NBER WORKING PAPER SERIES

SEGMENTED HOUSING SEARCH

Monika Piazzesi
Martin Schneider
Johannes Stroebel

Working Paper 20823
http://www.nber.org/papers/w20823

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2015

We thank Ed Glaeser, Bob Hall, Giuseppe Moscarini, Robert Shimer, Vincent Sterk and Silvana Tenreyro for helpful comments. We also thank conference and seminar participants at the AEA meetings, Arizona State, Berkeley, Chicago, Mannheim, LBS, Michigan, Minneapolis Fed, Minnesota, NBER Summer Institute, Northwestern, Tsinghua, NYU, Philadelphia Fed, SITE, SED, Stanford, Yale, UCLA, and USC. This research is supported by a grant from the National Science Foundation. We are grateful to Trulia for providing the data, which was obtained through a consultancy agreement. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w20823.ack

Segmented Housing Search
Monika Piazzesi, Martin Schneider, and Johannes Stroebel
NBER Working Paper No. 20823
January 2015
JEL No. E00,G12,R30

ABSTRACT

This paper studies housing markets with multiple segments searched by heterogeneous clienteles. We document market and search activity for the San Francisco Bay Area. Variation within narrow geographic areas is large and differs significantly from variation across those areas. In particular, search activity and inventory covary positively within cities and zip codes, but negatively across those units. A quantitative search model shows how the interaction of broad and narrow searchers drives housing market activity at different levels of aggregation and shapes the response to shocks as well as price discounts due to market frictions.

Monika Piazzesi
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305-6072
and NBER
piazzesi@stanford.edu

Martin Schneider
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305-6072
and NBER
schneidr@stanford.edu

Johannes Stroebel
New York University
Leonard N. Stern School of Business
44 West 4th Street, KCM 9-98
New York, NY 10012
johannes.stroebel@nyu.edu

Segmented Housing Search*

Monika Piazzesi
Stanford & NBER

Martin Schneider
Stanford & NBER

Johannes Stroebel
New York University

December 2014

Abstract

This paper studies housing markets with multiple segments searched by heterogeneous clienteles. We document market and search activity for the San Francisco Bay Area. Variation within narrow geographic areas is large and differs significantly from variation across those areas. In particular, search activity and inventory covary positively within cities and zip codes, but negatively across those units. A quantitative search model shows how the interaction of broad and narrow searchers drives housing market activity at different levels of aggregation and shapes the response to shocks as well as price discounts due to market frictions.

1 Introduction

Home buyers typically look for properties in a search range that depends on their geographic preferences, budget, or family size. For example, they might only consider houses in a certain price range that are also within reasonable commuting distance from their workplace. A family with children might in addition require that the house be located in a good school district. An individual property that comes on the market is thus considered by a clientele of potential buyers whose search ranges contain that property. The interaction of these clienteles then determines how turnover, inventory and prices differ across segments of the housing market.

Existing studies of housing search typically assume that clienteles are homogeneous. In particular, two common assumptions are that the market under consideration is either fully integrated – that is, the clientele for each house consists of all potential buyers – or that it is perfectly segmented, that is, it can be partitioned into submarkets, each with its own buyer type who considers all houses in the submarket. However, when studying, say, a metro area, homogeneity of clienteles

*Addresses: piazzesi@stanford.edu, schneidr@stanford.edu, johannes.stroebel@nyu.edu. We thank Ed Glaeser, Bob Hall, Giuseppe Moscarini, Robert Shimer, Vincent Sterk and Silvana Tenreyro for helpful comments. We are grateful to Trulia for providing data. We also thank conference and seminar participants at the AEA meetings, Arizona State, Berkeley, Chicago, Mannheim, LBS, Michigan, Minneapolis Fed, Minnesota, NBER Summer Institute, Northwestern, Tsinghua, NYU, Philadelphia Fed, SITE, SED, Stanford, Yale, UCLA, and USC. This research is supported by a grant from the National Science Foundation.

is not a priori obvious. For example, a neighborhood with good schools might see competition between families with children – who search narrowly in that neighborhood – and other potential buyers, who search more broadly. More generally, the distribution of workplaces, preferences over amenities, and commuting costs in the population is likely to generate clienteles whose search ranges only partially overlap.

This paper considers housing search, trading and valuation in interconnected housing market segments with heterogeneous clienteles. We introduce a novel dataset on housing search behavior in the San Francisco Bay Area to document stylized facts on the search ranges of home buyers. When we divide the Bay Area into market segments based on these observed search ranges, we find substantial heterogeneity not only for market outcomes across segments but also for clienteles both within and across segments. We then use a search model with multiple segments to relate market outcomes to the distribution of preferences and the matching technology. We show that the interaction of heterogeneous clienteles is a quantitatively important force in the housing market.

We infer search ranges from online housing search via the real estate website trulia.com. Home searchers on trulia.com can set an email alert that triggers an email whenever a house with their desired characteristics comes on the market. We find that housing search occurs predominantly along three dimensions: geography, price and, to a lesser extent, house size as captured by the number of bathrooms. Most searchers look for houses in contiguous areas that differ in geographic breadth. In cheaper urban areas, there are fewer searchers per house; those searchers consider broad search ranges to look for low prices. In contrast, clienteles in more expensive and more suburban areas tend to be larger but also more selective.

To analyze market activity, we divide the San Francisco Bay Area into 564 distinct market segments along dimensions suggested by the observed search behavior. We then measure the cross section of turnover and inventory at the segment level by matching search alert data to deeds and assessment records as well as feeds of "for sale"listings. We find that about half of the variation in market activity occurs within zip codes, our finest geographic unit. Inventory and turnover comove strongly, both at the segment level and when we aggregate to the zip code or city level. In particular, in cheaper areas or segments, houses turn over faster, but there is also more inventory for sale.

To relate market and search activity, we express search ranges as subsets of the set of all segments, resulting in 11,503 distinct ranges. We measure search activity at the segment level in terms of searchers per house. We find that the relationship between inventory and search activity depends critically on the level of aggregation. Across cities, inventory and search activity are inversely related – in other words, the "Beveridge curve"slopes *down across cities*. For example, in expensive cities like San Francisco, many people search scarce inventory, while in cheaper cities like San Jose, plenty of inventory is considered by few searchers. In contrast, the Beveridge curve slopes *up within most cities*: for example, cheaper segments within San Francisco have higher inventory and are considered by more searchers.

Our model exercise builds on a version of the Diamond-Mortensen-Pissarides random matching model with fixed numbers of both houses and agents. Moving shocks induce agents to sell their current house (at a cost) and search for another house. What is new in the model is the presence of multiple market segments as well as heterogeneous agent types identified by search ranges – subsets of the set of all segments, as in our data. While matching is random, agents are more likely to match in those segments within their search range where inventory is higher. Prices reflect the present value of housing services less a discount due to search and transaction costs.

The equilibrium of the model relates the cross sectional distribution of turnover, inventory, price and search activity to the distribution of preferences, moving shocks and the matching technology. The distribution of preferences – including search ranges – allows the model to capture the rich clientele patterns we observe in the data. The key theoretical effect added by heterogeneous clienteles is that broad searchers flow to high inventory segments and compete with narrow searchers there. This effect is stronger in more integrated areas, for example within cities that have a larger share of broad searchers. It implies that the nature of clientele patterns matters both for how market activity responds to changes in the economic environment and for what we can infer about parameters from the cross section of market activity.

If there is perfect segmentation, then identification of the three forces that drive segment heterogeneity is straightforward. In more stable segments – where moving shocks arrive less frequently – turnover and inventory are both lower. In more liquid segments – where matching is faster holding fixed the buyer and seller pools – turnover is also higher but inventory is lower. The same is true in more popular segments that have more potential buyers per house. However, more popular segments also see more search activity.

Our quantitative exercise suggests that patterns at the city level are driven by differences in popularity and stability. More expensive cities like San Francisco are both more stable and more popular than cheaper cities like San Jose. The former explains why turnover and inventory are both lower in San Francisco. The latter explains why search activity is higher there as well; it also helps generate a downward sloping Beveridge curve in the cross section of cities.

Within cities, the Beveridge curve is affected not only by the correlation of exogenous forces, but also by the endogenous interaction of heterogeneous clienteles. Indeed, consider two segments that are equally popular and differ only in stability. Broad searchers who scan both segments will tend to flow to the less stable segment where inventory is higher. As a result, narrow searchers in the unstable segment find it harder to find a house and must search more. Within partially integrated areas such as cities, differences in stability alone thus generate an upward sloping Beveridge curve. We show that this endogenous response of broad searchers is quantitatively important.

We use our estimated parameters to infer liquidity discounts for houses across segments. These discounts are quantitatively large, between 10 percent and 40 percent of the frictionless house value (defined as the present discounted value of future housing services by the house.) The liquidity discounts are larger in segments that are less stable, where houses turn over more often. They are

also larger in illiquid segments, where houses take a long time to sell for whatever reason. High turnover and high time on market increase the value of the trading frictions that the current and future buyers face, which amount to the liquidity discount.

Finally, we illustrate the role of search patterns for the transmission of shocks with comparative statics exercises. In particular, we ask how time on market and inventory change if a segment becomes more popular in the sense that more buyers want to narrowly buy there. The answer crucially depends on the number of searchers and what other markets those searchers look at. For example, shocks to a downtown San Francisco segment with many searchers who search broadly are transmitted widely across the city. In contrast, shocks to a suburban segment close to the San Francisco city boundary have virtually no effect on the market in the city itself.

These results suggest that the impact of policies on local housing markets will depend crucially on the characteristics of the local clientele. For example, whether local zoning regulations to increase the available housing stock in a neighborhood have a large impact on local prices and time on market will depend on whether a neighborhood is considered primarily by narrow or broad searchers.

Related Literature

Our paper contributes to a growing body of empirical work that analyzes housing market activity in the cross section. The typical approach is to sort houses within geographic units (such as cities) based on similarity along property characteristics, and to then compare measures of market activity such as turnover and time on market across these segments (see, for example, Goodman and Thibodeau (1998), Leishman (2001), as well as the survey by Islam and Asami (2009)). There is, however, little direct evidence on housing search behavior. An exception is Genesove and Han (2012) who build a time series for search activity at the city level from survey data on buyers' house hunting experience. Our paper offers a new source of demand side information and uses the distribution of online searchers' criteria to define segments for which market activity is then measured. Our approach emphasizes the heterogeneity of market and search activity within cities.

There is a large literature on search models of housing (see Han and Strange (2014) for a comprehensive survey). Recent studies have turned to quantitative evaluations using micro data, for example Diaz and Perez (2013), Burnside, Eichenbaum and Rebelo (2014), Head, Lloyd-Ellis and Sun (2014), Ngai and Tenreyro (2014), Anenberg and Bayer (2014), Halket and Pignatti (2014) and Guren and McQuade (2014). Most of these authors are interested in how search modifies the time series dynamics of house prices and market activity in homogeneous housing markets.[1] In contrast, our focus is on steady states of markets with rich cross sections of houses and buyers. Our theoretical model builds on earlier random matching models such as Wheaton (1990), Krainer (2001), Albrecht, Anderson, Smith and Vroman (2007), Novy-Marx (2009), Piazzesi and Schneider

[1] Davis and Van Nieuwerburgh (2014) survey the broader literature on business cycles, asset prices and housing, including studies that do not rely on search frictions.

(2009); the new element is that we allow for matching across multiple interconnected segments.

In the labor search literature, Manning and Petrongolo (2011) estimate a search and matching model for local labor markets. They also divide space into many small areas (U.K. wards in their case, as supposed to U.S. zip codes in ours) and allow searchers to look in many areas simultaneously, thus generating spillover effects across areas. Their demand side data consists of addresses for both a job seeker's home as well as the vacancies he applies to. They infer a distribution of preferences that consists of each searcher's home address as well as a common parameter that determines how quickly utility declines with commuting time. Our approach puts less structure on utility in order to exploit our detailed search data: we infer the distribution of preferences from the nature of online search behavior, using both spatial and quality information to define the commodity space.

Our paper also contributes to a recent literature that considers house valuation within metro areas. For example, Poterba, Weil and Shiller (1991) consider the role of demographics for prices, Bayer, Ferreira and McMillan (2007) look at school quality, and Hurst, Guerrieri and Hartley (2014) study the effects of gentrification. Stroebel (2014) and Kurlat and Stroebel (2014) investigate asymmetric information about property and neighborhood characteristics, respectively. Stein (1995) as well as Mian and Sufi (2011) emphasize the role of credit constraints. Landvoigt, Piazzesi and Schneider (2014) study the effect of credit constraints on prices in an assignment model with many quality segments. They consider competitive equilibria of a model with homogeneous preferences. In contrast, this paper emphasizes liquidity discounts due to search frictions and transactions costs, as well as heterogeneity in preferences within a metro area.

2 Dimensions of Housing Search

In this section we document housing search behavior in the San Francisco Bay Area using email alerts set on the real estate website trulia.com. We first describe the data and then provide summary statistics on the major search dimensions. The discussion provides answers to three broad questions. First, can search ranges inferred from trulia.com email alerts be plausibly interpreted as reflecting the considerations of a typical home buyer? Second, how much heterogeneity in search behavior do we observe? Third, is there a simple way to describe search ranges as subsets of a space of characteristics (including geography and quality)?

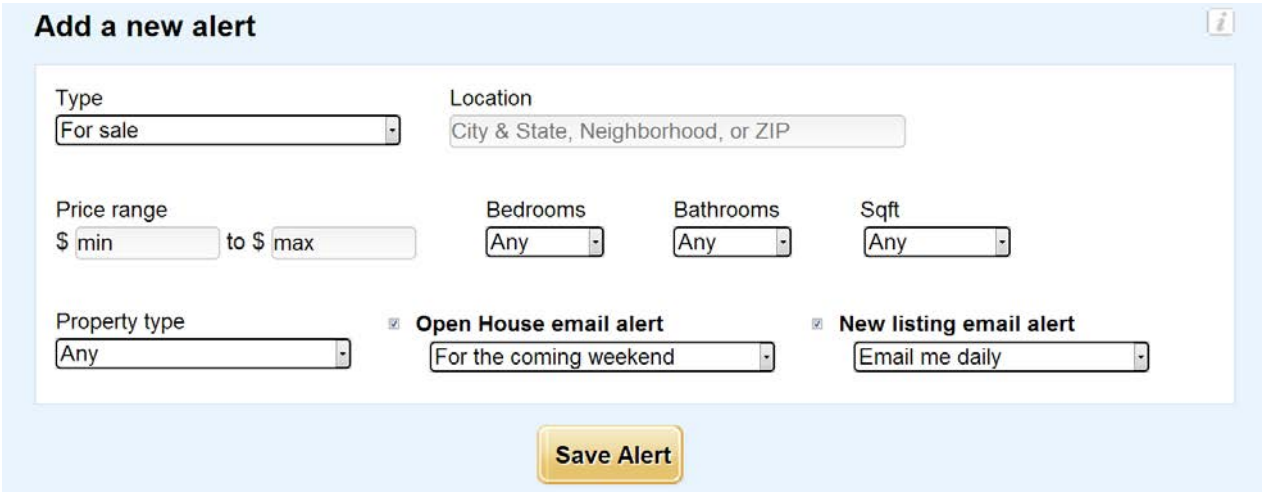
The San Francisco Bay Area is a major urban agglomeration in Northern California that includes the cities of San Francisco, San Jose and Oakland. Our analysis combines data on two Metropolitan Statistical Areas (MSAs) bordering the San Francisco Bay. The San Francisco-Oakland-Hayward, CA Metropolitan Statistical Area comprises Alameda, Contra Costa, San Francisco, San Mateo, and Marin counties. The San Jose-Sunnyvale-Santa Clara, CA Metropolitan Statistical Area consists of Santa Clara and San Benito counties. As of the 2010 Census, these counties were home to about 6 million people who live in about 2.2 million housing units.

2.1 Email Alerts

Visitors to trulia.com can set alerts that trigger regular emails when houses with certain characteristics come on the market. The web form for setting alerts is shown in Figure 1. Every alert must specify the fields in the first line: "Type" is either "For sale," "For rent," or "Recently sold." The field "Location" allows for a list of zip codes, neighborhoods, or cities. Neighborhoods are geographic units commonly listed on realtor maps that are often aligned with zip codes. When users fill out the form, an autocomplete function suggests names of neighborhoods or cities.

The second row in the form provides the option of specifying house characteristics beyond geography. Price ranges may be set by providing a lower bound, an upper bound or both. For bedrooms, bathrooms and house size, there is the option to set a lower or an upper bound. In the third row, "Property type" allows narrowing the search to "Single family home," "Condo" and several smaller categories. The remaining fields govern how emails are processed: for the "New listing email alerts" relevant to our study, the options are "Email me daily" or "Email me weekly."[2]

Figure 1: Setting Email Alerts on Trulia.com



Pooling email alerts to obtain search ranges

We observe a random subset of 40,525 "For sale" search alerts set between March 2006 and April 2012. Those alerts were set by 23,597 unique trulia.com users, identified by the (scrambled) email address to which emails triggered by the alert are sent. Almost 70 percent of searchers set only one alert, and more than 90 percent of searchers set three or fewer alerts. Since we are interested in search ranges rather than individual alerts, we pool alerts set by the same searcher, as described in the Appendix.

[2]Trulia also provides a second way for potential home buyers to set an email alert. After looking at results from normal searches on their website, generally along the same dimensions as shown in Figure 1, users can press a single button: "Send me an email whenever houses with these characteristics come on the market."

Representativeness

The interpretation of our findings depends to some extent on whether home searchers who use trulia.com are representative of the overall pool of searchers. In particular, their use of the internet in home search might signal that they are younger and richer than the average home buyer. While we do not have direct demographic information on the searchers in our sample, recent surveys conducted by the National Association of Realtors provide some useful background information on modern home search.

The internet has now become the most important tool in the home buying process, with over 90 percent of home buyers using the internet in their home search process (National Association of Realtors, 2013). In particular, for 35 percent of home buyers, looking online is the first step taken in the home purchase process. The fraction of people who deemed real estate websites "very important" as a source of information was 76 percent, larger than the 68 percent who found real estate agents "very important". Moreover, use of the internet is not concentrated among younger buyers: 86 percent of home buyers between the ages of 45 and 65 go online to search for a home. The median age of home buyers using the internet is 42, the median income is $83,700 (National Association of Realtors, 2011). This is only slightly younger than the median of all home buyers (which is 45), and slightly wealthier (the median income of all home buyers was $80,900).

In addition to showing that online real estate search is almost universal, this suggests that we can learn from online real estate search about overall housing search behavior. Moreover, trulia.com, with approximately 24 million unique monthly visitors during our sample period (71 percent of whom report to plan to purchase in the next 6 months), has similar demographics to those of the overall online home search audience (Trulia, 2013).

Major dimensions of search

All email alerts specify at least a geographic dimension of the potential home buyer's search range. Roughly a third of the queries do not specify any fields in addition to geography. The other fields that are specified regularly include listing price and the number of bathrooms. Table 1 shows the distribution of our sample of search alerts across whether these dimensions are specified. Just under a third of queries specify both price and the number of bathrooms, while another third specify just a price range. The remaining 5 percent of queries specify just a bathroom criterion in addition to the geographic restriction. Other fields in Figure 1 are used much less. For example, only 1.3 percent of queries specify square footage while 2.7 percent of queries specify the number of bedrooms. While the latter two fields are alternative measures of size, the minimum number of bathrooms is the most commonly used filter to place restrictions on the size of homes.

In Section 2.2, we establish stylized facts on the geographic breadth of housing search. In Appendix A.1 we discuss the key characteristics of housing search along the price and size dimensions; we also provide information about how the size and price dimension correlate with geographic breadth of the search range. For example, we document that searchers that are more

Table 1: Distribution of Email Alert Parameters

	Price not specified	Price specified	Total
Baths not specified	13,019	13,777	26,796
Baths specified	1,848	11,881	13,729
Total	14,867	25,658	40,525

Note: Table shows the distribution of parameters that trulia.com users specify in addition to geography.

specific about the price range and home size cover larger geographic areas.

2.2 Search by Geography

Each email alert defines the desired search geography by selecting one or more city, zip code or neighborhood. About 61 percent of alerts define the finest geographic unit in terms of cities, 18 percent in terms of zip codes, and the remaining 21 percent of alerts in terms of neighborhoods. Some queries include geographies in terms of cities, zip codes, and neighborhoods in the same query. In order to compare search queries that specify geography at different levels of aggregation, we translate every query into the set of zip codes that are approximately covered by that query, as discussed in Appendix A.2. The search queries cover a total of 191 unique Bay Area zip codes.

Distance

To summarize how search ranges reflect geographic considerations, we construct various measures of size of the area considered. 26 percent of searchers consider only a single zip code. For the remaining searchers, we measure the average and maximal geographic distance and travel time between all zip codes contained in their search ranges. We focus on distances between population-weighted zip code centroids. Population weighting is useful, since we are interested in the distance between agglomerations within zip codes that might reflect searchers' commutes.

Table 2 reports these measures of the geographic breadth of housing search. Geographic distance is measured in miles, and is direct "as the crow flies." We also report the maximum and average travel times by car or public transport between the population-weighted zip code centroids. Travel times are calculated using Google Maps, and are measured as of 8am on Wednesday, March 20, 2013.[3] The size of the typical search range is consistent with reasonable commuting times guiding geographic selections. For example, the median search range includes zip codes with a maximum travel time by car of about 20.5 minutes, and a maximum geographic distance of 6.8 miles. Moreover, there is sizable heterogeneity in the geographic breadth of housing search.

[3] A few zip code centroids are inaccessible by public transport as calculated by Google. Public transport distances to those zip code centroids were replaced by the 99th percentile of travel times between all zip code centroids for which this was computable. This captures that these zip codes are not well connected to the public transport network.

Table 2: Distribution of Distances Across Zip Codes

	Population-Weighted Zip Code Centroids					
	Min	Bottom Decile	Median	Top Decile	Max	Mean
Max Geographic Distance	0.5	2.3	6.8	21.1	103.3	9.7
Mean Geographic Distance	0.5	1.8	3.2	8.9	74.0	4.7
Max Car Travel Time	4.0	9.5	20.5	38.5	143.5	22.8
Mean Car Travel Time	3.8	8.9	13.1	19.7	132.5	14.0
Max Public Transport Time	10.5	40.5	79.0	375.0	573.5	140.1
Mean Public Transport Time	9.3	27.3	48.0	120.0	375.0	169.9

Note: Table shows summary statistics of geographic distance and travel time between the population-weighted centroids of all zip codes selected by a searcher. We focus on searchers who select more than one zip code. Travel times are measured in minutes, distances are measured in miles.

In Appendix A.3 we show that the geographic patterns of housing search are constant across the years in our sample, as well as across the seasonality of the housing market, suggesting that they truly represent time-invariant measures of household preferences.

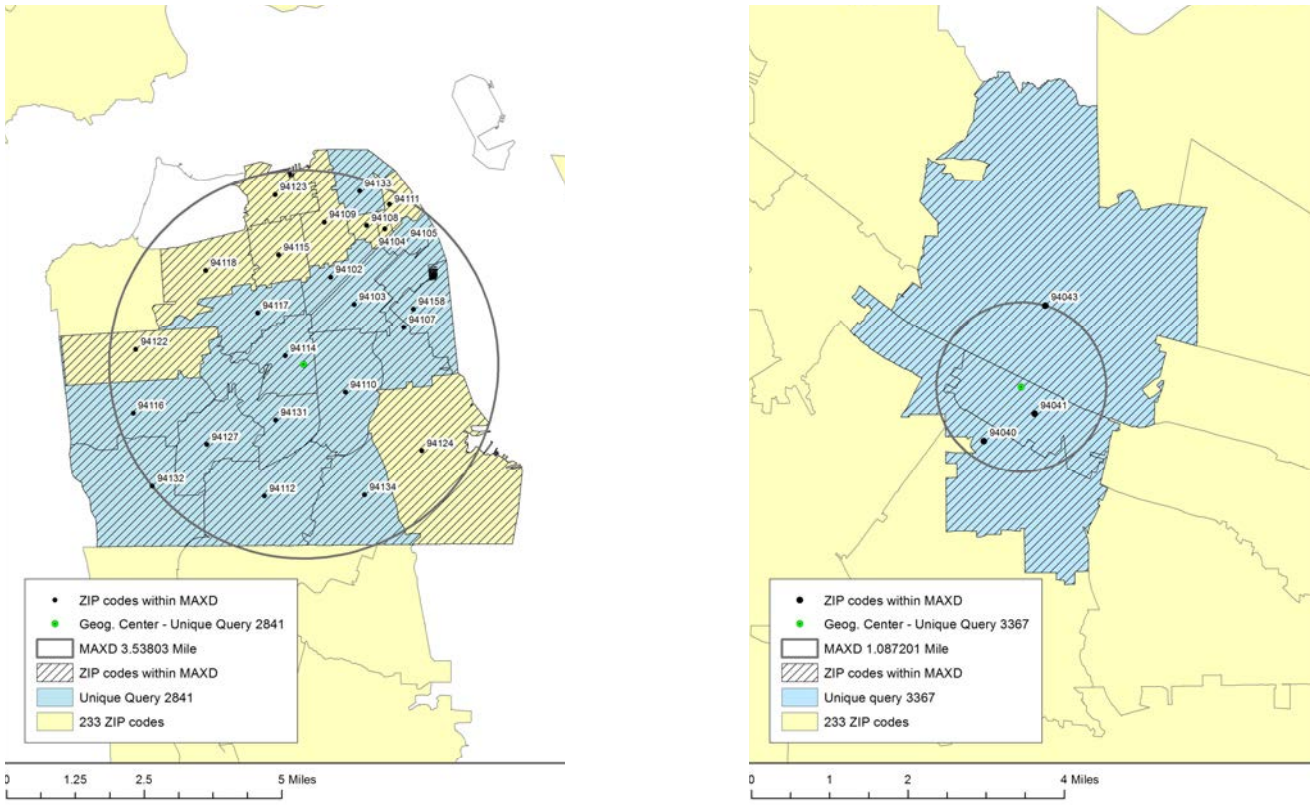
Contiguity & Circularity

To guide our modeling of clientele heterogeneity, we analyze whether there is a simple organizing principle for observed search ranges, namely that searchers consider contiguous areas, possibly centered around a focal point such as a place of work or a school. We say a search range is contiguous if it is possible to drive between any two zip code centroids in the range without ever leaving the range. Here we allow for travel across one of the six Bay Area bridges. Details on the construction of our measure of contiguity, full summary statistics, and examples of contiguous and non-contiguous searches are provided in Appendix A.4. About 18 percent of searchers have non-contiguous search ranges. These tend to be broad searchers, who consider more than five distinct zip codes and hence provide market integration across neighborhood and city boundaries.

A stylized model of geographic search might view a search range as being circular around a central point such as a job or a school. We ask whether the observed search ranges can be suitably approximated by such a model. To do this, we compute, for each searcher, the geographic center of the search range as the average longitude and latitude of all zip code centroids selected by that searcher. We then determine the maximum distance to this center of any zip code centroid contained in the search range. On average, the maximum distance is 3.95 miles, while the 10th percentile is 1.31 miles and the 90th percentile is 12.78 miles. We next compute the number of zip code centroids (not necessarily contained in the search range) that are within the maximum distance to the center. We say a search range is circular if all zip codes within maximum distance to the center are also contained in the search range. Figure 2 illustrates this procedure.

Overall, 47 percent of all searchers that cover more than one zip code have circular search ranges. This number is highest, at 83 percent, for ranges that only cover two zip codes, and

Figure 2: Explanation of Circularity Test



Note: Figure provides examples of the circularity measure. All zip codes that are part of the search set are shown in blue. The geographic center of each search set is given in green. The circle is centered around this geographic center and has radius equal to the furthest distance of any zip code centroid in the search set. All zip codes whose center lies within the circle (and who are thus at least as close as the furthest zip code center in the search set) are shaded. The left panel shows a non-circular search set, the right panel a circular search set.

declines for queries that cover more zip codes. In addition, for search sets with a larger maximum distance, the proportion of searches that cover all zip codes within this maximum distance from the center declines. On average, searchers cover 78 percent of all zip codes within maximum distance of their search range center. For non-contiguous ranges, the share of zip codes covered falls to 33 percent.

We conclude from these results that it is difficult to arrive at a parsimonious description of the geographic preferences defined by the search ranges. In particular, a modeling approach that describes search ranges in terms of contiguous and/or circular subsets of the plane will fail to account for the behavior of broad searchers who integrate markets. This finding guides our approach in the next section, where we define a discrete grid of market segments, using zip codes as the geographic units. Geographic selection can then be represented as subsets of the set of all zip codes. This approach allows us to accommodate non-contiguous and non-circular search patterns.

3 Housing Market Segments

In this section, we divide the San Francisco Bay Area into a finite number of housing market segments, motivated by the search ranges inferred from email alerts. We then establish stylized facts on market and search activity at the segment level.

3.1 Data

To measure housing market activity, we combine three main datasets. We start with the universe of ownership-changing deeds in the Bay Area between 1994 and 2011. The property to which a deed relates is uniquely identified by the Assessor Parcel Number (APN). From the deeds data, we obtain the property address, transaction date, transaction price, type of deed (e.g. Intra-Family Transfer Deed, Warranty Deed), and the type of property (e.g. Apartment, Single-Family Residence). We identify armslength transactions using information on the type of deed (see Appendix A.6).

We combine the transaction deeds with the universe of tax assessment records in the Bay Area for the year 2009. Properties are again identified by their APN. This dataset includes information on property characteristics such as construction year, owner-occupancy status, lot size, building size, and the number of bedrooms and bathrooms.

Finally, we use a dataset of all property listings on trulia.com between October 2005 and December 2011. The variables we use here are listing date, listing price, and the listing address. The latter can be used to match listings data to deeds data. We can then construct a measure of time on market for each property that eventually sells, as well as the average inventory that is for sale in a market segment at each point in time.

Throughout, we pool observations for the period 2008-2011, a time period for which we observe information on both housing search and housing market activity. The goal of this paper is to understand the cross section of market activity. Pooling observations across years helps us achieve a finer description of cross sectional heterogeneity. In particular, there are sufficiently many observations to measure market activity in segments with few listings and low housing turnover rates. To make prices comparable across years, we convert all prices to 2010 dollars using zip code level repeat sales price indices.

3.2 Defining Segments

The next step is to partition the Bay Area housing stock into different market segments. The finest partition that can be motivated by search data is obtained by the join of all search ranges in our sample. Any division of houses into segments would then characterize the preferences of at least one searcher. Moreover, the preferences of any one searcher could be expressed exactly as a subset of the set of all segments. However, the problem with this approach is sample size: the number of houses per segment would be too small to accurately measure moments such as time on the market and inventory.

Our approach, therefore, is to get as close as possible to the finest partition, but subject to the constraint that segments must be sufficiently large in terms of volume and housing stock. This leads us to a set H of 564 segments and a set Θ of 11,503 search ranges that can each be represented as a subset of H. These segments contain houses within a zip code that are of similar quality (based on price) and size (based on bathrooms). We provide a detailed description of the algorithm for constructing these segments in Appendix A.5. In what follows we only sketch the main steps.

We start from our earlier finding that people search mostly according to (i) quality, by specifying price ranges, (ii) geography, where zip code is generally the finest unit, and (iii) size, by specifying the number of bathrooms. Facts (ii) and (iii) lead us to first divide the Bay Area by zip code, and to then divide each zip code into two size categories, either "up to 2 bathrooms" or "more than 2 bathrooms."

To accommodate search by quality, we further divide – zip code by zip code – each size group into four price groups. Here we start from a set of candidate price cutoffs: \$200K, \$300K, \$400K, \$500K, \$750K and \$1 million. We then select the three cutoffs from these candidates that are most often close to price cutoffs appearing in our email alerts. The idea is that the resulting set of segments is close to the ideal partition implied by the ranges. In particular, high price zip codes will typically have higher cutoffs than lower price zip codes.

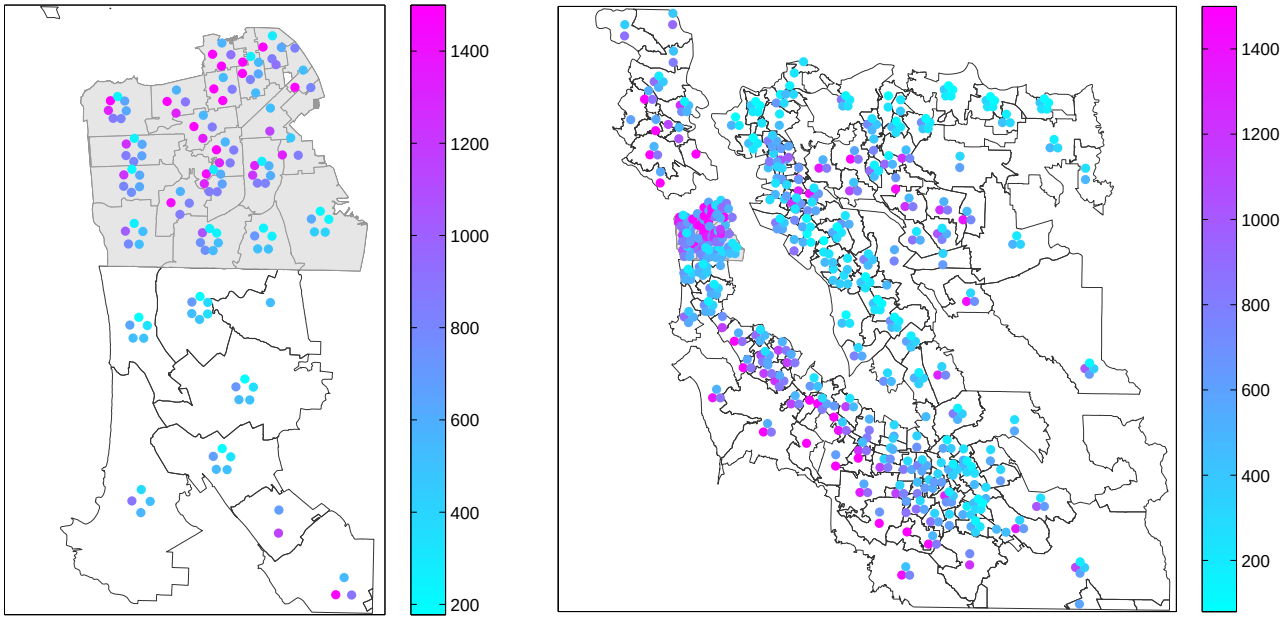
At this point, we have divided each zip code into eight size-price groups. It is possible, however, that some of these groups are too small to provide accurate measures of segment level moments. Our criteria here are that a segment must have enough transactions as well as a sufficiently large housing stock. If this is not the case, we merge candidate segments to form a larger joint segment, as described in the appendix. As a result, some zip codes that have very thin housing markets will have fewer than eight segments. This procedure splits the Bay Area into 564 segments, ranging in housing stock between 1,211 and 13,167, with a median housing stock of 3,298 (see Table 3).

The left panel of Figure 3 shows a map of the city of San Francisco in gray in addition to the area south of the city. The black lines delineate zip codes. The white areas without boundaries are water. Within each zip code, there are up to eight dots that represent segments. The segments are aligned clockwise starting with the cheapest segment at the top of the clock. The colors of the dots are the average house price in the segment, with the Dollar amounts in thousands indicated by the colored legend. The map illustrates the substantial heterogeneity of house prices in a city like San Francisco. There is also important heterogeneity within zip codes: the variance of log prices across zip codes captures only 60 percent of the variance at the (more disaggregated) segment level.

The map in the right panel of Figure 3 shows the entire Bay Area. The busy agglomeration of segments in the upper-left quadrant of the Bay Area map is San Francisco. The mostly light blue segments in the lower-right quadrant are the cheaper city of San Jose. The pink segments between these two cities are Silicon Valley cities like Palo Alto and Atherton, while the light blue

segments across the water are Oakland and other cheaper East Bay cities. The light blue dots in the upper-right corner belong to the Sacramento Delta.

Figure 3: Maps with segments in San Francisco and the Bay Area



Note: The left plot shows a map of downtown San Francisco as shaded area and areas south of downtown. The right plot shows a map of the entire Bay Area. The color bar indicates the price of the segment in thousands of Dollars.

Given the final set of segments, we express each search range as a subset of these segments. Here we start from the raw search range, specified along the dimensions quality, size and geography, ignoring other dimensions (which are rarely specified). We then determine the set of segments that is (approximately) covered by the specified range, using the algorithm described in Appendix A.2. There are 11,503 distinct search patterns.

3.3 Market Activity and Search Activity

We next present segment-level facts about market and search activity. These facts reveal a number of interesting patterns that motivate the subsequent quantitative exercise. The following notation is useful to organize facts reported at the segment level. Let H denote the set of all segments. The measure μ^H counts houses, so $\mu^H(h)$ is the housing stock in segment h. Let $V(h)$ denote the average monthly turnover rate in segment h, defined as the number of armslength transactions divided by total housing stock. Let $T(h)$ denote the mean time on the market in segment h, defined as months between listing and sales date, less one month for the typical escrow period. Our measure of average inventory in segment h is $\mu^S(h) := T(h) V(h) \mu^H(h)$.[4] We also define the

[4]This measure of inventory for houses conditions on houses that are eventually sold, since time on market T is based on actual sales. Alternatively, one can construct measures of inventory directly from listings data. The resulting series are noisy because they require assumptions on when the few listings that do not sell are removed.

inventory share $I(h) = \mu^S(h)/\mu^H(h)$, which is the fraction of the housing stock that is currently for sale.

Every search range in our sample is a subset of the set of all segments H. We index the ranges by $\theta \in \Theta$ and refer to the set Θ as the set of searcher "types."[5] A searcher of type θ scans inventory in the set of segments $\tilde{H}(\theta) \subset H$. The total housing stock that is of interest to searcher θ is

$$\nu^H(\theta) = \sum_{h \in \tilde{H}(\theta)} \mu^H(h).$$

Similarly, we define the total inventory considered by searcher θ, denoted $\nu^S(\theta)$, as the sum over all inventory μ^S for sale in segments in θ's search range $\tilde{H}(\theta)$.

The *clientele* of segment h consists of all searchers who consider segment h as part of their search range, that is,

$$\tilde{\Theta}(h) = \left\{\theta \in \Theta : h \in \tilde{H}(\theta)\right\}. \tag{1}$$

The pattern of clienteles reflects the interconnectedness of segments. As an extreme example, in a perfectly segmented market, there are $\#H$ types with search ranges each consisting of a single segment, and each segment has a homogeneous clientele of one type who searches only that segment, that is $\tilde{H}(\theta) = \{h\}$. In contrast, in a perfectly integrated market, there is a single type with $\tilde{H}(\theta) = H$, and all clienteles are identical and contain only that type. More generally, clienteles are heterogeneous and may consist of distinct types with only partially overlapping search ranges. Let $\beta(\theta)$ denote the relative frequency of search range θ in the data sample. The distribution of searchers interested in segment h follows by integrating the distribution $\beta(\theta)$ over $\tilde{\Theta}(h)$.

As one summary statistic of overall search activity in segment h, we compute the weighted number of searchers per house

$$\sigma(h) = \sum_{\theta \in \tilde{\Theta}(h)} \frac{\beta(\theta)}{\nu^H(\theta)}. \tag{2}$$

Weighting here captures the idea that search effort is somewhat diluted if it is broader. Indeed, if every searcher was looking only at one segment, then $\sigma(h)$ simply reflects the number of searchers per house in h. More generally, some searchers θ in the clientele of h may consider segments other than h. Dividing the number $\beta(\theta)$ of type θ by the housing stock $\nu^H(\theta)$ that this type is interested in makes broader searchers (who are interested in more housing stock) count less towards search activity in h.

So far, all summary statistics have been defined at the segment level only. We are also interested in how market and search activity vary at different levels of aggregation. Since V, I and σ are all

[5] For the presentation of facts in this section, "type" is no more than a label for search ranges. The notation is motivated by our model below, where each search range will indeed correspond to a different type of agent (with the search range a feature of preferences).

defined as ratios relative to housing stock, aggregation uses housing stock as weights. For example, the turnover rate over some subset $G \subset H$, such as a zip code or city, is computed as

$$\sum_{h \in G} \frac{\mu^H(h) V(h)}{\sum_{h \in G} \mu^H(h)}.$$

In addition to administrative geographic units, we are also interested in aggregating to sets of segments that are more closely integrated, in the sense that there is is a sufficiently large common clientele. We define the area connected to h as the set of segments $\tilde{h}$ such that the weighted share of searchers scanning both h and $\tilde{h}$ is at least a fraction ϕ of searchers scanning h,

$$A_\phi(h) = \left\{ \tilde{h} \in H : \sum_{\theta: h, \tilde{h} \in \tilde{H}(\theta)} \frac{\beta(\theta)}{\nu^H(\theta)} \geq \phi \sum_{\theta: h \in \tilde{H}(\theta)} \frac{\beta(\theta)}{\nu^H(\theta)} \right\}. \tag{3}$$

The distribution of market and search activity

Table 3 presents summary statistics on market and search activity for the Bay Area as a whole, as well as by segment. The housing market is illiquid: on average only 1.1 percent of Bay Area housing stock is for sale at any point in time, and the average monthly turnover rate is 0.24 percent, so that the typical house turns over once every $100/(0.24 \times 12) = 35$ years. The cross-sectional variation in market activity at the segment level is substantial. For example, the upper quartile inventory share is 1.5 percent, over twice as much as the 0.6 percent inventory share in the lower quartile. Houses in the upper quartile volume segments turn over more than twice as fast as at the lower quartile segments.

Table 3: Summary Statistics of Market and Search Activity

	Inventory Share I (in percent)	Turnover Rate V (in percent)	Search Activity σ	Mean Price (in thous.)	Housing Stock
Bay Area	1.14	0.24	1.00	649	2,216,021
Min	0.00	0.00	0.05	72	1,221
Q25	0.61	0.16	0.51	306	2,333
Q50	0.94	0.21	0.81	518	3,298
Q75	1.51	0.29	1.31	790	4,858
Max	6.66	0.97	7.05	2,491	13,167

Note: Table shows summary statistics for market and search activity, both for the entire Bay Area as well as across the 564 housing market segments. We consider inventory share, volume share, search activity σ (see equation 2 for the definition), mean price and total housing stock.

The measure of search activity in equation (2) has an average of one by construction. Its distribution is positively skewed: the majority of segments have less than one weighted searcher per house. The minimum of 0.05 is achieved by a segment in Martinez in the Sacramento Delta,

which is primarily considered by a few very broad searchers. In contrast, some segments have substantially more search activity, all the way to a maximum of 7.05 in a segment in central San Francisco, which attracts many narrow searchers in addition to broad searchers.

Variation across submarkets

Table 4 reports cross sectional variation in observables at different levels of aggregation. There are 564 segments, 191 zip codes and 96 cities in our data. The three left-hand panels show volatilities and correlation coefficients *across* segments, zip codes and cities, respectively. Comparison of volatilities shows that there is substantial variation across segments which is below the zip code and city level. Indeed, the zip code-level movements account for only 46, 44, and 64 percent of the segment-level variance in inventory share I, turnover rate V, and search activity σ, respectively.

Our market activity indicators – inventory share and turnover rate – comove strongly across any type of "submarket": segment, zip code, or city. Both variables also tend to be higher in cheaper submarkets. In contrast, the comovement of search activity and market activity depends crucially on the level of aggregation. While it is close to zero at the segment level, it turns negative at the zip code, and even more negative at the city level. At the same time, the relationship with price also changes: while more expensive segments do not see higher search activity on average, more expensive zip codes and cities are searched more.

Variation within submarkets

The three right hand panels of Table 4 consider segment-level variation within submarkets. The bottom two panels report volatilities and correlations within the average zip code and city. There are 191 zip codes and 96 cities in our data. All moments are weighted using housing stock.[6] The top right panel shows segment-level variation within areas connected by significant common clienteles $A_\phi(h)$, evaluating (3) with $\phi = 0.3$. There are as many such areas as there are segments.[7]

For market activity indicators and prices, the nature of covariation across and within submarkets is essentially the same: inventory share and turnover rate move together and are both negatively correlated with price. For zip code and city, the within-correlation coefficients in the right hand panels are also quantitatively close to the across-correlation coefficients in the left hand panels.

In contrast, the sign of the comovement between search activity on the one hand and market activity and prices on the other depends crucially on whether we look within or across submarkets. Indeed, for both zip codes and cities, search activity moves together with market activity and against price across units, but it moves against market activity and with price within units. The signs of within correlations are the same for connected areas that are defined on the basis of common clienteles as opposed to geographic closeness.

[6] The unweighted median number of segments for both zip codes and cities is equal to 3. However, the distribution of number of segments for cities is skewed. For example, San Francisco and San Jose contain 79 segments each.

[7] The median and 75th percentile connected area by number of segments consist of 5 and 10 segments, respectively. The main qualitative message below is not sensitive to the choice ϕ.

Table 4: Cross-Sectional Variation in Market and Search Activity

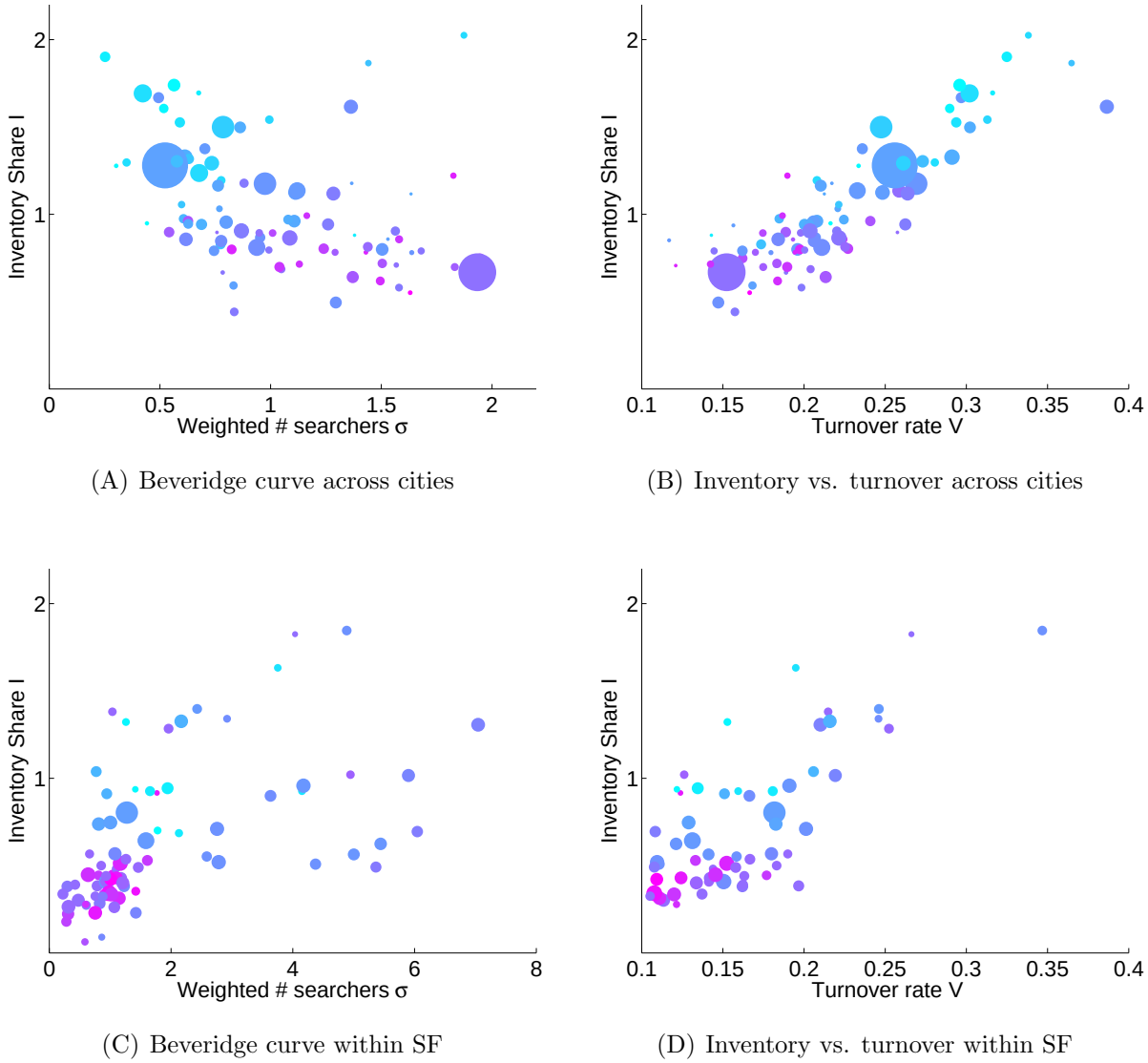
VARIATION ACROSS UNITS					AVG. VARIATION WITHIN UNITS				
	I	V	σ	$\log(p)$		I	V	σ	$\log(p)$
SEGMENT					CONN. AREA				
vol	0.74	0.12	0.81	0.64	vol	0.56	0.09	0.75	0.33
corr	1	0.93	0.03	-0.63	corr	1	0.87	0.34	-0.47
		1	0.01	-0.48			1	0.27	-0.36
			1	0.08				1	-0.18
				1					1
ZIP CODE					ZIP CODE				
vol	0.50	0.08	0.65	0.50	vol	0.60	0.09	0.55	0.41
corr	1	0.91	-0.18	-0.67	corr	1	0.83	0.58	-0.76
		1	-0.15	-0.52			1	0.41	-0.55
			1	0.41				1	-0.54
				1					1
CITY					CITY				
vol	0.42	0.07	0.48	0.46	vol	0.54	0.09	0.40	0.42
corr	1	0.93	-0.42	-0.74	corr	1	0.88	0.51	-0.67
		1	-0.34	-0.59			1	0.42	-0.52
			1	0.53				1	-0.51
				1					1

Note: Table shows cross-sectional variation in market and search activity at different levels of aggregation: 564 segments, 191 zip codes and 96 cities. The left column presents statistics across units, the right column presents statistics across segments within units. We consider inventory share, volume share, search activity σ (see equation 2 for definition) and log(price). We present both volatilities (standard deviations) as well as within and across unit correlations. Connected areas are defined as in equation (3), with $\phi = 0.3$.

The relationship between inventory and measures of search activity is reminiscent of the "Beveridge curve" that relates vacancies and unemployment in labor market statistics. The stylized fact here is that the housing market Beveridge curve is *downward sloping* across broad units of aggregation, while it is on average *upward sloping* within broad units. Panel A in Figure 4 shows this relationship across cities, and Panel C across all segments within the city of San Francisco. The Beveridge curve is upward sloping within 64 out of the 74 cities with at least two segments, representing 84% of the total Bay Area housing stock. One notable exception is Oakland, which has a correlation coefficient of -0.05. The fact is not primarily driven by small cities, however. The within-city Beveridge curve slopes up for 23 of the largest 25 cities by housing stock, with an average correlation of 0.44 and a 25th percentile correlation of 0.23. The correlation for San Francisco is 0.61. Negative correlations among the top 25 cities are close to zero, at -0.05 and -0.04.

Panel B in Figure 4 shows the positive correlation between inventory and turnover across cities. Panel D shows the positive correlation between inventory and turnover within San Francisco.

Figure 4: Inventory, Search and Volume



(A) Beveridge curve across cities

(B) Inventory vs. turnover across cities

(C) Beveridge curve within SF

(D) Inventory vs. turnover within SF

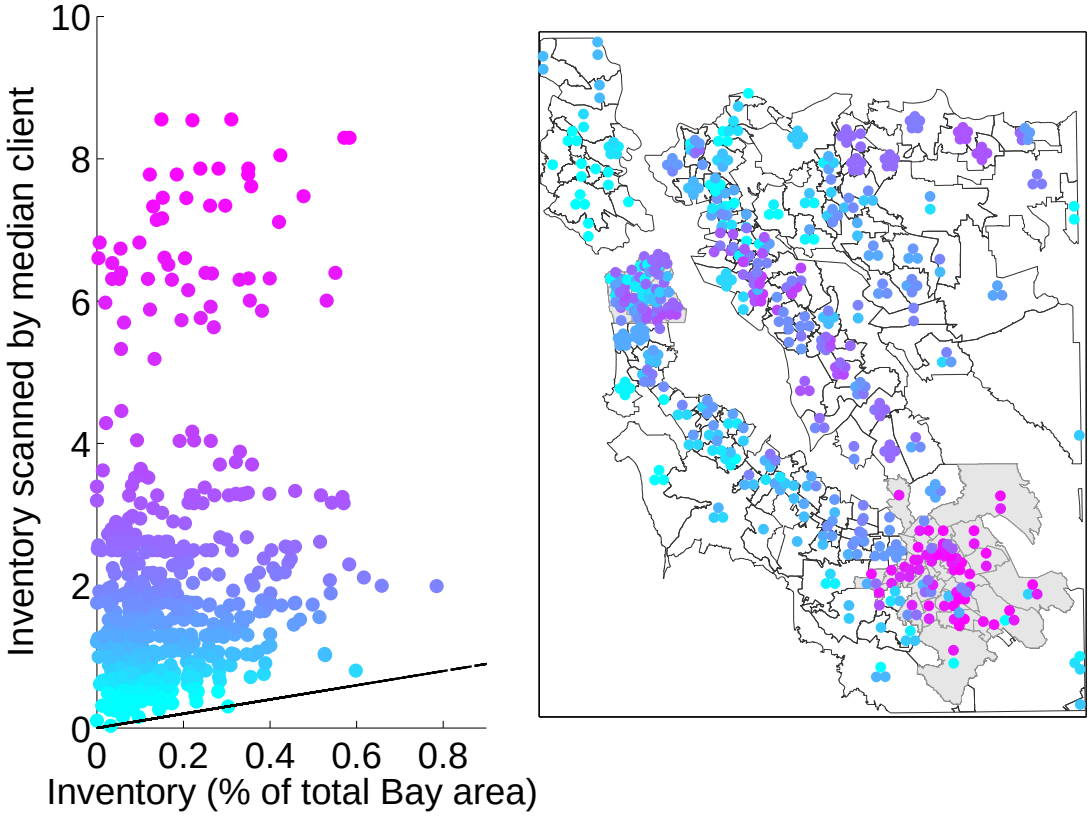
Note: Figure shows inventory I, weighted number of searchers σ, and turnover rate V across and within cities. Panel A shows inventory shares and weighted number of searchers (the Beveridge curve) across cities. Panel B shows inventory share and turnover rate across cities. Panel C shows the Beveridge curve within the city of San Francisco. Panel D shows inventory share and turnover rate within San Francisco. The size of the dots reflects the size of the housing stock (from small to large). The color of the dots reflects housing prices from cheap (blue) to expensive (pink).

Breadth of search & integration

The summary measure $\sigma(h)$ reflects the average search activity in a segment, but it does not tell us whether that activity is due to narrow local searchers or due to broader searchers who provide connection to other segments. To summarize interconnectedness, we now compare segments in terms of the inventory scanned by their typical client. The left-hand panel of Figure 5 plots the share of inventory in segment h in total Bay Area inventory (measured along the horizontal axis) against the share of Bay Area inventory scanned by the median client of segment

h (measured along the vertical axis). Every dot represents a segment, and color reflects the value on the vertical axis so the segments can be recognized in the map in the right-hand panel.

Figure 5: Scanned Inventory



Note: Figure shows scanned inventory by median household searching in each segment. In both panels, each dot corresponds to one segment. In the left panel, the horizontal axis shows the inventory in that segment as a fraction of entire Bay Area inventory. The vertical axis shows the fraction of total Bay Area inventory scanned by the median searcher in that segment. The right panel shows the geographic distribution of these segments.

If the Bay Area were perfectly segmented, then any given segment would only have clients who scan that particular segment. As a result, all points would have to line up along the 45-degree line. At the opposite extreme, if the Bay Area were perfectly integrated, then every client of every segment would scan all houses, so all points should line up along a horizontal line at 100 percent of total inventory. Not surprisingly, the truth is in the middle: the median searcher in a segment scans multiple times more inventory than is available in the segment itself, but far less than 100 percent of the Bay Area total.

Areas with a large common clientele appear in the plot as near-horizontal clusters: if any subset of segments were perfectly integrated but not connected to other segments, then it would form a horizontal line at the level of its aggregate inventory. This effect is visible for the top cluster of pink dots. The map in the right hand panel shows that those dots represent cheaper segments in the city of San Jose. More generally, clusters of dots with high scanned inventory correspond to cheap urban areas, where broad search appears to be more common.

The first column of Table 5 summarizes the distribution of inventory scanned by the median searcher. In the average segment, the median searcher scans 2.1 percent of the total inventory in that segment, or about 47 houses. The table also clarifies that most dots in Figure 5 are clustered in the bottom left; the upper quartile is at only at 2.5 percent of total inventory. The second column in Table 5 shows the distribution of within segment interquartile ranges for scanned inventory. There is substantial within-segment heterogeneity in the clientele's breadth of housing search. Indeed, the average within-segment IQ range of inventory scanned by different searchers is, at 1.75 percent, larger than the across-segment IQ range of inventory scanned by the median searcher. Interestingly, clientele heterogeneity comoves strongly with overall connectedness: the correlation coefficient between the first and second columns is 65 percent. In other words, in segments that are, on average, more integrated with other segments, there are larger within-segment differences between the interacting narrower and broader searchers.

Table 5: Variation in Scanned Inventory by Clienteles

	Segment: total inv. scanned (in percent)		Zip code: share of search ranges (in percent)				City: share of search types (in percent)			
	median	IQ range	one	many	subset	other	one	many	subset	other
Mean	2.10	1.75	5.9	4.4	5.8	84.5	10.7	8.1	16.8	64.4
Q25	0.94	1.04	0	0	0.8	72.3	0	0	3.1	53.2
Q50	1.55	1.56	1.2	1.2	2.4	93.0	8.0	1.7	12.4	66.6
Q75	2.53	2.17	6.5	4.7	6.5	97.1	17.0	6.3	25.4	77.2

Note: Table provides information on the inventory scanned by clienteles at different levels of aggregation. The second column measures the inventory scanned by the median searcher in a segment. The third column is the interquartile range of inventory scanned across all searchers in a segment. The other columns report mean and quartiles for the share of different searcher categories across segments. For each of the two geographies, zip code and city, a searcher is a category "one" searcher if the reason he searches the segment is because he uniquely selected that geography. The searcher is a category "many" searcher, if he only selected on that geography, but included more than one unit. A "subset" searcher covers the segment, but only selected a subset of one zip code or city to be included. "Other" searchers cover subsets of multiple zip codes or cities.

Search at the city and zip code level

The right-hand columns of Table 5 demonstrate the importance of detailed segment-level information for understanding search patterns. For each zip code and city, we consider all searchers who are active in that zip code or city. We then separate these searchers into how they searched in that geography. We first classify the share of searchers who scan exactly one zip code or city in its entirety (column "one"). The category "many" captures searchers that scan only based on the particular geography, but consider more than one zip code or city. Together, they indicate the share of searchers for whom detailed segment level information beyond geography is not important. The category labeled "subset" collects searchers who scan only a subset of the geography (i.e. they scan less than one zip code, or less than one city), for example because they select that zip code in addition to a price cutoff. The final category "other" collects searchers for whom seg-

ment information matters because their range intersects with multiple zip codes or cities, without covering them in their entirety.

The table reports mean and quartiles for the shares of each category of searcher in the cross section of segments. For example, in the average segment, only 5.9 percent of searchers select exactly the zip code containing that segment. An additional 4.4 percent of searchers specify their search query only in terms of zip codes, but specify more than just one zip code. The distribution is highly skewed: in 75 percent of segments, the share of searchers scanning exactly the zip code is 6.5 percent or less. The order of magnitude of the numbers is larger at the city level but still relatively small. We therefore conclude that the clientele patterns at work in our data are not simply driven by searches selecting a single zip code or city. Instead, other characteristics defining a segment, in particular size and quality, play an important role.

4 Model

We now introduce a search model with multiple segments to interpret the stylized facts presented above. In particular, we show how the interaction of heterogeneous clienteles helps to understand the slope of the Beveridge curve within and across cities. More generally, the model relates observed market outcomes (inventory, turnover and search) to the distribution of searcher preferences and the matching technology. We can then use it to perform counterfactuals as well as to assess the effect of search frictions and transactions costs on house values. We can also use our model to analyze the transmission of shocks to local housing markets. We do this in Section 6.

4.1 Setup

The model describes a small open economy, such as the San Francisco Bay Area. Time is continuous and the horizon is infinite. Agents live forever and discount the future using the risk-free rate r.

Segments, search ranges and clienteles

We assume that different household types search across different segments of the housing market, as in our data. We use the notation introduced in Section 3.3. There is a finite set H of market segments. The measure μ^H on H counts the number of houses in each segment. We normalize the total number of houses in the economy to one:

$$\sum_{h \in H} \mu^H(h) = 1.$$

Agents have quasilinear utility over two goods: numeraire and housing services. Agents own at most one house. When an agent moves into a house in segment h, he obtains housing services $v(h) > 0$, until the house falls out of favor, which happens at the rate $\eta(h)$. After the house falls out of favor, the agent no longer receives housing services from that particular house. The agent

can then put the house on the market in order to sell it and subsequently search for a new house. We assume that search for a new house is costless. While putting a house on the market is also costless, sellers incur a proportional cost c upon sale of a house.

Agent type θ is identified by a *search range*, a subset $\tilde{H}(\theta) \in H$ of market segments that he is interested in. Search ranges are part of the description of preferences, and type θ never moves into a house outside of $\tilde{H}(\theta)$. We use a measure μ^{Θ} on the set of all types Θ to count the number of agents of each type. The total number of agents is

$$\bar{\mu}^{\Theta} = \sum_{\theta \in \Theta} \mu^{\Theta}(\theta) > 1.$$

Since there are more agents than houses and agents own at most one house, some agents are always searching. These $\bar{\mu}^{\Theta} - 1$ agents rent or stay in a hotel while they search for a house to buy.

The *clientele* $\tilde{\Theta}(h)$ of segment h is the set of all agents who are interested in segment h, as defined in (1). It is helpful to consider two extremes. The market is perfectly segmented if every segment is searched by a single type who is interested only in that segment. The clienteles $\tilde{\Theta}(h)$ are then disjoint sets that each contain a single type θ. In contrast, the market is fully integrated if there is only one type who searches all segments; all clienteles $\tilde{\Theta}(h)$ are then identical and consist only of that type. The *inventory scanned* by type θ is $\nu^S(\theta)$.

Matching

Matching in the housing market involves searchers who scan inventory, identify suitable properties and make contact with sellers. We capture the entire process by a random matching technology. We make two assumptions. First, searchers flow into segments within their search range in proportion to segment inventory. This assumption is natural if searchers are equally likely to find a favorite house anywhere in their search range. Formally, let $\tilde{\mu}^B(\theta)$ denote the number of buyers of type θ. We define the number of buyers in segment h as

$$\mu^B(h) = \sum_{\theta \in \tilde{\Theta}(h)} \frac{\mu^S(h)}{\nu^S(\theta)} \tilde{\mu}^B(\theta). \tag{4}$$

For a given segment h, buyers can belong to any type in the clientele $\tilde{\Theta}(h)$. If a type θ searches only in segment h, then $\nu^S(\theta) = \mu^S(h)$ and all buyers $\tilde{\mu}^B(\theta)$ of type θ are in fact buyers in h. In particular, if the market is perfectly segmented, then the only buyers in h are searchers of the type who looks exclusively in h. More generally, the more inventory is available in h relatively to the total inventory in type θ's search range, the larger the share of type θ buyers who flow into h. If the market is perfectly integrated, there is only one type and the number of buyers is always proportional to inventory; put differently, the buyer-seller ratio is equated across segments.

Our second assumption is the presence of a matching function. The match rate in segment h

is given by

$$m(h) = \tilde{m}\left(\mu^B(h), \mu^S(h), h\right),$$

where $\tilde{m}$ is increasing in the number of buyers and sellers and satisfies $\tilde{m}\left(0, \mu^S, h\right) = \tilde{m}\left(\mu^B, 0, h\right) = 0$. At this point, we do not make further assumptions on the functional form of the function $\tilde{m}$. What is important is that it is allowed to depend on the segment h directly (other than through the number of buyers and inventory). For example, the process of scanning inventory could be faster in some segments because the properties there are more standardized, or because more open houses are available to view properties.

Once a buyer and seller have been matched, the seller makes a take-it-or-leave-it offer. If the buyer rejects the offer, the seller keeps the house and the buyer continues searching. If the buyer accepts the offer, the seller pays the offer price. The seller receives the offer price net of the proportional cost c which goes, say, to a real estate agent. The seller then starts to search, whereas the buyer moves into the house and begins to receive utility $v(h)$.

Equilibrium

In equilibrium, agents choose optimally given the distribution of others' choices. In particular, owners decide whether or not to put their houses on the market, sellers choose price offers and buyers choose whether or not to accept those offers. In what follows, we focus on steady state equilibria in which (i) owners put their house on the market if and only if their house has fallen out of favor, so that the owners do not receive housing services from it, and (ii) all offers are accepted.

4.2 Characterizing Equilibrium

We derive a system of equations that determines the steady state distribution of agent states (that is, searching for a house, listing one for sale, or owning without listing). Since there are fixed numbers of agents and houses, that distribution can be studied independently of prices and value functions. We need notation for the number of agents in each state. Let $\mu^H(h;\theta)$ denote the number of type θ agents who are homeowners in segment h, and let $\mu^S(h;\theta)$ denote the number of type θ agents whose house is listed in segment h. In steady state, all those numbers, as well as the numbers of buyers by type $\tilde{\mu}^B(\theta)$ and by segment $\mu^B(h)$, are constant.

The first set of equations uses the fact that $\mu^S(h)$, the number of houses for sale in segment h, is constant in steady state. As a result, the number of houses newly put on the market in segment h must equal the number of houses sold in segment h:

$$\eta(h)\left(\mu^H(h) - \mu^S(h)\right) = \tilde{m}\left(\mu^B(h), \mu^S(h), h\right). \tag{5}$$

The left-hand side shows houses coming on the market, given by the rate at which houses fall out of favor multiplied by the number of houses that are not already on the market. The right-hand

side shows the number of matches and thus the number of houses sold.

The second set of equations uses the fact that the rate at which houses fall out of favor in segment h is the same for all types in the clientele of h. As a result, the share of houses owned by type θ agents in h must equal the share of houses bought by type θ agents in h:

$$\frac{\mu^H(h;\theta)}{\mu^H(h)} = \frac{\mu^S(h)}{\nu^S(\theta)} \frac{\tilde{\mu}^B(\theta)}{\mu^B(h)}. \tag{6}$$

On the right-hand side, the share of type θ buyers in segment h equals the number of type θ buyers that flow to h in proportion to inventory, as in (4), divided by the total number of buyers in segment h. The equation also says that the buyer-owner ratio for any given type θ in segment h is the same and equal to the segment level buyer-owner ratio $\mu^B(h)/\mu^H(h)$.

Finally, the number of agents and the number of houses must add up to their respective totals:

$$\begin{aligned} \mu^H(h) &= \sum_{\theta \in \tilde{\Theta}(h)} \mu^H(h;\theta), \\ \mu^\Theta(\theta) &= \tilde{\mu}^B(\theta) + \sum_{h \in \tilde{H}(\theta)} \mu^H(h;\theta). \end{aligned} \tag{7}$$

Equations (5), (6) and (7) jointly determine the unknown objects $\mu^H(h;\theta)$, $\mu^S(h)$, $\mu^B(h)$ and $\tilde{\mu}^B(\theta)$, a system of $2\#H + \#\Theta(1+\#H)$ equations in as many unknowns.

4.3 Parameters and Identification

The model distinguishes three forces that determine market activity and prices in the cross section. Two forces operate at the segment level. First, the rate $\eta(h)$ at which houses fall out of favor represents differences in the supply of houses across segments. In what follows, we refer to $\eta(h)$ as a measure of *instability*: a more unstable segment is one where more houses come on the market per period. The second force is the segment-specific effect on match rates summarized by $\tilde{m}(.,.,h)$, which represents differences in market frictions across segments. The third force is the demand for housing, which is captured by the distribution of search ranges $\tilde{H}(\theta)$ and the number of agents $\mu^\Theta(\theta)$ of type θ.

Housing demand parameters are the most complex, since their effect depends on the entire clientele pattern. It is nevertheless helpful to consider a summary measure at the segment level. We define the *popularity* of a segment by

$$\pi(h) = \sum_{\theta \in \tilde{\Theta}(h)} \frac{\mu^\Theta(\theta)}{\nu^H(\theta)}. \tag{8}$$

A segment is more popular if there are more agents per house who include it in their search ranges. The weighting of agents is the same as for our measure of search activity $\sigma(h)$ defined

in (2): broad searchers who are interested in multiple segments other than h (and thus a larger housing stock) count less towards the popularity of segment h than narrow searchers. A key difference between $\sigma(h)$ and $\pi(h)$ is that the latter is an exogenous determinant of demand for segment h; it depends only on the distribution of types $\mu^{\Theta}(\theta)$. In contrast, $\sigma(h)$ is determined endogenously; in particular, it depends on the equilibrium share of agents of each type that are searching at any point in time.

Observables

The observables described in Section 3.3 all have model counterparts. The *inventory share* is $I(h) = \mu^S(h)/\mu^H(h)$ and the *turnover rate* is $V(h) = m(h)/\mu^H(h)$. Search alerts represent a sample of buyers. The relative frequencies of the search ranges $\tilde{H}(\theta)$ in the model are given by $\beta(\theta) = \tilde{\mu}^B(\theta)/\left(\bar{\mu}^{\Theta} - 1\right)$, and are thus observable up to the constant $\bar{\mu}^{\Theta} - 1$. Our measure of search activity at the segment level can be written as

$$\sigma(h) = \frac{1}{\bar{\mu}^{\Theta} - 1} \sum_{\theta \in \tilde{\Theta}(h)} \tilde{I}(\theta) \frac{\tilde{\mu}^B(\theta)}{\nu^S(\theta)}, \tag{9}$$

where $\tilde{I}(\theta) = \nu^S(\theta)/\nu^H(\theta)$ is the inventory share measured over the search range of type θ.

Exact identification

The structure of the model implies that the supply and demand parameters — $\eta(h)$ and $\mu^{\Theta}(\theta)$, respectively — can be identified without taking a stand on the exact shape of the matching function. All that is required is that the dependence of the matching function on the segment is sufficiently flexible that the model can jointly match inventory, turnover and search behavior. Our results on the relative importance of heterogeneous clienteles for understanding the Beveridge curve as well as for determining price discounts will be independent of the exact matching function.

Our formal identification result, derived in Appendix A.7, says that, under the assumption stated above, the model implies a one-to-one mapping between two sets of numbers. The first set consists of the parameters $\eta(h)$ and $\mu^{\Theta}(\theta)$ as well as the vector of rates at which buyers find houses in a given segment, defined as $\alpha(h) = m(h)/\mu^B(h)$. The second set consists of the inventory share $I(h)$, the turnover rate $V(h)$, the relative frequencies of search ranges $\beta(\theta)$ and the *average* time it takes for a buyer to find a house.

To quantify the model, we can therefore use the second set of numbers as targets and derive the first set in closed form; this delivers the complete vector of supply parameters $\eta(h)$ and demand parameters $\mu^{\Theta}(\theta)$. Numbers for inventory, turnover and searcher frequencies come directly from the data reported in Section 3. We do not have information on the overall number of buyers $\bar{\mu}^{\Theta} - 1$. As an additional target moment, we set the average match rate for a buyer to 20 percent per month, the average match rate for inventory in our data. The average time it takes for a buyer to find a house is therefore about 5 months. This choice does not affect the relative behavior of

market and search activity across segments, and therefore is not particularly important for most of our results.

4.4 Search with heterogeneous clienteles

Before reporting numerical estimation results, we illustrate the main economic mechanisms using example markets with stylized clientele patterns. There are two broad themes. First, local conditions in a segment – for example, its popularity – matter relatively more if the segment is less integrated with other segments. Second, the interaction of broad and narrow searchers generates patterns for observables – in particular, an upward sloping Beveridge curve – in ways that cannot occur in either perfectly segmented or perfectly integrated markets.

Perfect segmentation

To illustrate the role of local conditions, we start with the extreme case of perfect segmentation. Suppose there are exactly as many types as segments and each type scans exactly one segment. We use the label $\theta = h$ for the type scanning segment h and otherwise drop θ arguments. The number of buyers $\mu^B(h)$ is the difference between the number of types interested in h and the number of houses in segment h, $\mu^\Theta(h) - \mu^H(h)$. Substituting into equation (5), equilibrium inventories are determined segment by segment by

$$\eta(h)\left(\mu^H(h) - \mu^S(h)\right) = \tilde{m}\left(\mu^\Theta(h) - \mu^H(h), \mu^S(h), h\right). \tag{10}$$

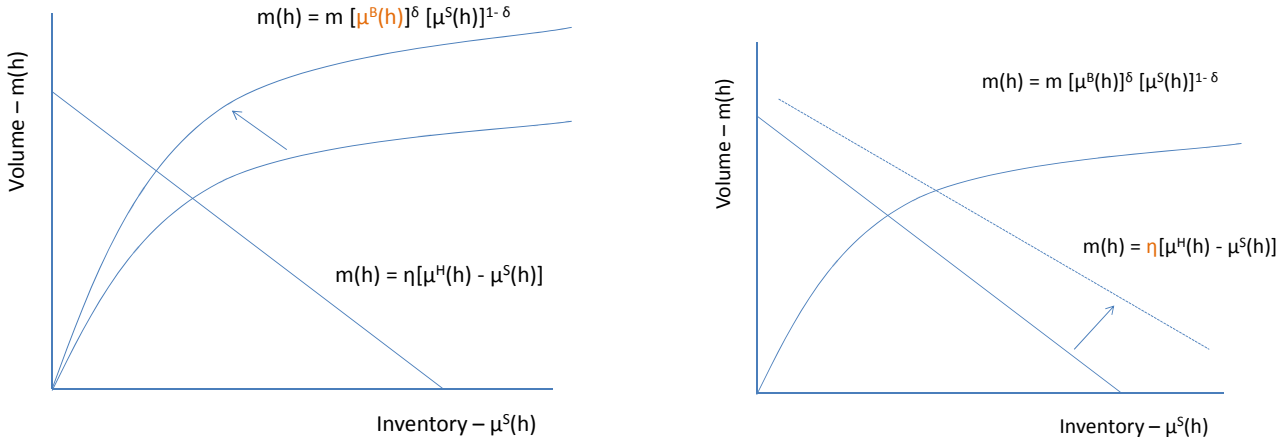
The equation determines equilibrium inventory $\mu^S(h)$ and volume $m(h)$ or equivalently the inventory share $I(h)$ and the turnover rate $V(h)$. [8]

Figure 6 plots both sides of the equilibrium condition (10) for segment h. The left-hand side is the rate at which houses come on the market in segment h. It is strictly decreasing in inventory: if inventory is higher, fewer agents live in their favorite house and fewer houses come on the market each instant. The right-hand side of (10) is the rate at which houses are sold. It is strictly increasing in inventory: if inventory is higher, matches occur at a faster rate. It follows that there is a unique equilibrium level of inventory $\mu^S(h)$ – if inventory is too low, then too many houses come on the market whereas if inventory is too high, then too many houses are sold.

Comparative statics show how the cross section of segments depends on local conditions. In the left panel of Figure 6, we consider a more popular segment by increasing $\mu^\Theta(h)$ and thus $\pi(h) = \mu^\Theta(h)/\mu^H(h)$ as defined in (8). There are now more buyers per house; the upward-sloping curve shifts up. The increased demand for houses implies higher equilibrium volume (and turnover rate) but lower equilibrium inventory (and inventory share). From (10), a segment with a matching technology that allows more matched per period for given buyer and seller pools similarly has higher volume and lower inventory. Finally, in the right hand panel of Figure 6, we consider

[8] Indeed, after dividing by the housing stock $\mu^H(h)$ in the segment, we can view both sides as representing the turnover rate as a function of inventory share, with $\mu^H(h)$ as a parameter.

Figure 6: Equilibrium and Comparative Statics with Perfect Segmentation



Note: The plots show the curves on the left and right hand side of the equilibrium condition (10) together with comparative statics exercises. In the left panel, the exercise increases the popularity of the segment by increasing $\mu^{\Theta}(h)$ and thus the number of buyers $\mu^{B}(h)$. The left panel increases its instability $\eta(h)$.

a more unstable segment by increasing $\eta(h)$. Houses now come on the market more quickly in h; the downward-sloping curve shifts to the right. The increased supply of houses implies that equilibrium volume and inventory are both higher.

Suppose we want to explain the data in Table 4 with a perfectly segmented model. At any level of aggregation, we observe a strong positive relationship between the inventory share and the turnover rate. It follows that differences in supply – the parameter $\eta(h)$ – must be important, whereas differences in popularity or the matching technology must be weak enough so as not to overturn the positive relationship between inventory share and turnover rate. At the same time, equation (9) says that search activity in the perfectly segmented case simply reflects the relative number of buyers scanning segment h

$$\sigma(h) = \frac{\mu^{\Theta}(h) - \mu^{H}(h)}{(\bar{\mu}^{\Theta} - 1)\,\mu^{H}(h)} = \frac{\pi(h) - 1}{\bar{\mu}^{\Theta} - 1}. \tag{11}$$

Variation in search activity across segments is thus driven only by variation in popularity; instability or the matching technology do not matter. Since more popular segments have lower inventory shares, such variation in itself will always generate a downward sloping Beveridge curve. To generate instead an upward sloping Beveridge curve with a perfectly segmented model would thus require comovement of $\pi(h)$ and $\eta(h)$: if more unstable segments are also more popular, then high inventory and turnover driven by ample supply could in principle go along with more search activity driven by high demand.

Partial integration and the role of broad searchers

We now extend the example by adding one additional type: a "broad searcher" who scans

all segments $h \in H$. We denote this type by 0 so $\nu^S(0)$ is total inventory. The new buyer pool of segment h thus contains both narrow searchers of type h and broad searchers who flow into segment h depending on the share of segment h inventory in total inventory (by the definition of buyers (4)). Equilibrium inventories again adjust to equate the flow of houses coming on the market to the volume of sales:

$$\eta(h)\left(\mu^H(h) - \mu^S(h)\right) = \tilde{m}\left(\tilde{\mu}^B(h) + \frac{\mu^S(h)}{\nu^S(0)}\tilde{\mu}^B(0), \mu^S(h), h\right). \tag{12}$$

The key new feature with partial integration is that the buyer pool is endogenous and tends to move positively with inventory in equilibrium. Two effects are relevant here. The first is that a larger share of broad type 0 searchers flows to segments with higher inventory. This effect comes from integration only; it is present also with perfect integration. Indeed, if $\tilde{\mu}^B(h) = 0$, then buyers are proportional to inventory; in equilibrium all markets are equally tight. The second effect comes from the interaction of broad and narrow searchers: competition from broad searchers implies that the number of narrow searchers looking for a house in h also increases with inventory. To see this, use the implication of (6) that buyer-owner ratios in any given segment are equated across all types in the clientele. Comparing narrow and broad buyers and owners in the clientele of h, we have

$$\frac{\tilde{\mu}^B(h)}{\mu^\Theta(h) - \tilde{\mu}^B(h)} = \frac{\mu^S(h)}{\nu^S(0)} \frac{\tilde{\mu}^B(0)}{\left[\mu^H(h) - \left(\mu^\Theta(h) - \tilde{\mu}^B(h)\right)\right]}, \tag{13}$$

where the second denominator on the right hand side determines broad owners in h as a residual. Holding fixed the number of houses and narrow types, a segment with a higher share of inventory must have more narrow buyers.[9]

How could a model with partial integration account for the facts in Table 4? For the cross section of inventory shares and turnover rates, the role of parameters is qualitatively similar to the perfect segmentation case. Indeed, in more unstable segments, inventory and volume both tend to be higher. However, the effect on inventory is typically weaker because more searchers flow into segment h as more houses come on the market there. In other words, greater supply endogenously gives rise to an offsetting increase in demand. An increase in the speed of matching or a larger number of types interested in h will increase turnover and decrease inventory. Of course, the interdependence of segments implies that the magnitude of effects is now more complicated to assess. For example, how the relative inventory of two segments depends on their stability now depends on the stability of other segments as well.

An important difference to the perfectly segmented case is that the presence of broad searchers

[9]The remaining equations determining equilibrium is the definition of $\nu^S(0)$ as total inventory and the requirement that buyers add up to the correct total, that is,

$$\bar{\mu}^\Theta - 1 = \tilde{\mu}^B(0) + \sum_{h \in H} \tilde{\mu}^B(h).$$

alters the mapping between parameters and search activity. With partial integration, search activity (2) becomes

$$\sigma(h) = \frac{1}{\bar{\mu}^{\Theta} - 1}\left(\frac{\tilde{\mu}^{B}(h)}{\mu^{H}(h)} + \tilde{\mu}^{B}(0)\right). \tag{14}$$

Differences in search activity across segments are driven only by differences in narrow buyers per house, since the contribution of broad searchers is the same for all segments. It then follows from (12)-(13) that for two equally popular segments (which have identical $\mu^{\Theta}(h)$ and $\mu^{H}(h)$), the segment with more instability or slower matching must have higher inventory together with higher search activity. In other words, with partial integration, differences in stability or the speed of matching can account for an upward sloping Beveridge curve.

5 Quantitative Results

Table 6 summarizes properties of the estimated demand and supply parameters. We report moments of instability $\eta(h)$ as well as popularity $\pi(h)$, our segment-level summary statistic of demand. Panel A provides information on the distribution of the parameters across segments. Panel B reports correlations both among the parameters themselves and between parameters and observables. Here we compare variables at the segment level, city-level averages as well as variation across segments within the city of San Francisco. Table 6 also reports the properties of the inferred match rate $\alpha(h)$. While $\alpha(h)$ is an endogenous object rather than a parameter, it contains information about the role of matching frictions.[10] Since we do not have information to identify the shape of the matching function, we do not directly draw conclusions about the matching technology. We only record what can be inferred about the distribution of match rates from search behavior.

5.1 Instability and popularity at the segment and city level

Instability $\eta(h)$ tracks turnover almost exactly: its moments in Table 6 are essentially the same as those reported for the turnover rate in Table 3. The intuition follows from the equilibrium condition (5) together with the summary statistics in Table 3. Indeed, inventory shares are so small – their 75th percentile is at 1.5% – that $\eta(h) \approx V$. Intuitively, because the time a house remains on the market is much shorter than the time that it is occupied by an owner, turnover is almost entirely accounted for by the frequency of moving shocks. We thus infer that unstable segments are cheaper and have more turnover and larger inventories. This is true not only across segments, but also across cities and within San Francisco.

Popularity at the segment level, $\pi(h)$, ranges between 0.2 and 2.4, with an IQ range between 0.80 and 1.15. The fact that popularity is below one for many segments is indicative of the

[10] For example, with a Cobb-Douglas matching function $m(h) = \bar{m}(h)\,\mu^{B}(h)^{\delta}\,\mu^{S}(h)^{1-\delta}$, we would have $\log \bar{m}(h) = \delta \log \alpha(h) + (1-\delta)\log\left(V(h)/I(h)\right)$.

Table 6: Quantitative Results from Estimating the Model

Panel A: Estimated parameters vs. hypothetical segmentation

	estimated parameters			hypothetical perfectly segmented benchmark	
	$100 \times \eta(h)$	$\pi(h)$	$\alpha(h)$	$\hat{\sigma}(h) - \sigma(h)$	$\hat{\pi}(h) - \pi(h)$
Mean	0.24	1.00	0.19	0	0.02
Q25	0.16	0.80	0.07	−0.17	−0.13
Q50	0.21	1.00	0.13	−0.04	0.03
Q75	0.29	1.15	0.23	0.12	0.21

Panel B: Correlation between estimated parameters

	$\eta(h)$	$\pi(h)$	$\alpha(h)$	$I(h)$	$V(h)$	$\sigma(h)$	$\log(p(h))$
Across segments							
$\eta(h)$	1	−0.01	0.24	0.93	1.00	0.13	−0.48
$\pi(h)$		1	−0.53	−0.03	−0.01	0.73	0.07
$\alpha(h)$			1	0.27	0.23	−0.43	−0.22
$\hat{\sigma}(h) - \sigma(h)$	0.44	0.10	−0.04	0.46	0.44	0.28	−0.16
Across cities							
$\eta(h)$	1	−0.23	0.52	0.93	1.00	−0.24	−0.55
$\pi(h)$		1	−0.62	−0.27	−0.23	0.69	0.36
$\alpha(h)$			1	0.57	0.52	−0.70	−0.65
$\hat{\sigma}(h) - \sigma(h)$	0.37	−0.01	−0.00	0.39	0.37	0.09	−0.11
Within San Francisco							
$\eta(h)$	1	0.16	−0.01	0.89	1.00	0.26	−0.35
$\pi(h)$		1	−0.59	0.18	0.16	0.86	−0.04
$\alpha(h)$			1	−0.17	−0.01	−0.57	0.07
$\hat{\sigma}(h) - \sigma(h)$	0.67	0.16	−0.12	0.67	0.67	0.32	−0.14

Note: The estimated parameters from the model exactly match three moments from the data: the inventory share $I(h)$, the turnover rate $V(h)$ and the relative frequency of search ranges $\beta(\theta)$. The total number of buyers $\bar{\mu}^{\Theta} - 1$ is set to match a 5 month average buyer search time.

importance of partial integration. Indeed, if segments were either perfectly segmented or perfectly integrated, then the number of weighted buyers would be larger than the number of houses in all segments, and popularity would have to be above one.[11] Popularity also comoves strongly with search activity at all levels of aggregation.

Does the correlation between segment characteristics contribute to the slope of the Beveridge

[11] To see how partial integration can imply $\pi(h) < 1$, consider a simple example: assume there are two equally large and equally unstable segments 1 and 2, say, as well as an equal number of narrow searchers who scan only segment 1 and broad searchers who scan both segments. We then have $\pi(1) = 3/2$ and $\pi(2) = 1/2$. Intuitively, a segment that is considered largely by broad searchers will tend to have low popularity.

curve? From our discussion in Section 4, this would require that the correlation between popularity and instability takes on different signs at different levels of aggregation. In fact, this correlation is weakly positive within San Francisco, zero across segments and negative across cities. We conclude that the Beveridge curve slopes down across cities in part because more popular cities are also less stable. Similarly, the Beveridge curve slopes up within San Francisco in part because more popular segments within the city are less stable. However, the within-city correlation is only 0.15 which already suggests that other forces are at work as well. We discuss those in the next section.

5.2 The role of partial integration: a perfectly segmented benchmark

How does the interaction of heterogeneous clienteles contribute to the shape of the Beveridge curve? What goes wrong if a modeler ignores that interaction? We now provide quantitative answers to both questions by comparing the actual economy – for which the estimated parameters reflect partial integration – with a hypothetical perfectly segmented benchmark economy. For the benchmark economy, we change demand parameters so as to remove integration, but hold all other parameters fixed. At the same time, we require that the benchmark economy still matches the same observed inventory shares and turnover rates as the actual economy. This pins down a unique vector of hypothetical demand parameters $\hat{\mu}^{\Theta}$. From the equilibrium condition (5) and the fact that the matching function remains unchanged, the benchmark economy also predicts the same number of buyers and buyer match rates $\alpha(h)$ as the actual economy. We can therefore view the benchmark economy as a model estimated by an econometrician who observes $I(h)$ and $V(h)$ as well as buyer match rates by segment (or, equivalently, who knows the segment-specific matching functions), but who does not have information on integration and proceeds by assuming that the economy is perfectly segmented.

The key source of error for an econometrician who observes inventory, turnover and buyer match rates (or matching functions) is that he may misjudge the relative popularity of segments. Indeed, the equilibrium condition (5) implies that the parameter $\eta(h)$ closely tracks turnover regardless of the clientele pattern. However, the benchmark economy can differ from the actual economy in the demand parameters $\mu^{\Theta}(h)$ and will typically imply a different cross section of the popularity parameters, denoted by $\hat{\pi}(h)$. We use the distribution of the difference $\hat{\pi}(h) - \pi(h)$ to assess the error made by an econometrician who ignores the possibility of partial integration. Moreover, the benchmark economy may predict a different cross section of search activity, denoted by $\hat{\sigma}(h)$, and hence a different Beveridge curve. We use the difference in search activity $\hat{\sigma}(h) - \sigma(h)$ to assess the contribution of partial integration to the Beveridge curve: if integration were not important, then the actual and hypothetical Beveridge curves would coincide.

An analytical example

We illustrate properties of the benchmark economy using the example from section Section 4.4, for which we can derive the Beveridge curve in closed form. Indeed, if there are only narrow

searcher types who look at one segment as well as one broad searcher type who looks at all segments, the type distribution for the benchmark economy is given by $\hat{\mu}^{\Theta}(0) = 0$ and $\hat{\mu}^{\Theta}(h) = \mu^{H}(h) + \mu^{B}(h)$. Combining equations (11) and (14), search activity in the benchmark economy is

$$\hat{\sigma}(h) = \sigma(h) + \beta(0)\left(\frac{I(h)}{\tilde{I}(0)} - 1\right), \tag{15}$$

where $\tilde{I}(0)$ is the average inventory share in broad searchers' range (that is, the economy as a whole). The actual Beveridge curve consists of the locus $(\sigma(h), I(h))$. The hypothetical benchmark Beveridge curve coincides with the actual one only if the second term is zero. This could be either because there are no broad searchers or because all inventory shares are the same, so that broad searchers flow equally into all segments and their effect is neutral.

More generally, the benchmark economy contains more searchers per house than the actual economy in a segment if and only if that segment has an inventory share that is larger than the economy average. This is because both economies must generate the same equilibrium inventory and number of buyers via different mechanisms. In the actual partially integrated economy, the endogenous flow of broad searchers toward inventory endogenously generates more buyers in high inventory segments. This mechanism is absent from the perfectly segmented benchmark economy, where a segment has more buyers if and only if it is more popular. Since search activity is driven in turn by the number of narrow searchers, high (low) inventory segments must also see more (less) search activity. It follows that in the (I, σ) plane, the benchmark Beveridge curve is flatter than the actual Beveridge curve. Moreover, the correlation between inventory and the difference in search activity $\hat{\sigma}(h) - \sigma(h)$ measures the impact of partial integration: if it is zero, then the perfectly segmented benchmark also explains the data.

Quantitative importance of partial integration

Table 6 reports key moments of the differences in search activity as well as popularity between the benchmark and actual economies. There are two main results here. First, the error incurred by an econometrician who ignores partial integration is large. This follows from the quantiles for the difference in popularities $\hat{\pi}(h) - \pi(h)$ reported in the rightmost column in panel A. Indeed, the interquartile range for this difference is of the same order of magnitude as that for the estimated parameter $\pi(h)$ itself, reported on the top left. The reason is that the range of popularities is much narrower for the benchmark economy: the 25th percentile for $\hat{\pi}(h)$ is at 1.01, while the 75th percentile is at 1.03.

The intuition here is again the absence of an endogenous flow of broad searchers. In order to explain the data on market activity, the perfectly segmented benchmark economy must assume small differences in narrow searchers, rather than broad searchers who flow towards inventory in unpopular segments. An econometrician who postulates a perfectly segmented model in the absence of data on clientele patterns will mistakenly infer that unpopular segments have a sizable

clientele who look for housing only there. He will miss the fact that buyers in those segments are broad searchers who are happy to buy elsewhere. The difference matters also for considering the response to shocks, which we consider in Section 6.2 below.

The second result is that partial integration is an important force that works towards an upward-sloping Beveridge curve. This follows from the lines in panel B that report cross-sectional correlations between the difference in search activity $\hat{\sigma}(h) - \sigma(h)$ and inventory at different levels of aggregation. From the expression for search activity (15), we would expect this correlation to be higher in integrated areas (that is, shutting down integration flattens the Beveridge curve for such areas). In fact, the correlation coefficient within San Francisco is 0.67. A one standard-deviation increase in inventory thus implies that $\hat{\sigma}(h) - \sigma(h)$ is higher by 0.67 standard deviations; this corresponds to 0.45 standard deviations of the search activity measure $\sigma(h)$ itself. Put differently, the absence of broad searchers flattens the Beveridge curve within San Francisco by adding about half a standard deviation worth of search activity per unit of inventory. Across cities or segments, the effect is smaller because there is less integration and the effects average across areas. For example, across cities, the absence of broad searchers adds only about 10 percent worth of search activity per unit of inventory.

What exogenous forces drive broad searchers towards high inventory segments? In principle, differences in instability, popularity or matching frictions could all generate difference in inventory. In fact, all forces are unconditionally correlated with inventory share. The flow of broad searchers, however, is directed to a large extent by differences in instability alone. Indeed, the correlation of $\hat{\sigma}(h) - \sigma(h)$ with $\eta(h)$ is much larger than that with $\pi(h)$ or $\alpha(h)$. We can therefore sum up the key mechanism as follows: in more unstable segments, more houses come on the market; if they are part of an integrated area, then broad searchers are attracted to those houses and crowd out narrow searchers, generating an upward sloping Beveridge curve.

6 Prices and spillovers

6.1 Equilibrium prices & liquidity discounts

Our model captures two distinct housing market frictions. The first is search: owners whose house falls out of favor spend time first looking for a buyer and then for a new house; during this time they forgo the flow utility of living in their favorite house. The second friction is the transaction cost paid upon sale. In equilibrium, both costs are capitalized and reduce the house price: every buyer takes into account that both he and all potential future buyers may have to sell and hence search and pay transactions costs.

We now derive a convenient formula to show how the resulting *liquidity discount* reflects both frictions. Let $V^F(h;\theta)$ be the utility of a type θ agent who obtains housing services from a house in segment h. Since sellers make take-it-or-leave offers and observe buyers' types, they charge prices equal to buyers' continuation utility. The price paid by a type θ buyer in segment h is thus

$p(h,\theta) = V^F(h;\theta)$. We now show that prices are the same in all transactions in segment h. We start from the Bellman equation of a seller who puts his house on the market

$$rV^S(h;\theta) = \frac{m(h)}{\mu^S(h)}\left(E\left[p(h,\theta)(1-c)|h\right] - V^S(h;\theta)\right),$$

where the expectation uses the equilibrium distribution of buyers $\mu^B(h;\theta)/\mu^B(h)$. It follows that the value function of the seller is independent of type. Intuitively, a seller knows that once he becomes a buyer, his continuation value is zero. Seller utility thus derives only from the expected sale price, about which all seller types care equally.

Consider next the Bellman equations of an owner who does not put his house up for sale

$$rV^F(h;\theta) = v(h) + \eta(h)\left(V^S(h;\theta) - V^F(h;\theta)\right).$$

Since utility $v(h)$ and the arrival of moving shocks are also independent of type, so is $V^F(h;\theta)$. As a result, the same price $p(h)$ is paid in all transactions in segment h. We can combine these equations and determine the price from

$$p(h) = \frac{v(h)}{r} - \frac{\eta(h)}{r + m(h)/\mu^S(h) + \eta(h)}\left(\frac{v(h)}{r} + cp(h)\frac{m(h)}{\mu^S(h)}\right). \tag{16}$$

The first term on the right-hand side is the present value of a permanent flow of housing services. This frictionless price obtains if houses never fall out of favor ($\eta = 0$) or else if sales involve no transaction costs ($c = 0$) as well as no search costs. Search costs are zero if matching is infinitely fast ($m/\mu^S \to \infty$). Indeed, if a new buyer can be found instantaneously once a house is listed, then the price must also reflect the permanent flow of services. More generally, the house price reflects a liquidity discount – the second term – that capitalizes search and transaction costs. The liquidity discount is larger if houses fall out of favor more quickly (η higher) or if it is more difficult to sell a house in the sense that time on market $\mu^S(h)/m(h)$ is longer or the cost c is higher.

To decompose the liquidity discount, we use our knowledge of the relative magnitudes of parameters to derive an approximate formula. The key observation is that the fraction multiplying the big bracket in equation (16) is approximately equal to the inventory share I. Indeed, both the monthly interest rate r and the parameter $\eta \approx V(h)$ are small relative to $m(h)/\mu^s(h) = V(h)/I(h)$. We thus obtain a useful shortcut to interpret the numerical results below: solving out, we write the price (up to an approximation error of at most 15 basis points) as

$$p(h) \approx \frac{v(h)(1-I(h))}{r + cV(h)} = \frac{v(h)}{r}(1-I(h))\frac{r}{r+cV(h)}. \tag{17}$$

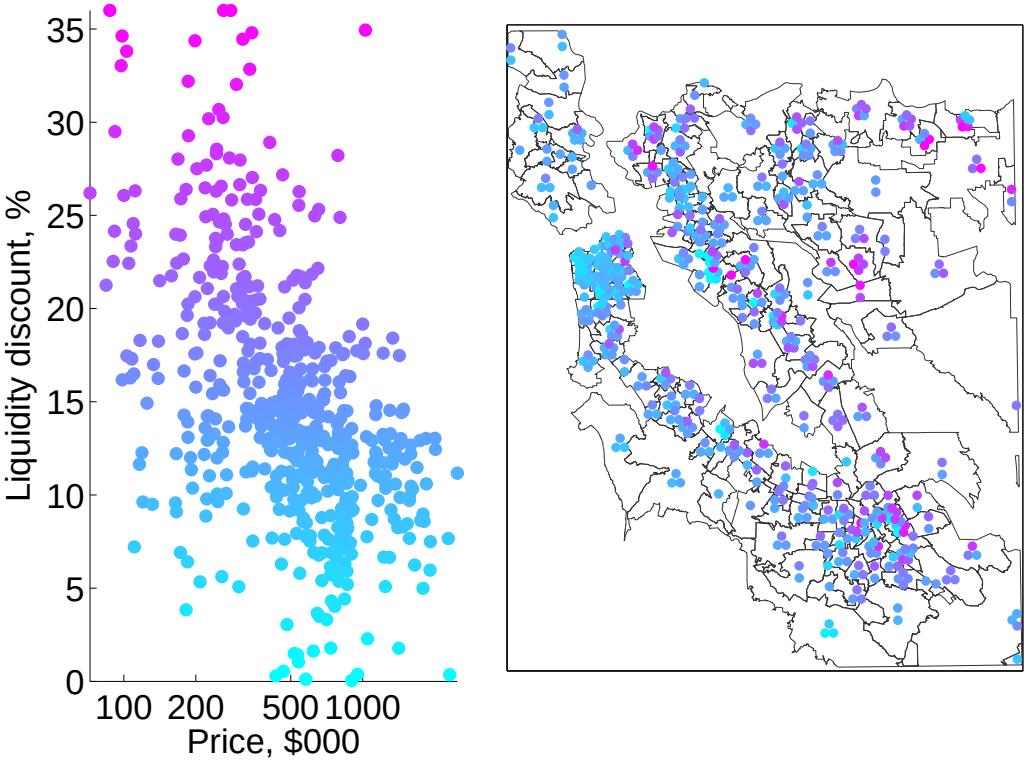
Frictions thus modify the frictionless price v/r by first reducing housing services proportionately by I and then increasing the discount rate to $r + cV$. The second expression distinguishes

two separate discounts. First, the inventory share measures the loss in value due to search frictions. It is zero in the absence of search frictions when there is no inventory as all houses are sold immediately. From Table 3, the size of the search discount is typically a few percentage points. The interest rate does not matter for its size (at least approximately) because time on market is fairly short. The second discount measures the present value of transaction costs: it is zero if there is no turnover or if selling houses is costless. Here the interest rate is important: if future transaction costs are discounted at a lower rate, then the discount is larger.

Liquidity discounts by segment

We now ask how market frictions identified by our estimation quantitatively affect the dispersion of house prices across segments. To compute the liquidity discount in the pricing equation (16), we need an estimate of the fundamental value. Our equilibrium computations imply values for matches $m(h)$ and houses for sale $\mu^S(h)$. Together with a real interest rate of 1% and a transaction cost equal to 6% of the resale value of the house, we can back out the utility values $v(h)$ such that the model exactly matches the cross section of transaction prices on the left-hand side of (16).

Figure 7: Liquidity Discounts



Note: Left panel: mean segment price versus segment liquidity discount in percent of frictionless price. Color coding reflects liquidity discount. Right panel: map with dots for each segment in same color as in left panel. Dots for segments within the same zip code arranged clockwise by price with lowest priced segment at noon.

Figure 7 shows the results. The left-hand panel plots transaction prices by segment against the liquidity discounts, stated as a percentage of frictionless price. The right-hand panel shows

the geographic distribution of the liquidity discounts. There are two notable results here. First, liquidity discounts are large. The median discount is 14 percent and 90th percentile is at 24 percent. From Table 3 and the approximating formula above, both search and transactions costs contribute to this result, but transactions costs are quantitatively more important. While search costs generate discounts up to 6 percent, the capitalized value of transaction costs is what leads to double digit discount numbers.

The second result is that liquidity discounts differ widely by segment, often within the same zip code. Table 4 shows that inventory and turnover exhibit about the same amount of variation within and across zip codes. The search and transaction costs inherit these properties, respectively. In poor segments with high volume and high inventory, both search and transaction costs are high. As a consequence, prices are significantly lower than they would be in a frictionless market. In rich segments discounts are still significant, but they are considerably smaller.

6.2 Comparative statics

In this section, we perform comparative statics exercises to show how clientele patterns matter for the transmission of shocks across segments. Motivated by recent debates on gentrification in San Francisco, we ask what happens when a neighborhood becomes more popular, that is, more searchers are specifically interested in that neighborhood. We compare two neighborhoods that are similar in size and price, but differ in clientele patterns.

The first neighborhood is zip code 94015 in Daly City, a suburb right outside of the San Francisco city limits. It contains about 11,000 houses with an average price of $480K. The average inventory is 105 houses; at our estimated parameters, this inventory is considered by 430 active searchers. We choose Daly City because narrow buyers are prevalent: depending on the segment within the zip code, between 12 and 35 percent of buyers were looking only in 94015, with the largest narrow buyer share in the most expensive segment.

The second neighborhood is the Outer Mission area of San Francisco, zip code 94112. For comparability with 94015, we select only the cheapest three segments in this zip code; we thereby obtain the same total housing stock and average price. However, the population of searchers is quite different: there are now 2000 searchers looking at an inventory of 84 houses. Moreover, the share of broad searchers is large: at our parameter values at most 3 percent of buyers are looking exclusively within zip code 94112.

We now recompute equilibrium under the assumption that 500 additional agents are interested either in 94015 or in the Outer Mission. In other words, we identify the two types in Θ whose search ranges consist of the five segments of 94015 or the three cheapest segment in 94112, respectively. We increase the corresponding entries in μ^{Θ} by 500 agents. This counterfactual captures long-run effects on inventory, turnover and prices if a neighborhood becomes more popular.

Greater popularity of a segment implies that inventory there declines: more buyers stand ready to snap up houses that go on sale. At the same time, turnover rates remain essentially unchanged

since houses come on the market at the same rates $\eta(h)$ as before. The time to sell a house thus moves proportionally with inventory. Moreover, price effects follow from equation (17): while changes in the transaction cost discount are negligible, changes in inventory directly move the search discount. The left-hand panels of Figure 8 show the magnitude of inventory share declines (and hence also declines in time on market and increases in price) for Daly City on the top and Outer Mission on the bottom. Both figure display only the San Francisco peninsula where both zip codes are located. The colors range from light blue for no effect to pink for the largest effect, which obtains in both cases within the zip codes that become more popular, shaded in gray.

Clientele patterns imply dramatically different behavior of inventory shares across the two local shocks. If 94015 becomes more popular, then the market reacts strongly in 94015 itself. The average inventory share decline within the zip code is 18 percent. Houses sell about 5 times faster, reducing time on market by about 1 month in the long run. Strong spillover effects are present to the adjacent zip code 94014, with some smaller effects transmitting to the towns to the south; here the change is around 1-2 percent of inventory. There is no notable effect on the city of San Francisco to the north, shaded in light gray.

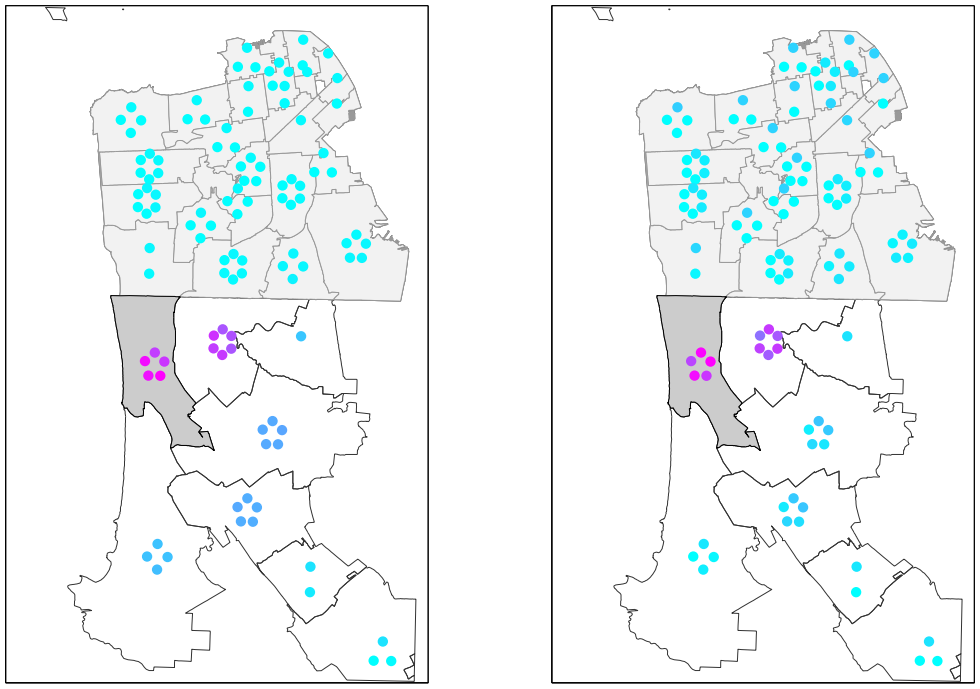
In contrast, if similar segments in the Outer Mission become more popular, then there are sizable spillover effects within the entire city of San Francisco. The average decline in inventory in the segments that have become more popular is only 8 percent. However, we have declines of 3 percent or more in all San Francisco neighborhoods. Since changes in inventory translate directly into search discounts, price responses are also relatively small and diluted if the shock hits Outer Mission, but large and concentrated in Daly City.

Data on clientele patterns are critical for the conclusions of this section. A perfectly segmented economy would not be able to distinguish the spillover effects of more versus less integrated neighborhoods. It would also provide misleading summary information about the demand for housing. For example, at our estimated parameters, mean popularity is 1.08 for 94015 and 0.82 for the three cheapest segments of 94112: since the Outer Mission has more broad searchers, it has a more elastic demand which leads to smaller responses to shocks. In contrast, in the perfectly segmented benchmark economy introduced in Section 5, mean popularities are 1.01 and 1.03, respectively. An econometrician using a perfectly segmented economy to assess the inflow of new searchers would therefore expect more competition for scarce housing in the Outer Mission compared to Daly City.

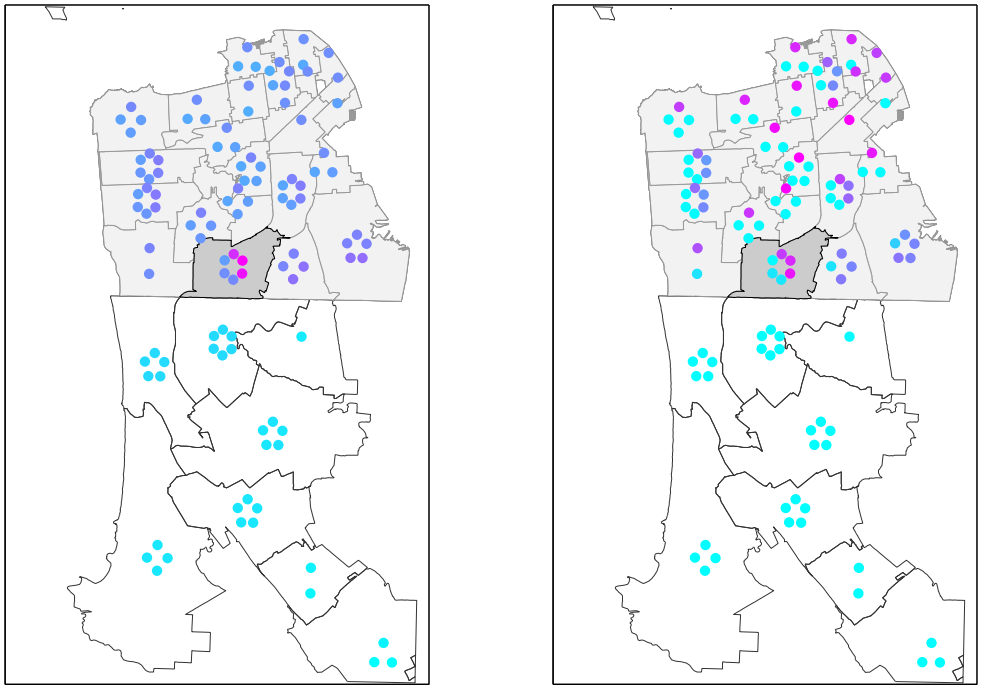
Spillovers along chains of integration

According to our model, the difference in spillovers is due to differences in clientele patterns. One immediate effect is that if a segment that becomes more popular is connected to a second segment by a broad searcher, then one would expect inventory to decline in both segments. Indeed, as more narrow searchers reduce inventory in the first segment, broad searchers turn to the second segment and reduce inventory there also. This need not be the end of the story, however: suppose

Figure 8: Inventory Response to an Increase in Popularity; Joint Searchers



(A) More searchers in Daly City



(B) More searchers in San Francisco Outer Mission

Note: Top left panel: percent change in inventory with 500 more searchers in zip code 94015 (Daly City). Top right panel: joint searcher share with 94015. Lower left panel: percent change in inventory with 500 more searchers in the three cheapest segments of zip code 94112 (San Francisco Outer Mission). Lower right panel: joint searcher share with 94112.

the second segment is connected to a third segment via another broad searcher, then inventory might decline also in that third segment even though it might not share any joint searchers with the first segment where the original shock occurred.

To explore the relevance of these effects, we compute a measure that isolates the first-round spillover effects. Starting from an initial segment, we compute, for each other segment, the share of its clientele that also searches in the initial segment. This *joint searcher share* is 100 percent if the segments are perfectly integrated and zero if they share no common searchers. The measure is plotted in the right-hand panels of Figure 8. For the case of Daly City, we find that the joint searcher share tightly matches the inventory response. In the neighboring segments where the inventory response is large, the joint searcher share is around 70 percent. Spillovers from a shock to Daly City are therefore largely transmitted through first-round effects.

The situation is very different for the case of San Francisco. Here the joint searcher share shows that the Outer Mission is directly connected only to the relatively cheaper segments of other San Francisco zip codes. Indeed, for a given zip code pink dots appear at noon and the first few dots in a clockwise directions. At the same time, the inventory response is large throughout the city. This is because richer neighborhoods are in turn connected to somewhat cheaper neighborhoods by common searchers so that higher-round spillover effects occur. Altogether the chain of connections generates strong spillovers throughout the city.

References

Albrecht, James, Axel Anderson, Eric Smith, and Susan Vroman, "Opportunistic matching in the Housing Market," *International Economic Review*, 2007, *48*, 641–664.

Anenberg, Elliot and Patrick Bayer, "Endogenous Sources of volatility in housing markets: the joint buyer-seller problem," *National Bureach of Economic Research*, 2014, *18980.*

Bayer, Patrick, Fernando Ferreira, and Robert McMillan, "A unified framework for measuring preferences for schools and neighborhoods," *Journal of Political Economy*, 2007, *115* (4), 588–638.

Burnside, Craig, Martin Eichenbaum, and Sergio Rebelo, "Understanding booms and busts in housing markets," *National Bureau of Economic Research*, 2014.

Davis, Morris and Stijn Van Nieuwerburgh, "Housing, finance and the macroeconomy," 2014.

Diaz, Antonia and Belen Perez, "House prices, sales, and time-on-market: a search-theoretic framework," *Journal of International Economics*, 2013, *53*, 837–872.

Genesove, David and Lu Han, "Search and matching in the housing market," *Journal of Urban Economics*, 2012, *72* (1), 31–45.

Goodman, Allen C and Thomas G Thibodeau, "Housing market segmentation," *Journal of housing economics*, 1998, *7* (2), 121–143.

Guren, Adam and Tim McQuade, "How do foreclosures exacerbate housing downturns?," 2014.

Halket, Jonathan and Matteo Pignatti, "Housing tenure choices with private information," 2014.

Han, Lu and Will Strange, "The microstructure of housing markets: search, bargaining, and brokerage," *Forthcoming Handbook of Regional and Urban Economics*, 2014, *5*.

Head, Allen, Huw Lloyd-Ellis, and Hongfei Sun, "Search and the dynamics of house prices and construction," *American Economic Review*, 2014, *104*, 1172–1210.

Hurst, Erik, Veronica Guerrieri, and Dan Hartley, "Endogenous gentrification and housing price dynamics," *Forthcoming Journal of Public Economics*, 2014.

Islam, Kazi Saiful and Yasushi Asami, "Housing market segmentation: A review," *Review of Urban & Regional Development Studies*, 2009, *21* (2-3), 93–109.

Krainer, John, "A theory of liquidity in residential real estate markets," *Journal of Urban Economics*, 2001, *49* (1), 32–53.

Kurlat, Pablo and Johannes Stroebel, "Testing for information asymmetries in real estate markets," 2014.

Landvoigt, Tim, Monika Piazzesi, and Martin Schneider, "The housing market(s) of San Diego," *Forthcoming American Economic Review*, 2014.

Leishman, Chris, "House building and product differentiation: An hedonic price approach," *Journal of Housing and the Built Environment*, 2001, *16* (2), 131–152.

Manning, Alan and Barbara Petrongolo, "How local are labor markets? Evidence from a spatial job search model," 2011.

Mian, Atif and Amir Sufi, "House prices, home equity-based borrowing, and the US household leverage crisis," *American Economic Review*, 2011, *101* (5), 2132–56.

National Association of Realtors, "2011 Profile of home buyers and sellers," Technical Report 2011.

_ , "Digital house hunt," Technical Report 2013.

Ngai, L Rachel and Silvana Tenreyro, "Hot and cold seasons in the housing market," *Forthcoming American Economic Review*, 2014.

Novy-Marx, Robert, "Hot and cold markets," *Real Estate Economics*, 2009, *37*, 1–22.

Piazzesi, Monika and Martin Schneider, "Momentum traders in the housing market: Survey evidence and a search model," *The American Economic Review*, 2009, *99* (2), 406–411.

Poterba, James M, David N Weil, and Robert Shiller, "House price dynamics: The role of tax policy and demography," *Brookings Papers on Economic Activity*, 1991, *1991* (2), 143–203.

Stein, Jeremy, "Prices and trading volume in the housing market: A model with down-payment effects," *Quarterly Journal of Ecomomics*, 1995, *110*, 379–406.

Stroebel, Johannes, "Asymmetric information about collateral values," *Forthcoming Journal of Finance*, 2014.

Trulia, "The largest audience of real estate consumers," Technical Report 2013.

Wheaton, William C, "Vacancy, search, and prices in a housing market matching model," *Journal of Political Economy*, 1990, pp. 1270–1292.

A Online Appendix – Not for Publication

A.1 Search by Price and Size

Out of the 63 percent of email alerts that specify a price criterion, 50 percent specify both an upper and a lower bound, whereas 48 percent specify only an upper bound; only 2 percent select just a lower bound. Panels A and B of Figure A.1 show the distribution of minimum and maximum prices selected in the email alerts. Price range bounds are typically multiples of $\$50,000$, with particularly pronounced peaks at multiples of $\$100,000$.

There is significant heterogeneity in the breadth of the price ranges selected by different home searchers. Among those searchers who set both an upper and a lower bound, the 10th percentile selects a price range of $100,000, the median a price range of $300,000, and the 90th percentile a price range of $1.13 million. Panel C of Figure A.1 shows the distribution of price ranges both for those agents that select an upper and a lower bound, as well as for those agents that only select an upper bound.

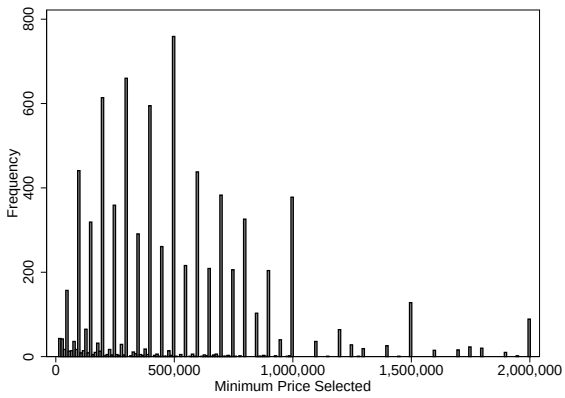
Panel D shows that searchers who consider more expensive houses specify wider price ranges. We bin the midprice of price ranges into 10 groups. The solid line (with values measured along the right-hand vertical axis) shows that the price range considered increases monotonically with the midpoint of the price range. One simple hypothesis consistent with this is that searchers set price ranges by choosing a fixed percentage range around a benchmark price. The bar chart (with percentages measured on the left hand vertical axis) shows that this is not the case: the percentage range is in fact U-shaped in price.

In addition to geography and price, the third dimension that is regularly populated in the email alerts is a constraint on the number of bathrooms. Panels E and F of Figure A.1 show the distribution of bathroom cutoffs selected for the Bay Area. 68% of all bathroom limits are set a value of 2, most of them as a lower bound. This setting primarily excludes 1 and 2 bedroom apartments and very small houses.

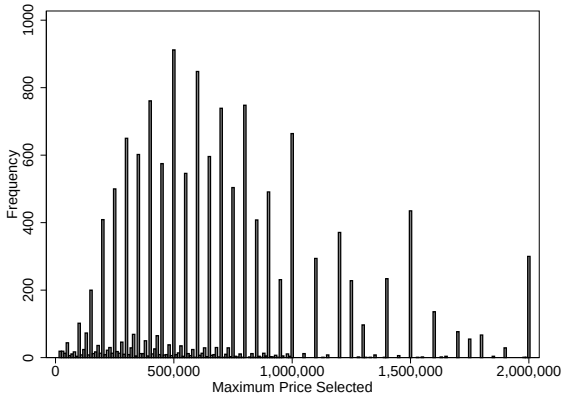
Tradeoffs between search dimensions

The three major search dimensions we have identified (geography, price, and size) are not necessarily orthogonal. For example, one can search for houses in a particular price range by looking only at zip codes in that price range or only at homes of a certain size. Table A.1 provides evidence on how different search dimensions interact. It shows that searchers who are more specific on price or home size search more broadly geographically. For example, searchers who specify a price restriction cover an average of 10.3 zip codes with an average maximum distance between centroids of 10.6 miles, while other searchers cover only 7.3 zip codes with an average maximum distance of 7.9 miles. Our segment construction discretizes the space of search characteristics and deals with searchers expressing their budget constraint or size preferences via geographic restrictions.

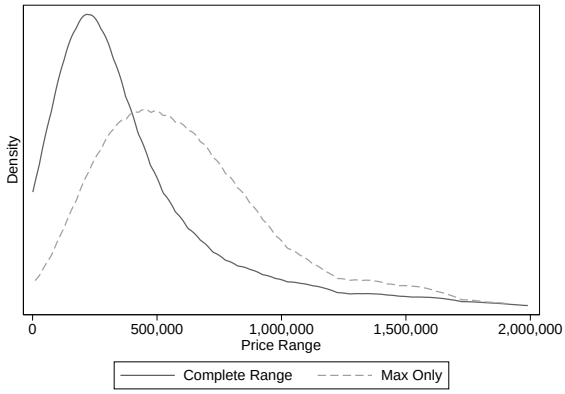
Figure A.1: Price and Size Criteria of Housing Search



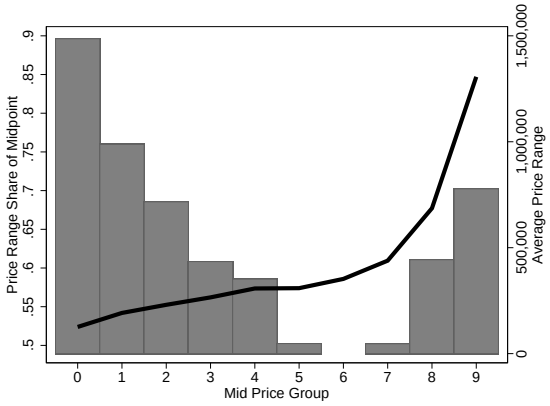
(A) Minimum House Price



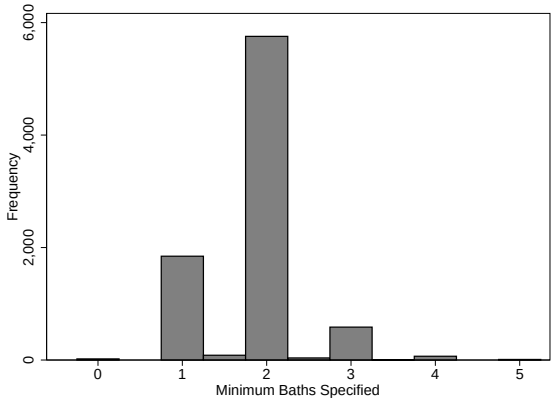
(B) Maximum House Price



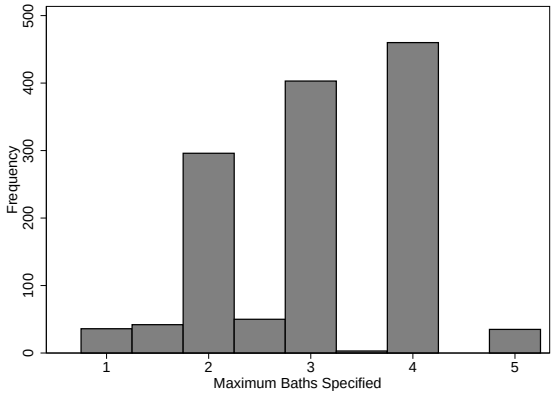
(C) Price Range



(D) Price Range by Midprice



(E) Minimum Number of Bathrooms



(F) Maximum Number of Bathrooms

Note: Panels A and B show histograms in steps of $10,000 of the minimum and maximum listing price parameters selected in email alerts. Panel C shows the distribution of price ranges across queries both for queries that only select a price upper bound (dashed line), as well as for those queries that select an upper bound and a lower bound (solid line). Panel D shows statistics only for those alerts that select an upper and a lower bound. The line chart shows the average price range by for different groups of mid prices, the bar chart shows the average of the price range as a share of the mid price. Panels E and F show histograms in steps of 0.5 of the minimum and maximum bathroom selected.

Table A.1: Geography, Price and Bath Parameter Interaction

	No Price	Price	No Bath	Bath
# Zips Covered	7.3	10.3	8.8	10.0
Max Distance (Miles)	7.9	10.6	8.9	11.1
Max Time Car (Min)	20.8	24.5	22.5	24.7
Max Time Public Transport (Min)	75.8	92.4	82.1	95.9
Is Contiguous	54%	62%	59%	60%

Note: Table shows summary statistics across queries that cross-tabulate moments across different search parameters.

A.2 Assigning Zip Codes and Segments to Search Alerts

To analyze segments in terms of their search clientele we need to analyze each query and determine which segments are covered by that query. We first focus on the geographic dimension of the search query, and determine which zip codes are covered by each query. To implement this, we need to deal with alerts that specify geography at a unit that might not perfectly overlap with zip codes. For alerts that select listings at the city level, we assign all zip codes that are at least partially within the range of the city that is covered by the search query (i.e. for a searcher who is looking in Mountain View, we assign the query to cover the zip codes 94040, 94041 and 94043). Neighborhoods and zip codes also do not line up perfectly, and so for each neighborhood we again consider all zip codes that are at least partially within the neighborhood that is covered by the search query (i.e. for a searcher who is looking in San Francisco's Mission District, we assign the query to cover zip codes 94103 and 94110). This provides us, for each search query, with a list of zip codes that are covered by that search alert.

We next incorporate the price and size dimensions of housing search to determine the set of segments covered by each query. The challenge is that price ranges selected by queries will usually not overlap perfectly with the price cutoffs of the individual segments. For those queries that specify a price dimension, we assign a query to cover a particular segmented in one of three cases:

1. When the query completely covers the segment (that is, when the query lower bound is below the segment cutoff and the query upper bound is above the segment cutoff).

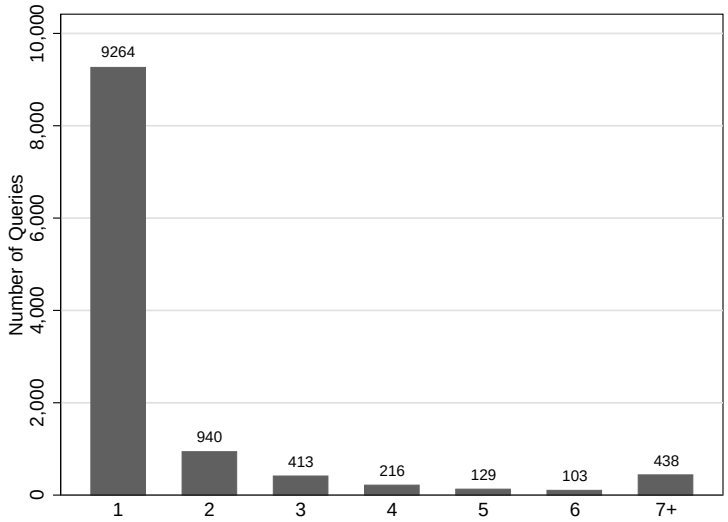
2. When the segment is open-ended (e.g. $1 million +), and the upper bound of the query exceeds the lower bound (in this case, all queries with an upper bound in excess of $1 million).

3. For queries that partially cover a non-open ended segment, we determine the share of the segment price range covered by the query. For example, for a segment $300k - $500k, the query 0-$250k covers 25% of the segment, and the query $300k - $700k covers 50% of the segment. We assign all queries that include at least 50% of the price range of a segment to cover that segment.

To incorporate the bathroom dimensions, we let a query cover a segment unless it is explicitly excluded. For example, queries that want at least two bathrooms will not cover the < 2 bathroom segments and vice versa.

The housing market segments constructed above allow us to pool across all search alerts set by the same individual. In particular, we add all segments that are covered by at least one search alert of an individual to that individual's search set. After pooling all segments covered by at least one email alert set by each searcher, we arrive at a total of 11,503 unique search profiles, set by the 23,597 unique users that comprise our dataset. Figure A.2 shows the distribution of how often each search profile is selected. A total of 9,264 search profiles are selected by only a single searcher, 940 queries are selected by 2 searchers. 438 queries are selected by at least 7 searchers, with the most common query getting selected 788 times.

Figure A.2: Number of Searchers per Unique Profile



Note: Figure shows how often each of the 11,503 individual search profiles is selected.

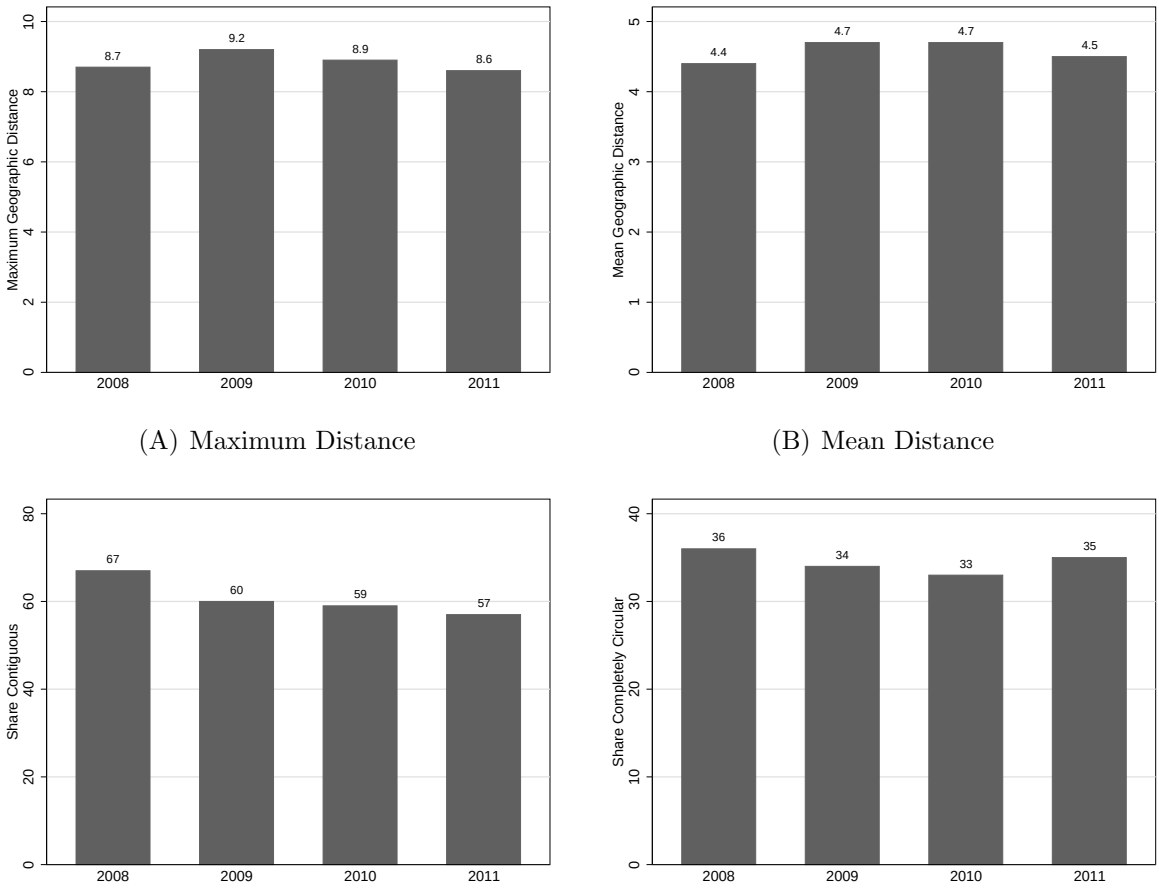
A.3 Stability of Search Patterns

Our model interprets the observed search ranges as a time-invariant feature of buyer preferences. It is then interesting to ask whether empirically the search ranges are indeed invariant to changes in market conditions. In particular, do searchers narrow the range of houses they consider when market activity is higher, and there is more inventory in each segment? We provide two tests that show no evidence that the parameters of housing search vary with market activity.

In a first test, we analyze the important summary statistics on the geographic breadth of each query between 2008 and 2011, split by the year in which the query was set. The results are presented in Figure A.3. We include the maximum distance between geographic zip code centroids (Panel A), the average distance between geographic zip code centroids (Panel B), the share of searches that yield contiguous search sets (Panel C), and the share of "circular" queries

(Panel D). We find that these important parameters of search activity are very stable across the years in our sample. This suggests that they do indeed capture time-invariant preference parameters of households over the Bay Area housing stock.

Figure A.3: Non-Cyclicality of Search Parameters



(A) Maximum Distance

(B) Mean Distance

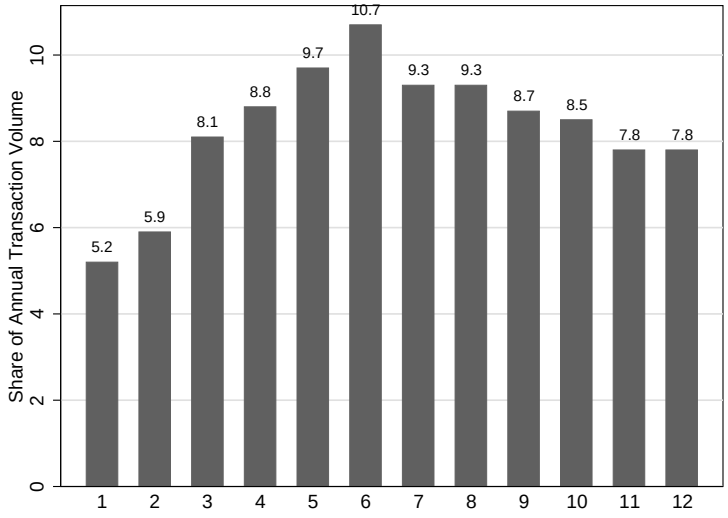
(C) Share Contiguous Queries

(D) Share Circular Queries

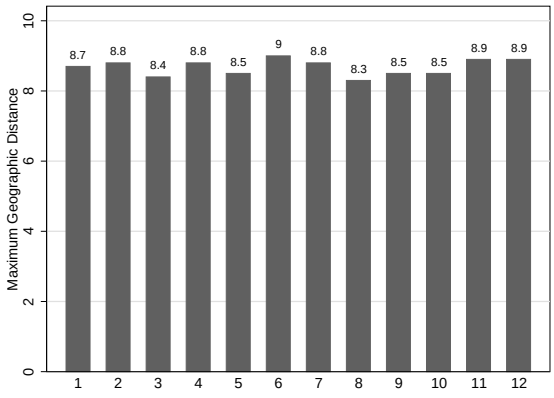
Note: Figure shows average values of search parameters by year of search query. We report the maximum distance between zip code centroids (Panel A), the mean distance between zip code centroids (Panel B), the share of contiguous queries (Panel C) and the share of circular queries (Panel D).

A second test of this hypothesis exploits seasonal variation in housing market activity: more houses typically trade in the summer as compared to the winter. This can be seen in Panel A of Figure A.4, which shows the share of total annual transaction volume over our sample in each month. Market activity is twice as high in June than it is in January. Panels B to E of Figure A.4 show averages of the same summary statistics on the search queries as Figure A.3, split by the month of the year when the query was set. As before, none of the search dimensions exhibit meaningful seasonality, consistent with an interpretation of search parameters as time-invariant measures of preferences that do not vary with market activity.

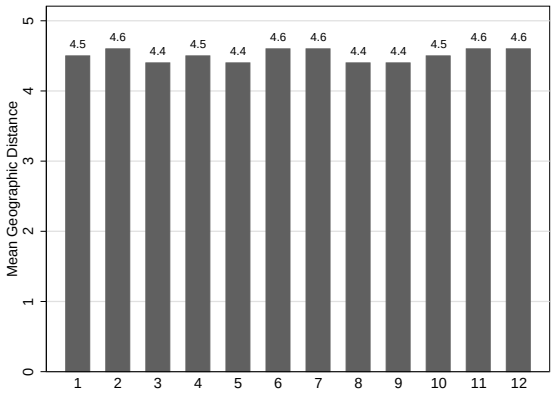
Figure A.4: Non-Seasonality of Search Parameters



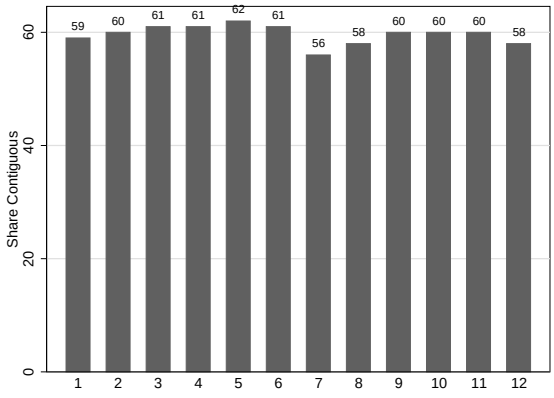
(A) Share of Annual Transactions



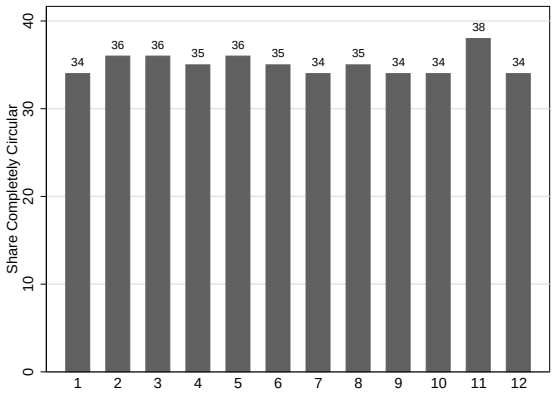
(B) Maximum Distance



(C) Mean Distance



(D) Share Contiguous Queries



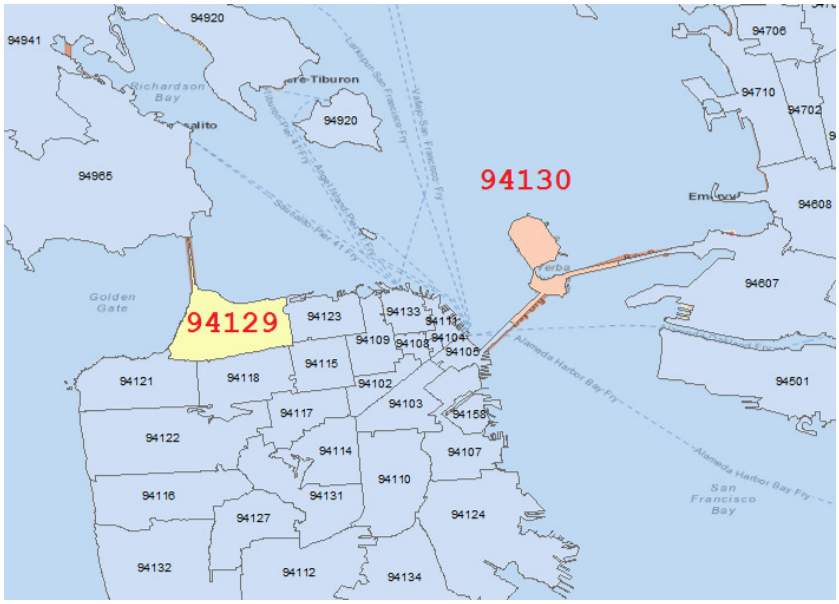
(E) Share Circular Queries

Note: Panel A shows the share of total annual transaction volume in each month. Panels B - E show average values of search parameters by month of search query. We report the maximum distance between zip code centroids (Panel B), the mean distance between zip code centroids (Panel C), the share of contiguous queries (Panel D) and the share of circular queries (Panel E). Months increase from January to December on the horizontal axis.

A.4 Constructing Contiguity Measures

To analyze whether all zip codes covered by a particular search query are contiguous, one challenge is provided by the San Francisco Bay. The location of this body of water means that two zip codes with non-adjacent borders should sometimes be considered as contiguous, since they are connected by a bridge such as the Golden Gate Bridge. Figure A.5 illustrates this. Zip codes 94129 and 94965 should be considered contiguous, since they can be traveled between via the Golden Gate Bridge. To take the connectivity provided by bridges into account, we manually adjust the ESRI shape files to link zip codes on either side of the Golden Gate Bridge, the Bay Bridge, the Richmond-San Rafael Bridge, the Dumbarton Bridge and the San Mateo Bridge. In addition, there is a further complication in that the bridgehead locations are sometimes in zip codes that have essentially no housing stock, and are thus never selected in search queries. For example, 94129 primarily covers the Presidio, a recreational park that contains only 271 housing units. Similarly, 94130 covers Treasure Island in the middle of the SF Bay, again with only a small housing stock. These zip codes are very rarely selected by search queries, which would suggest, for example, that 94105 and 94607 are not connected. This challenge is addressed by manually merging zip codes 94129 and 94130 with the Golden Gate and Bay bridge respectively. This ensures, for example, that 94118 and 94955 are connected even if 94129 was not selected.

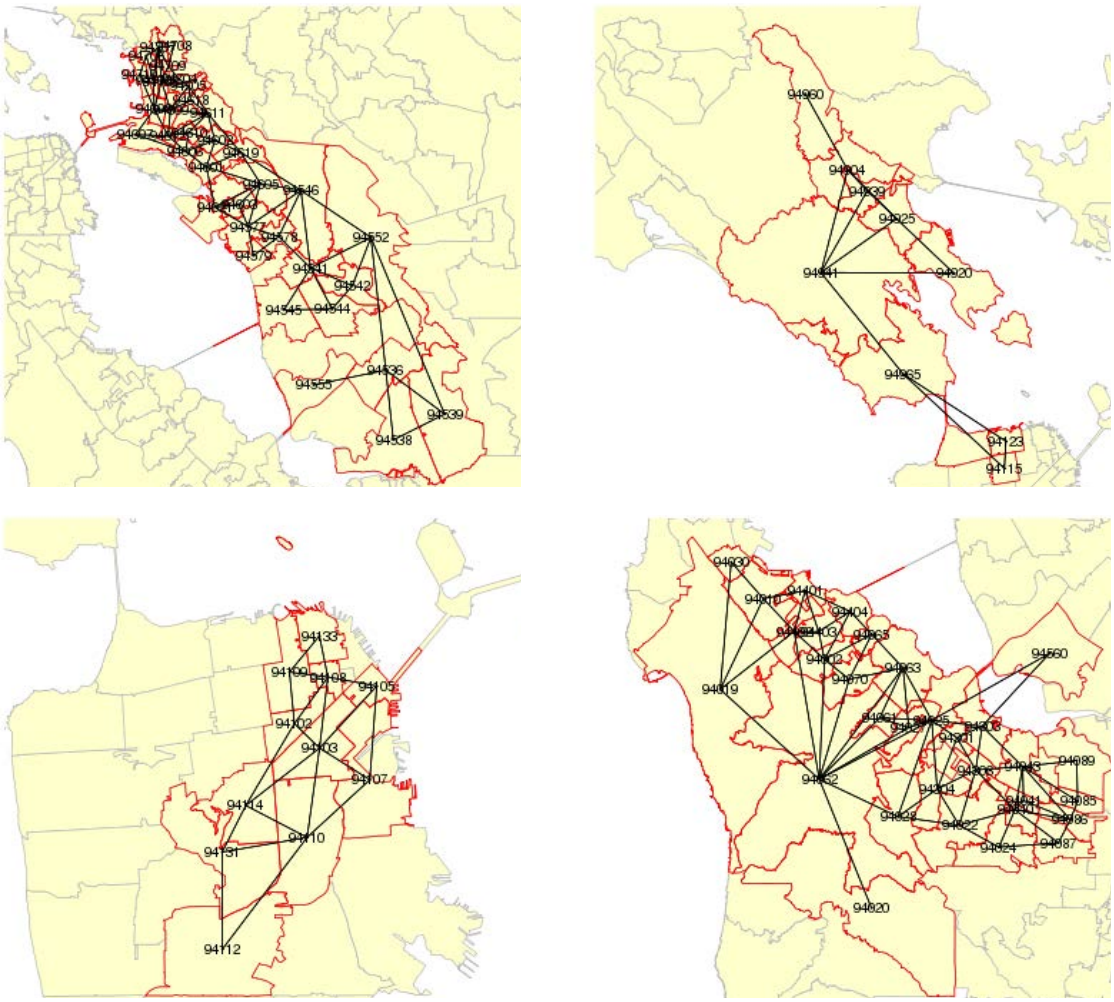
Figure A.5: Bridge Adjustments - Contiguity Analysis



Note: This figure shows how we deal with bridges in the Bay Area for the contiguity analysis.

In the following we provide examples of contiguous and non-contiguous search sets. In Figure A.6 we show four actual contiguous search sets. The top left panel shows all the zip codes covered by a searcher that searched for homes in Berkeley, Fremont, Hayward, Oakland and San Leandro.

Figure A.6: Sample Contiguous Queries

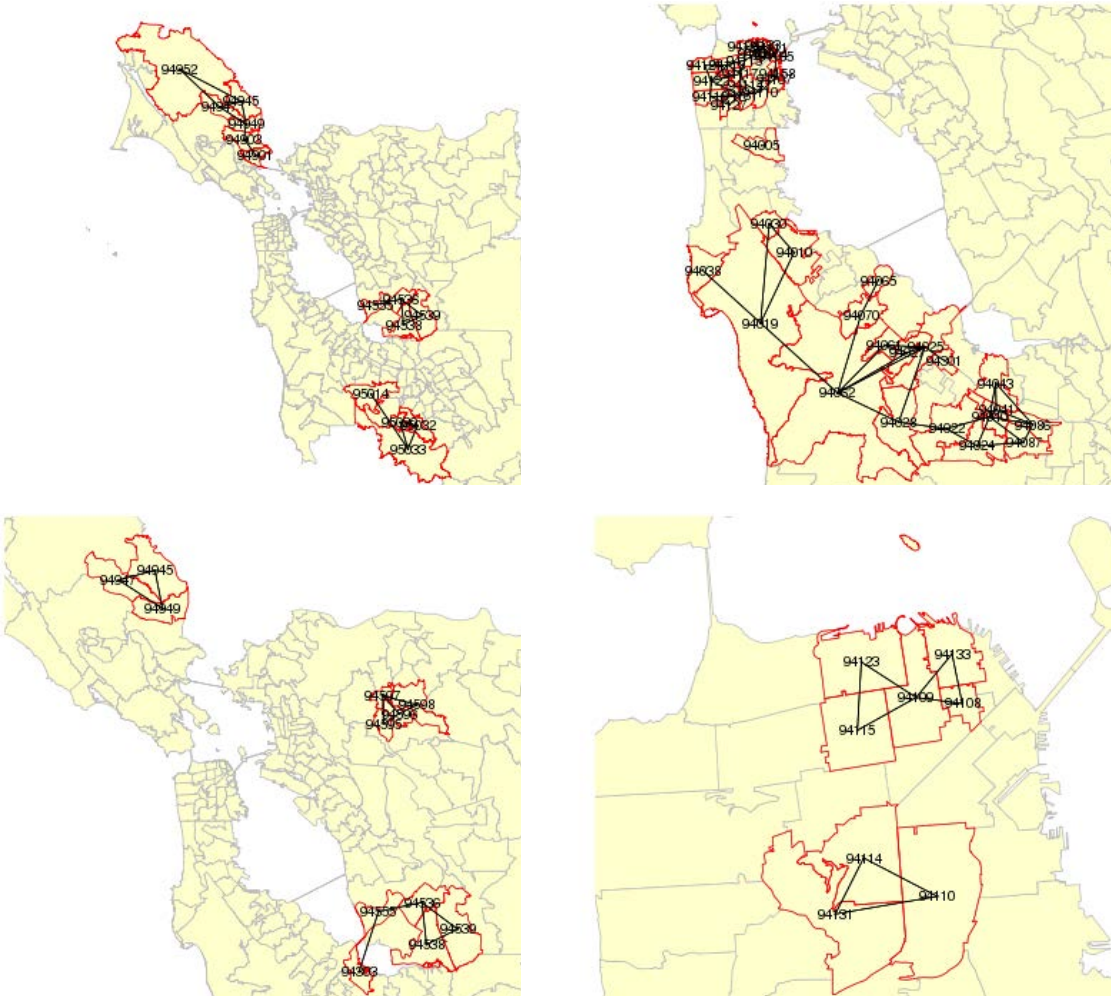


Note: This figure shows a sample of contiguous search sets. The zip codes selected by the searcher are circled in red. Zip code centroids of contiguous zip codes are connected.

This is a relatively broad set, covering most of the East Bay. The top right panel shows a contiguous set of jointly searched zip codes, with connectivity derived through the Golden Gate Bridge. The searcher queried homes in cities north of the Golden Gate Bridge (Corte Madera, Larkspur, Mill Valley, Ross, Kentfield, San Anselmo, Sausalito and Tiburon), but also added zip codes 94123 and 94115. The bottom left panel shows the zip codes covered by a searcher that selected a number of San Francisco neighborhoods. The final contiguous search set (bottom right panel) was generated by a searcher that selected a significant number of South Bay cities.[12] These are all locations with reasonable commuting distance to the tech jobs in the Silicon Valley. Notice how the addition of Newark adds zip code 94550 in the East Bay, which is connected to the South Bay via the Dumbarton Bridge.

[12] Atherton, Belmont, Burlingame, El Granada, Emerald Hills, Foster City, Half Moon Bay, Hillsborough, La Honda, Los Altos Hills, Los Altos, Menlo Park, Millbrae, Mountain View, Newark, Palo Alto, Portola Valley, Redwood City, San Carlos, San Mateo, Sunnyvale, Woodside.

Figure A.7: Sample Non-Contiguous Queries



Note: This figure shows a sample of non-contiguous search sets. The zip codes selected by the searcher are circled in red. Zip code centroids of contiguous zip codes are connected.

In Figure A.7 we show four actual non-contiguous search sets. The top left panel shows the zip codes covered by a searcher that selects the cities of Cupertino, Fremont, Los Gatos, Novato, Petaluma and San Rafael. This generates three contiguous set of zip codes, rather than one large, contiguous set. The zip codes in the bottom right belong to a searcher that selected zip code 94109 and the neighborhoods Nob Hill, Noe Valley and Pacific Heights. Again, this selection generates more than one set of contiguous zip codes.

Table A.2 shows summary statistics of our measure of contiguity by the number of zip codes included in the search range. The second column reports the share of searchers who select contiguous geographies. While overall only 18 percent of searchers have non-contiguous search ranges, they tend to be broad searchers, who consider more than five distinct zip codes and hence provide market integration across neighborhood and city boundaries. The third and fourth columns report the mean and max number of contiguous areas covered by a search range. Broad searchers

often consider multiple distinct contiguous areas. Preference for certain cities plays a role here: the increase in the share of contiguous queries for the group with 21-30 zip codes selected can be explained by the prevalence of searches for "San Francisco" and "San Jose" in that category.

Table A.2: Contiguity Analysis – Summary Statistics

		Contiguous Areas		
Number of Zips Covered	Share Contiguous	Mean	Max	Total
2	91%	1.09	2	2,927
3	83%	1.18	3	1,761
4	91%	1.10	3	2,248
5	67%	1.37	4	844
6-10	71%	1.38	5	2,612
11-20	74%	1.38	8	2,071
21-30	91%	1.13	10	4,213
30+	48%	1.94	9	798
Total	82%	1.24	10	17,474

Note: Table shows summary statistics for contiguity measures across searchers that select different number of zip codes.

A.5 Segment Construction

This section describes the process of arriving at the set of 564 distinct housing market segments for the San Francisco Bay Area. As before, we select the geographic dimension of segments to be a zip code. Since we will compute average price, volume, time on market and inventory for each segment, we restrict ourselves to zip codes with at least 800 arms-length housing transactions between 1994 and 2012. This leaves us with 191 zip codes with sufficient observations to construct these measures.

We next consider how to further split these zip codes into segments based on a quality (price) and size dimension. Importantly, we will need to observe the total housing stock in each segment in order to appropriately normalize moments such as turnover and inventory. The residential assessment records do contain information on the universe of the housing stock. However, as a result of Proposition 13, the assessed property values in California do not correspond to true market value, and it is thus not adequate to divide the total zip code housing stock into different price segments based on this assessed value.[13] To measure the housing stock in different price segments we use the U.S. Census Bureau's 2011 American Community Survey 5-year estimates, which report the total number of owner-occupied housing units per zip code for a number of price bins. We combine a number of these bins to construct the total number of housing units in each of the following price bins: $< \$200\text{k}, \$200\text{k}-\$300\text{k}, \$300\text{k}-\$400\text{k}, \$400\text{k}-\$500\text{k}, \$500\text{k}-\$750\text{k},$

[13] Allocating homes that we observe transacting into segments based on value is much easier, since this can be done on the basis of the actual transaction value, which is reported in the deeds records.

$750k−$1m, > $1m. These bins provide the basis for selecting price cut-offs to delineate quality segments within a zip code. One complication is that the price boundaries are reported as an average for the sample years 2006-2010. Since we want segment price cut-offs to capture within zip code time-invariant quality segments, we need to adjust for average market price changes of the same-quality house over time. To do this, we adjust all prices and price boundaries to correspond to 2010 house prices.[14]

Not all zip codes have an equal distribution of houses in each price (quality) bin. For example, Palo Alto has very few homes valued at less than $200,000, while Fremont has very few million-dollar homes. Since we want to avoid cutting a zip code into too many quality segments with essentially no housing stock to allow us to measure segment-specific moments such as time on market, we next determine a set of three price cut-offs for each zip code by which to split that zip code. To determine which of the seven census price bin cut-offs should constitute segment cut-offs, we use information from the search queries. This proceeds in two steps: First we change the price parameters set in the email alerts to account for the fact that we observe queries from the entire 2006 - 2012 period. This adjusts the price parameters in each alert by the market price movements of homes in that zip code between the time the query was set and 2010.[15] Second, we determine which set of three ACS cutoffs is most similar to the distribution of actual price cut-offs selected in search queries that cover a particular zip code. For each possible combination of three (adjusted) price cut-offs from the list of ACS cut-offs, we calculate for every email alert the minimum of the absolute distance from each of the (adjusted) search alert price restrictions to the closest cut-off.[16] We select the set of segment price cut-offs that minimizes the average of this value across all queries that cover a particular zip code. This ensures, for example, that if there are many queries that include a high limit such as $1 million, $1 million is likely to also be a segment boundary.

[14]This is necessary, because the Census Bureau only adjusts the reported values for multi-year survey periods by CPI inflation, not by asset price changes. This means that a $100,000 house surveyed in 2006 will be of different quality to a $100,000 house surveyed in 2010. We choose the price that a particular house would fetch in 2010 as our measure of that home's underlying quality. To transform the housing stock by price bin reported in the ACS into a housing stock by 2010 "quality" segment, we first construct zip code specific annual repeat sales price indices. This allows us to find the average house price changes by zip code for each year between 2006 and 2010 to the year 2010. We then calculate the average of these 5 price changes to determine the factor by which to adjust the boundaries for the price bins provided in the ACS data. Adjusting price boundaries by a zip code price index that looks at changes in median prices over time generates very similar adjustments.

[15]This ensures that the homes selected by each query correspond to our 2010 quality segment definition. Imagine that prices fell by 50% on average between 2006 and 2010. This adjustment means that a query set in 2006 that restricts price to be between $500,000 and $800,000 will search for homes in the same quality segment as a query set in 2010 that restricts price to a $250,000 - $400,000 range.

[16]For example, imagine testings how good the the boundaries 100k, 300k and 1m fit for a particular zip code. A query with an upper bound of 500k has the closest absolute distance to a cut-off of $\min\{|500-100|, |500-300|, |500-1000|\} = 200$. A query with an upper bound of 750k has the closest absolute distance to a cut-off of 250. A query with a lower bound of 300k and an upper bound of 600k has the closest absolute distance to a cut-off of 0. For each possible set of price cut-offs, we calculate for every query the smallest absolute distance of a query limit to a cut-off, and then find the average across all search alerts.

To determine the total housing stock in each price by zip code segment, one additional adjustment is necessary. Since the ACS reports the total number of owner-occupied housing units, while we also observe market activity for non owner-occupied units, we need to adjust the ACS-reported housing stock for each price bin by the corresponding homeownership rate. To do this, we use data from all observed arms-length ownership-changing transactions between 1994 and 2010 as reported in our deeds records. We first adjust the observed transaction price with the zip code level repeat sales price index, to assign each house for which we observe a transaction to one of our 2010 price (quality) bins. For each of these properties we also observe from the assessor data whether they were owner-occupied in 2010. This allows us to calculate the average homeownership rate for each price segment within a zip code, and adjust the ACS-reported stock accordingly.[17] To assess the quality of the resulting adjustment, note that the total housing stock across our segments is approximately 2.2 million, very close to the total Bay Area housing stock reported in the 2010 census.

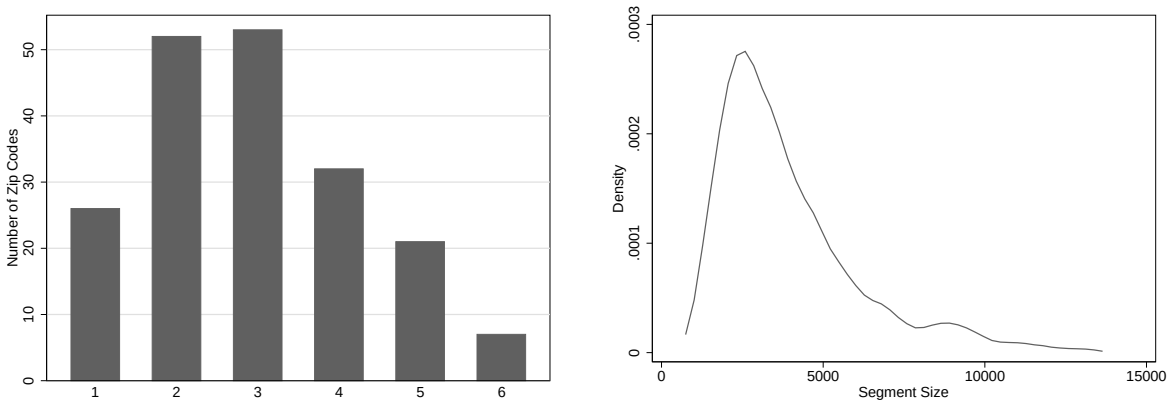
The other search dimension regularly specified in the email alerts, and that we hence want to incorporate in our segment definition, is the number of bathrooms as a measure of the size of a house conditional on its location and quality. Since Section A.1 showed that the vast majority of constraints on the number of bathrooms selected homes with either more or fewer than two bathrooms, we further divide each zip code by price bucket group into two segments: homes with less than two bathrooms, and homes with at least two bathrooms. Unfortunately the ACS does not provide a cross-tabulation of the housing stock by home value and the number of bathrooms. To split the housing stock in each price and zip code segment into the two groups by home size, we apply a similar method as above to control for homeownership rate. We use the zip code level repeat sales price index to assign each home transacted between 1994 and 2010 to a 2010 price (quality) bin. For these homes we observe the number of bathrooms from the assessor records. This allows us to calculate the average number of bathrooms for transacted homes in each zip code by price segment. We use this share to split the total housing stock in those segments into two bathroom size groups.

The approach described above splits each zip code into eight initial segments along three price cutoffs and one size cutoff. For each of these segments, we have an estimate of the total housing stock. Since we need to measure specific moments such as the average time on market with some precision, we need to ensure that each segment has a housing stock of at least 1,200 units. If this is not the case the segment is merged with a neighboring segment until all remaining segments have a housing stock of sufficient size. For price segments where either of the two size subsegments have a stock of less than 1,200, we merge the two size segments. We then begin with the lowest

[17] For example, the 2010 adjusted segment price cutoffs for zip code 94002 are \$379,079, \$710,775 and \$947,699. This splits the zip code into 4 price buckets. The homeownership rate is much higher in the highest bucket (95%) than in the lowest bucket (65%). This shows the need to have a price-bucket specific adjustment for the homeownership rate to arrive at the correct segment housing stock.

price segment, see whether it has a stock of less than 1,200, and merge it with the next higher price segment. This procedure generates 564 segments. Figure A.8 shows how many segments each zip code is split into. 26 zip codes are not split up further into segments. 52 zip codes are split into two segments, 53 zip codes are split into 3 segments. The right panel of figure A.8 shows the distribution of housing stock across segments. On average, segments have a stock of 3,929, with a median value of 3,298. The largest segment has a housing stock of 13,167.

Figure A.8: Segment Overview



Note: Figure shows summary statistics across segments. The left panel shows the number of segments that the 191 zip codes are split into. The right panel shows the distribution of the number of housing units across segments.

A.6 Construction of Segment-Level Market Activity

Our model links the characteristics of search patterns to segment specific measures of market activity such as price, volume, time on market and inventory. In this section we describe how we construct these moments at the segment level. We begin by identifying a set of arms-length transactions, which are defined as transactions in which both buyer and seller act in their best economic interest. This ensures that transaction prices reflect the market value (and hence the quality) of the property. This excludes, for example, intra-family transfers. We drop all observations that are not a Main Deed or only transfer partial interest in a property (see Stroebel, 2014, for details on this process of identifying arms-length transactions).

We can then calculate the total number of transactions per segment between 2008 and 2011, and and use this to construct annual volume averages. In order to allocate houses to particular segments, we adjust transaction prices for houses sold in years other than 2010 by the same price index we used to adjust listing price boundaries (see Appendix A.5). We arrive at our measure of "volume share" by dividing the annual transaction volume by the segment housing stock.

To calculate the average time on market, we use the dataset on all home listings on trulia.com, beginning in January 2006, and match those home listings with final transactions from the deeds

database. We find segment-specific measures of time on market by averaging the time on market across all transactions that sold between 2008 and 2011.[18]

A.7 Identification

This appendix shows that our model implies a one-to-one mapping between two sets of numbers. The first set consists of the parameters $\eta(h)$ and $\mu^{\Theta}(\theta)$ as well as the vector of rates at which buyers find houses in a given segment, defined as $\alpha(h) = m(h)/\mu^{B}(h)$. The second set consists of the inventory share $I(h)$, the turnover rate $V(h)$, the relative frequencies of search ranges $\beta(\theta)$ and the *average* time it takes for a buyer to find a house.

We use equations (5)-(7) to derive closed form expressions for the first set of numbers in terms of the second. The frequency of moving shocks $\eta(h)$ can be written directly as a function of inventory and turnover: we divide the market clearing condition (5) by the housing stock $\mu^{H}(h)$ to obtain

$$\eta(h)(1 - I(h)) = V(h). \tag{A.1}$$

The match rate for a buyer who flows to segment h is $\alpha(h) = m(h)/\mu^{B}(h)$. Using the definition of buyers (4), it can be expressed in terms of observables (up to a constant) as

$$\frac{1}{\alpha(h)} = \sum_{\theta \in \tilde{\Theta}(h)} \frac{I(h)}{\tilde{I}(\theta)} \frac{\beta(\theta)\left(\bar{\mu}^{\Theta} - 1\right)}{\nu^{H}(\theta)} \frac{1}{V(h)}. \tag{A.2}$$

Interpreting terms from the right, we have that matching is fast – at a high rate $\alpha(h)$ – in segment h if the volume share is high in h, if the buyer-owner ratio is high for types in the clientele of h, and if the inventory share is low in h relative to other segments in its clientele's search ranges.

It remains to identify the distribution of searcher types μ^{Θ}. We determine the constant $\bar{\mu}^{\Theta} - 1$ by setting the average of the buyer match rates $\alpha(h)$ to the average of the inventory match rate $I(h)/V(h)$. The average inventory-weighted match rate across types θ is the same as the average inventory-weighted match rate across segments. Indeed, let $\bar{\mu}^{S}$ denote total inventory and consider the identity

$$\left(\bar{\mu}^{S}\right)^{-1} \sum_{\theta \in \Theta} \frac{\nu^{S}(\theta)}{\tilde{\mu}^{B}(\theta)} \sum_{h \in \tilde{H}(\theta)} \frac{\mu^{S}(h)}{\nu^{S}(\theta)} \frac{\tilde{\mu}^{B}(\theta)}{\mu^{B}(h)} m(h) = \left(\bar{\mu}^{S}\right)^{-1} \sum_{h \in \tilde{H}(\theta)} \frac{\mu^{S}(h)}{\mu^{B}(h)} m(h).$$

Here the right hand side is the average match rate across segments and the left hand side is the average match rate across types. In particular, the second sum on the left hand side is number of matches entered by type θ which depends on the relative inventory available in θ's search range

[18] In the very few instances when the listing price and the final sales price would suggest a different segment membership for a particular house – i.e. cases where the house is close to being at a segment boundary and sells for a price different to the listing price – we allocate the house to the segment suggested by the sales price, not the listing price.

as well as the share of θ in each segment's buyer pool.

Once the total number of buyers is determined, the number of buyers by type $\tilde{\mu}^B(\theta) = \beta(\theta)\left(\bar{\mu}^\Theta - 1\right)$ and by segment $\mu^B(h) = \mu^H(h) V(h)/\alpha(h)$ follow immediately. Substituting for $\mu^H(h;\theta)$ in (7) using (6) we can solve out for the type distribution $\mu^\Theta(\theta)$ from

$$\frac{\mu^\Theta(\theta) - \tilde{\mu}^B(\theta)}{\tilde{\mu}^B(\theta)} = \sum_{h \in \tilde{H}(\theta)} \frac{\mu^S(h)}{\nu^S(\theta)} \frac{\mu^H(h)}{\mu^B(h)}. \tag{A.3}$$

The adding up constraint says that the owner-buyer ratio for type θ agents should be the inventory weighted average of owner-buyer ratios at the segment level. We will therefore infer the presence of more types θ not only if we observe more buyers of type θ (higher $\beta(\theta)$ and hence higher $\tilde{\mu}^B(\theta)$), but also if type θ's search range has on average relatively more owners relative to buyers. In the latter case, more types θ agents are themselves owners, so their total number is higher.

At this point we have identified the supply and demand parameters of the model without specific assumptions on the functional form of the matching function. If we postulate such a functional form, restrictions on its parameters follow from equation (5). For example, consider the Cobb-Douglas case with a multiplicative segment-specific parameter $\bar{m}(h)$ that governs the speed of matching

$$\tilde{m}\left(\mu^B(h), \mu^S(h), h\right) = \bar{m}(h)\, \mu^B(h)^\delta\, \mu^S(h)^{1-\delta}.$$

For a given weight δ, the speed of matching parameter $\bar{m}(h)$ can be backed out from observables as $\bar{m}(h) = \alpha(h)^\delta \left(V(h)/I(h)\right)^{1-\delta}$. The speed of matching parameter is thus a geometric average of the buyer and inventory match rates.