied units. In this analysis, we treat the tenure status as an attribute of the housing unit. For owner-occupied units, the price equals the value of the house divided by 100 and for rental units the price equals the monthly rent. This specification of the model has three advantages. First, it avoids making an arbitrary conversion between housing values and rent. In general, house values depend on both current and expected future streams of rents, and consequently, any conversion between these two measures would need to account for expectations about the future demand for the housing unit in question. Second, because the coefficients on price vary with tenure, the estimates provide separate measures of a household's willingness to pay for neighborhood attributes in terms of monthly rent if the household is renting and housing value if the household owns the house. Finally, by estimating separate terms for housing characteristics (such as Number of Rooms) depending on tenure status, this specification accounts for the fact that it may be difficult to compare the features of renter- versus owner-occupied units.

In examining the estimated parameters, it is useful to start with the parameters on price. The parameters describe how a household trades off between housing and all other consumption. The positive terms on the income price interactions, for example, imply that higher income households would be willing to spend more for an increase in housing quality (i.e., an increase in

the unobservable x) than an otherwise identical household with less income. In this way, these positive interactions imply that housing quality (which includes both house and neighborhood quality) is a normal good, and, consequently, the model predicts that higher income households purchase more expensive housing in equilibrium. Importantly, because the income-price interactions control for the general increase in willingness to pay associated with income, the income interactions with other housing choice characteristics characterize how a household's preferences for these characteristics change with income above and beyond a normal income effect that operates through the budget constraint. In this way, it is possible for the demand for some aspect of neighborhood or housing to decline with income despite the fact that demand for housing quality more generally increases with income.

Willingness to Pay Calculations

While a quick scan of Table 6.1 indicates that most of the included interactions are statistically significant and have the expected sign, in order to get a sense of the economic significance of these parameters estimates, it is helpful to convert the estimates of the indirect utility function parameters into willingness to pay measures. We present these calculations in Table 6.2a and 6.2b. In particular, the first row of Table 6.2a describes how much a household with the mean set of household characteristics included in the model would be willing to pay in terms of monthly rent for a one standard deviation increase in the housing choice characteristics associated with each column. Each of these willingness-to-pay calculations is conducted using the mean housing choice characteristics for all renter-occupied units in the sample. In this way, the first entry of Table 6.2a implies that the average household in the sample would be equally well off with a housing unit with the average characteristics of all renter-occupied houses in the sample or by paying an additional $XXX in monthly rent for the same house with a one standard deviation increase in the average test score of the neighborhood school. The remaining rows of the table describe the incremental willingness to pay associated with particular household characteristics. The second row of Table 6.2a shows, for example, that Asian households would be willing to pay $100 more in monthly rent for a one standard deviation increase in the average test score of the neighborhood school than otherwise identical non-Asian households. Table 6.2b repeats these willingness-to-pay calculations for owner-occupied units measured in terms of house value.

General Equilibrium Simulations

We now use the estimated utility function parameters to conduct a series of general equilibrium simulations designed to explore the causes and consequences of residential segregation on the basis of race. Each of these simulations begins by changing a key primitive of the model. In one of the simulations, for example, we set all preferences for the neighborhood racial composition to zero. In this new counterfactual environment, we calculate a new equilibrium for the model, the conditions for which are the same as those outlined in Section3. Specifically, an equilibrium consists of a set of housing choice probabilities for each household and a set of housing prices such that (i) each household's decision is optimal given the decisions of all other households, and (ii) the set of housing prices clears the market. To calculate a new equilibrium, then, we must find a set of housing prices that clears the market and find a fixed point of the neighborhood sociodemographic mapping defined in equation (3.9). That is, given a set of sociodemographic characteristics for each neighborhood and the vector of market clearing prices, the household location decisions predicted by the model must aggregate up to these same sociodemographic characteristics for each neighborhood.

The basic structure of the simulations, then, consists of a loop within a loop. The outer loop calculates the sociodemographic composition of each neighborhood given a set of prices and an initial sociodemographic composition of each neighborhood. The inner loop calculates the unique set of prices that clears the housing market given an initial sociodemographic composition for each neighborhood. In this way, for any change in the primitives of the model, we first calculate a new set of prices that clears the market. As discussed in Section 3, Berry (1994) ensures that there is a unique set of market clearing prices. Using these new prices and the initial sociodemographic composition of each neighborhood, we then calculate the probability that each household makes each housing choice and aggregating these choices to the neighborhood level calculate the predicted sociodemographic composition of each neighborhood. We then replace the initial neighborhood sociodemographic measures with these new measures and start the loop again – i.e., calculate a new set of market clearing prices with these updated neighborhood sociodemographic measures. We continue this process until the neighborhood sociodemographic measures converge. The set of household location decisions corresponding to these new measures along with the vector of housing prices that clears the market then represents the new equilibrium.

<u>Summary of the Calculation of the New Equilibrium</u>

1. Incorporate change to the primitives of the model corresponding to simulation, start with initial measures of sociodemographic characteristics for each neighborhood.
2. Calculate unique vector of housing prices that clears the housing market (i.e., ensures that a total of one household chooses each house).
3. Using new vector of housing prices, calculate housing choice probabilities for each household.
4. Aggregating these probabilities to the neighborhood level, calculate new sociodemographic characteristics for each neighborhood. Replace existing measures of neighborhood sociodemographics in utility function with these new measures.
5. Repeat steps (2)-(4) until the sociodemographic characteristics calculated in (4) converge.

It is important to point out that because the model itself does not perfectly predict the housing choices that individuals make, the neighborhood sociodemographic measures initially predicted by model, $\overline{Z}_n^{PREDICT}$, will not match the actual scoiodemographic characteristics of each neighborhood, $\overline{Z}_n^{ACTUAL}$. Consequently, before calculating the new equilibrium for any simulation we first solve for the initial prediction error associated with each neighborhood *n*:

$$(7.1) \qquad w_n = \overline{Z}_n^{ACTUAL} - \overline{Z}_n^{PREDICT}$$

In solving for the new equilibrium, we add this initial prediction error w_n to the sociodemographic measures calculated in each iteration before substituting these measures back into the utility function.

Simulation Results

Because the parameter estimates presented in the paper are preliminary, we present simulation results primarily to demonstrate the economic analysis that we intend to carry out. There are two principal changes associated with the calculation of the new housing market equilibrium. First, we solve for a new price for every house in the market. Second, we solve for a new set of location decisions for each household in the sample, which in turn affect the sociodemographic composition of each neighborhood. In comparing this new equilibrium to the initial one, we first consider the difference in the level of racial segregation under the two scenarios. This difference highlights the importance of the factor that we have changes, preferences for neighborhood racial composition in this case, in driving the observed pattern of racial segregation. By analyzing the results from a wide range of simulations, examining the

change in the level of segregation under each scenario provides insights into the underlying causes of racial segregation. Using the new location decisions of each household in the sample, we then examine how housing prices, housing quality, school quality, commutes, tenure rates, and overall utility are affected for each household. These measures provide a wide variety of measures of the consequences of the experiment in question for households with different characteristics.

The results of the simulations that we have conducted are presented in the first panel of Table 6.3. The first simulation calculates the new equilibrium that would arise if all household preference parameters associated with neighborhood racial composition were identically zero. The second simulation randomly draws an income for each household from the empirical income distribution and the third simulation eliminates commuting distance as a factor in the location decision.

7 Conclusion

In this paper, we provide a comprehensive framework for analyzing both the causes and consequences of segregation in a general equilibrium. Because households are able to respond to policy (e.g., school desegregation orders) through their decision of where to live, accounting for such equilibrium changes in the metropolitan area housing market is crucial for conducting meaningful policy analysis. While the results of the analyses conducted in this paper are still too preliminary to draw firm conclusions at this time, further analysis should provide a number of new insights into the underlying causes of racial segregation and the impact of sorting on the basis of race has for the welfare of households from different backgrounds.

References

Anas, Alex, (1980), "A Model of Residential Change and Neighborhood Tipping," *Journal Urban Economics*, 7: 358-70.

Anas, Alex, (1982), *Residential Location Markets and Urban Transportation: Economic Theory, Econometrics and Public Policy Analysis*, Academic Press, New York.

Anas, Alex, and Chausie Chu, (1984), "Discrete Choice Models and the Housing Price and Travel to Work Elasticities of Location Demand," *Journal of Urban Economics*, Vol 15, pp. 107-123.

Bayer, Patrick, (2001), "What Drives the Family-School Match? An Equilibrium Analysis of Community Sorting," mimeo, Yale University.

Bayer, Patrick, Robert McMillan, and Kim Rueben, (2001), "Household Sorting and Residential Segregation: Evidence from the San Francisco Bay Area," mimeo, Yale University.

Bayer, Patrick and Christopher Timmins, (2001a), "Identifying Social Interactions in Endogenous Sorting Models," mimeo, Yale University.

Bayer, Patrick and Christopher Timmins, (2001b), "," mimeo, Yale University.

Benabou, Roland, (1996), "The Workings of a City."

Berry, Steven, (1994), "Estimating Discrete-Choice Models of Product Differentiation," *RAND Journal of Economics*, Vol. 25, pp. 242-262.

Berry, Steven, James Levinsohn, and Ariel Pakes, (1995), "Automobile Prices in Market Equilibrium," *Econometrica*, Vol 63, pp. 841-890.

Berry, Steven, James Levinsohn, and Ariel Pakes, (1998), "Differentiated Products Demand Systems from a Combination of Micro and Macro Data: Autos Again," mimeo, Yale University.

Blackley, Paul and Jan Ondrich, (1988), "A Limiting Joint-Choice Model for Discrete and continuous Housing Characteristics," *The Review of Economics and Statistics*, Vol. 70, pp. 266-274.

Borjas, George J., (1998), "To Ghetto or Not to Ghetto: Ethnicity and Residential Segregation." *Journal of Urban Economics*, 44: 228-253

Cutler, David and Edward Glaeser, (1997), "Are Ghettos Good or Bad?" *Quarterly Journal of Economics*, August: 826-72.

Ellickson, B., (1981), "An Alternative Test of the Hedonic Theory of Housing Markets," *Journal of Urban Economics*, Vol. 9, pp. 56-79.

Epple, D., R. Filimon, and T. Romer, (1984), "Equilibrium Among Local Jurisdictions: Towards an Integrated Approach of Voting and Residential Choice," *Journal of Public Economics*, Vol. 24, pp. 281-304.

Epple, D., R. Filimon, and T. Romer, (1993), "Existence of Voting and Housing Equilibrium in a System of Communities with Property Taxes," *Regional Science and Urban Economics*, Vol. 23, pp. 585-610.

Epple, Dennis, and Holger Sieg, (1999), "An Approach for Estimating Spatial Models," *Journal of Political Economy*..

Harsman, Bjorn and John M. Quigley, (1995) "The Spatial Segregation of Ethnic and Demographic Groups: Comparative Evidence from Stockholm and San Francisco," *Journal of Urban Economics*, 37: 1-16.

Inman, Robert and Daniel Rubinfeld, (1979), "The Judicial Pursuit of Fiscal Equity," *Harvard Law Review*, 92: 1662-750.

Massey, Douglas S., and Nancy A. Denton, (1989), "Hypersegregation in United States Metropolitan Areas – Black and Hispanic Segregation along Five Dimensions," *Demography*, 26: 373-91.

McFadden, Daniel, (1973), "Conditional Logit Analysis of Qualitative Choice Behavior," in P. Zarembka, eds., *Frontiers of Econometrics*, Academic Press, New York.

McFadden, Daniel, (1978), "Modeling the Choice of Residential Location," in eds. Karlquist, A., et al., Spatial Interaction Theory and Planning Models, Elssevir North-Holland, New York.

Nechyba, Thomas J., (1997), "Existence of Equilibrium and Stratification in Local and Hierarchical Tiebout Economies with Property Taxes and Voting," *Economic Theory*, Vol. 10, pp. 277-304.

Nechyba, Thomas J., (1999), "School Finance Induced Migration and Stratification Patterns: the Impact of Private School Vouchers," *Journal of Public Economic Theory*, forthcoming.

Nechyba, Thomas J., and Robert P. Strauss, (1998), "Community Choice and Local Public Services: A Discrete Choice Approach," *Regional Science and Urban Economics*, Vol. 28, pp. 51-73.

Pollakowski, Henry O., (1973), "The Effects of Property taxes and Local Public Spending on Property Values: A Comment and Further Results," *Journal of Political Economy*, Vol. 81, pp. 994-1003.

Quigley, John M., (1985), "Consumer Choice of Dwelling, Neighborhood, and Public Services," *Regional Science and Urban Economics*, Vol. 15(1).

Rivkin, Steven, (1994), "Residential Segregation and School Integration," *Sociology of Education*, 67: 279-92.

Schelling, Thomas C., (1969), "Models of Segregation." *American Economic Review*, 59(2): 488-93.

Schelling, Thomas C., (1978), *Micromotives and Macrobehavior*, Norton: New York.

Tiebout, Charles M., (1956), "A Pure Theory of Local Expenditures," *Journal of Political Economy*, 64: 416-424.

Table 6.0: Summary Statistics for San Francisco Subsample (10,000 households)

	MEAN	SD	MIN	MAX
Housing Choice Characteristics				
% census block gr. black	0.08	0.16	0.00	1.00
% census block gr. asian	0.21	0.19	0.00	1.00
% census block gr. hispanic	0.09	0.13	0.00	0.80
1 if unit is owned	0.37	0.48	0.00	1.00
house value if owned	339,837	229,377	100,000	550,000
number of rooms if owned	5.69	3.63	1.00	9.00
number of rooms if rental	3.37	2.46	1.00	9.00
monthly rent if rental	605.41	428.32	100.00	1000.00
elevation	228.90	117.96	0.00	820.00
% census block gr. High education	35.15	24.49	0.00	137.00
% census block gr. High income	10.44	9.12	0.00	53.00
average math score	189.31	8.74	156.02	239.35
employment access index	21.61	69.15	0.00	2032.00
pupil teacher ratio	22.92	0.82	16.63	27.82
polution index	25.68	4.28	17.26	36.95
Household Characteristics				
1 if black	0.08	0.28	0.00	1.00
1 if hispanic	0.09	0.29	0.00	1.00
1 if asian	0.21	0.41	0.00	1.00
household income	46,544	60,858	-9,600	1,912,524
1 if working	0.64	0.48	0.00	1.00
1 if low education level	0.36	0.48	0.00	1.00
1 if high education level	0.45	0.50	0.00	1.00
age	48.86	18.16	16.00	111.00
number of persons	2.31	1.61	1.00	15.00
1 if spanish spoken in home	0.09	0.29	0.00	1.00
1 if asian language spoken in home	0.19	0.40	0.00	1.00
1 if kids<18y old in household	0.21	0.41	0.00	1.00
distance to work (miles)	3.28	6.30	0.00	72.44

Table 6.1: Utility Function Parameter Estimates

	Neighborhood Sociodemographic Characteristics					House Characteristics					Other Neighborhood Attributes				
	% Block Black	% Block Asian	% Block Hispanic	% Block High Education	% Block High Income	Owner-Occupied Unit	House Value (/100,000)	Rooms if owned	Rooms if rental	Monthly Rent (/1000)	Elevation (/1000)	Pollution Index	Average Math Score	Pupil-Teacher Ratio	Distance To Work
Black	**16.219**	**2.687**	**4.459**	**1.375**	-1.169	**-1.237**	**-0.179**	-0.004	0.030	**-1.627**		-0.049	-0.026	-0.102	
	(0.271)	(0.354)	(0.457)	(0.408)	(1.694)	(0.333)	(0.066)	(0.046)	(0.037)	(0.210)		(0.037)	(0.026)	(0.302)	
Hispanic	**4.113**	**1.878**	**12.874**	0.408	0.860	**-1.832**	-0.014	**-0.142**	**-0.404**	-1.057		-0.065	**0.059**	**-0.442**	
	(1.729)	(1.048)	(0.507)	(0.904)	(3.054)	(0.578)	(0.117)	(0.037)	(0.138)	(0.679)		(0.054)	(0.029)	(0.108)	
Asian	1.214	**7.489**	**0.966**	0.404	-0.258	-0.330	**0.212**	**-0.422**	**-0.384**	**-1.560**		0.012	0.018	0.034	
	(1.205)	(0.967)	(0.503)	(0.437)	(1.810)	(0.257)	(0.091)	(0.015)	(0.084)	(0.512)		(0.040)	(0.020)	(0.140)	
Hhld Income (/100,000)	-0.381	**-0.791**	**-1.381**	**-1.061**	**5.565**	**11.102**	**0.878**	**0.110**	**0.901**	**10.268**	-0.329	0.012	0.008	-0.193	-0.005
	(0.732)	(0.172)	(0.594)	(0.425)	(0.678)	(0.244)	(0.066)	(0.012)	(0.040)	(0.298)	(0.198)	(0.018)	(0.010)	(0.127)	(0.013)
Working															**-0.184**
															(0.007)
Low Education	-0.160	**1.417**	1.278	-0.394	-1.124							0.019	0.014	-0.314	
	(0.763)	(0.443)	(0.791)	(0.405)	(1.076)							(0.011)	(0.011)	(0.117)	
High Education	**-0.686**	**-1.292**	0.222	**1.340**	**-1.596**							**-0.020**	**0.036**	**-0.133**	
	(0.278)	(0.217)	(0.373)	(0.215)	(0.587)							(0.008)	(0.006)	(0.065)	
Age						**0.026**	**-0.005**	**0.007**	**0.002**	**-0.098**			**0.000**	**0.008**	
						(0.003)	(0.001)	(0.000)	(0.001)	(0.003)			(0.000)	(0.001)	
# of Persons						**0.877**	**-0.146**	**0.191**	**0.292**	**-0.079**					
						(0.068)	(0.013)	(0.010)	(0.008)	(0.045)					
Spanish Language	0.993	0.636	**4.841**										**-0.058**	0.225	
	(0.469)	(0.447)	(0.442)										(0.015)	(0.163)	
Asian Language	**3.470**	**4.335**	**3.707**										0.010	-0.039	
	(0.622)	(0.276)	(0.688)										(0.008)	(0.047)	
Kids < 18 yr old													**-0.036**	0.484	
													(0.006)	(0.065)	
BASELINE (IV)	**-3.728**	**-5.248**	**-3.756**	-0.643	-2.554	**8.243**	**-2.324**	**0.526**	0.072	1.301	0.140	**-0.126**	**0.031**	0.036	
	(0.775)	(0.695)	(1.015)	(0.559)	(1.447)	(0.717)	(0.286)	(0.104)	(0.085)	(1.285)	(0.536)	(0.024)	(0.018)	(0.223)	
BASELINE (OLS)	-0.822	-4.136	-1.925	-0.797	-0.538	5.057	-0.422	-0.076	0.065	1.296	-0.090	-0.155	-0.004	0.218	
	(0.388)	(0.337)	(0.500)	(0.376)	(0.877)	(0.382)	(0.071)	(0.058)	(0.046)	(0.280)	(0.481)	(0.018)	(0.010)	(0.118)	

Note: Standard Errors in Parentheses. Bold-Face inidcates statistical significance at 95% level.

Baseline coefficients are reported for household with following characteristics: (White, $46,500, Not Working, Middle Educ, 49 years old, 2.3 Persons, English, No Kids)

Table 6.2: Willingness to Pay Measures (Owner-Occupied Houses)

	Willingness to Pay for One Standard Deviation Increase										
Variable	% Block Black	% Block Asian	% Block Hispanic	% Block High Education	% Block High Income	One Additional Room	Elevation (/1000)	Pollution Index	Average Math Score	Pupil-Teacher Ratio	Distance To Work
Standard Dev.	0.16	0.19	0.13	0.2449	0.0912		0.11796	4.28	8.74	0.82	6.3
BASELINE (White, $46,500, Not Working, Middle Educ, 49 years old, 2.3 Persons, English, No Kids)											
	-25,661	-42,900	-21,004	-6,776	-10,020	22,626	711	-15,689	2,813	559	
SAME AS BASELINE BUT:											
Black	79,844	-19,439	3,652	7,164	-13,565	20,835		-20,257	423	-958	
Hispanic	2,633	-27,385	50,697	-2,466	-6,608	16,427		-23,610	8,058	-6,334	
Asian	-19,040	20,148	-17,165	-2,778	-12,138	4,938		-15,639	4,899	1,206	
Hhld Income (+10,000)	-26,941	-45,256	-22,631	-8,204	-8,144	24,005	565	-23,837	12,521	598	-3
Working											-934
Low Education	-26,765	-31,319	-13,856	-10,926	-14,430			-13,355	4,030	-4,354	
High Education	-30,382	-53,461	-19,763	7,339	-16,282			-18,131	6,033	-1,529	
Age (+10 years)						24,879			3,141	1,763	
Persons (+1)						29,014					
Spanish Language	-18,823	-37,704	6,069						-2,424	4,078	
Asian Language	-1,776	-7,466	-271						3,694	-55	
Kids < 18 yr old									-449	8,144	

Note: Figures represent willingness to pay in terms of house value for a one standard deviation increase in the housing choice characteristic shown in column.
Baseline figures are reported for household with following characteristics: (White, $46,500, Not Working, Middle Educ, 49 years old, 2.3 Persons, English, No Kids).
The other figures are for households with same characteristics as the baseline household except for the characteristic shown in the row.