Abstract

This paper studies housing markets with multiple segments searched by heterogeneous clientele. We explore 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 these differences in housing market activity at different levels of aggregation, determines price discounts due to market frictions, and shapes the response of local housing markets to supply and demand shocks.

1 Introduction

Analyses of search markets benefit from both sell-side and buy-side statistics. For example, the state of the labor market is typically summarized by both job vacancies and the unemployment rate. The empirical labor literature documents that the Beveridge curve slopes downwards, as higher rates of unemployment tend to occur when there are fewer vacancies. For the housing market, however, we usually do not observe any information about potential home buyers. A “housing shortage” in some market is thus often identified with low inventory, even if the number of potential buyers is unknown (e.g., Forbes, 2016; USA Today, 2016). This approach is justified if low inventory always goes along with a large number of searchers – that is, if the housing Beveridge curve is also downward sloping – but can otherwise lead to incorrect inference and policy decisions.

In this paper, we explore the relationship between housing vacancies and housing demand at various levels of geographic aggregation. We use a novel dataset on housing search behavior to show that the housing Beveridge curve indeed slopes down in the cross section of cities. Within cities, however, it slopes up: housing market segments with low inventory draw interest from fewer searchers. A quantitative search model attributes much of this contrast to the composition

*This version: May 1, 2017. Author contact information: piazzesi@stanford.edu, schneider@stanford.edu, johannes.stroebel@nyu.edu. We thank Ed Glaeser, Bob Hall, John Leahy, Giuseppe Moscarini, Robert Shimer, Vincent Sterk, Silvana Tenreyro, four anonymous referees, and conference and seminar participants at the AEA meetings, Arizona State, Berkeley, Chicago, LBS, Mannheim, Michigan, Minneapolis Fed, Minnesota, NBER Summer Institute, Northwestern, NYU, Philadelphia Fed, SITE, SED, Stanford, Tsinghua, UCLA, USC, and Yale. Trulia for providing data. This research is supported by a grant from the National Science Foundation.
of buyer clienteles. Indeed, we show that detailed information on homebuyers’ search ranges – as is available in our data – is important for understanding local housing market behavior. It is also important for understanding the effects of gentrification and for conducting inference about the consequences of possible local housing policies, such as changes in zoning regulations.

Our empirical analysis studies housing market activity in the San Francisco Bay Area. We infer search ranges of potential buyers from online housing search: on the popular real estate website trulia.com, home searchers can set an alert that triggers an email whenever a house with their desired characteristics comes on the market. We observe the search parameters in a large sample of such email alerts. Housing search occurs predominantly along three dimensions: geography, price and, to a lesser extent, house size as captured by the number of bathrooms. To relate search to other market activity, we divide the Bay Area into 564 distinct market segments along the dimensions suggested by the observed search ranges. We then measure the cross section of turnover and inventory at the segment level from deeds records, assessment data and “for sale” listings. We express search ranges as subsets of the set of all segments, and measure search activity at the segment level by the number of searchers per house. The key stylized fact about buyer clienteles is that market segments within cities feature competition between narrow searchers whose range contains only one neighborhood or a limited price range, and broad searchers who consider the entire city. In contrast, different cities share relatively few common searchers.

Our model relates these cross-sectional facts about searcher clienteles to differences in the Beveridge curve at different levels of aggregation; the key intuition is consistent with many search models. Across cities – which are searched by distinct clienteles – the key variation is that some cities are more popular, that is, their clientele is larger relative to the housing stock. In a popular city, the number of searchers is thus large, and any house that comes on the market is sold quickly, so inventory is low. With distinct clienteles, variation in city popularity thus generates the downward sloping Beveridge curve that is observed in the data. Moreover, building houses in cities where inventory is low addresses a housing shortage: it targets cities where local excess demand is high and all new houses satisfy local demand.

In contrast, across segments within a city – which attract both broad and narrow searchers – the key variation is that some segments are less stable, that is, houses come on the market at a faster rate there. Inventory is higher in unstable segments. This attracts broad searchers who are more likely to find their favorite house in those segments within their search range that have more inventory. In unstable high-inventory segments, broad searchers thus crowd out ownership by narrow searchers who have to search longer to find a house. This crowd-out increases the number of equilibrium searchers. Variation in stability can thus lead to upward sloping Beveridge curves in markets where broad and narrow searchers interact. Moreover, building houses where inventory is low selects segments with relatively low local excess demand. Instead, the additional inventory mostly attracts more broad searchers who have no specific preference for the location.

Formally, we build on a version of the Diamond-Mortensen-Pissarides random matching model
with fixed numbers of 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, an assumption that is motivated by observed search patterns in our data. Prices reflect the present value of housing services less an illiquidity discount due to search and transaction costs.

Our quantitative analysis proceeds in three steps. We begin by focusing on quantities. We show that the distribution of preferences, moving shocks, and a measure of matching frictions is identified from cross-sectional moments of turnover, inventory, and search activity. This identification result is independent of the details of price bargaining and the matching function. We estimate that distribution and derive summary statistics of supply and demand conditions at the segment level: in particular, instability is the rate at which moving shocks arrive in a segment, and popularity is the size of the segment clientele, with broader searchers counting less towards popularity in any segment they cover. As an over-identifying restrictions test on our model, we show that population flows between segments implied by our model are consistent with observed moves in the data.

We then relate the inferred segment-level measures of popularity to observable characteristics of these segments. We find that segments with better schools, better restaurants, and better weather all attract a larger search clientele. In addition, we explore the relationship between search breadth and searcher demographics. We find that younger households, poorer households, and households with children have broader housing search ranges.

To quantify the importance of detailed information on buyer search ranges, we compare our estimated model to a misspecified benchmark that assumes all searchers narrowly target just one segment. The benchmark exercise follows the common approach in the literature of using data on turnover, inventory, and the time it takes to find a house to pin down parameters of a search model. In this benchmark without broad searchers, all observed demand is attributed to local demand from narrow searchers. We show that this approach leads researchers comparing segments within a city to infer that unstable segments with high inventory are substantially more popular than they actually are. On the other hand, when analyzing housing markets across cities which share few broad searchers, the traditional approach in the literature is much less biased.

We further illustrate the role of heterogeneous clienteles through a number of counterfactual analyses. In particular, we ask how housing markets respond if a segment becomes more popular via an increase in the number of narrow searchers who want to buy there. The answer crucially depends on the overall number of searchers and what other markets those searchers look at. For example, shocks to downtown San Francisco segments with many searchers who search broadly are transmitted widely across the city, without a substantially larger effect in the segments that have actually become more popular. In contrast, the effects of shocks to suburban segments close to the San Francisco city boundary, which primarily attract narrow searchers, are concentrated
in those segments, with only small effects on nearby housing markets. These results highlight that the impact of local housing market policies will depend on the characteristics of the local clientele. For example, whether local zoning regulations to increase the available housing stock in a neighborhood will have a large impact on local prices and time on market will depend on whether that neighborhood is considered primarily by narrow or broad searchers.

The second part of our quantitative analysis considers price formation. We use our parameter estimates to infer illiquidity discounts across segments, which capture the capitalized value of trading frictions that current and future buyers face. 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). Illiquidity discounts are larger in segments where houses take a long time to sell, for example because these segments do not attract many broad searchers. Illiquidity discounts are also larger in less stable segments, where houses transact more often.

While our analysis highlights the importance of understanding the interaction of broad and narrow searchers in the housing market, our insights are likely to also be important in other search markets. For example, Treasury securities are sold in over-the-counter search markets. In these markets, some investors might be particularly interested in purchasing Treasuries of certain maturities (e.g., pensions funds engaging in duration matching might only buy long-duration Treasuries), but there might be other buyers, such as hedge funds, that consider a broader range of maturities. Understanding the segmentation of the buyer clientele, and the interaction of broad and narrow investors at different maturities, is important for determining the optimal maturity structure of newly-issued debt. For example, monetary policy interventions, such as the maturity extension program (“Operation Twist”), which aimed at flattening the yield curve by selling short-maturity Treasuries and buying long-maturity Treasuries, will be less effective in the presence of “broad” investors who are indifferent between buying Treasuries with a wide range of maturities (see Swanson, 2011; Greenwood and Vayanos, 2014).

Similarly, the degree of segmentation in labor market search is substantial and time-varying. For example, recent research has documented that job seekers in areas with depressed housing markets apply for fewer jobs that require relocation, because of the difficulties with selling underwater homes (see Brown and Matsa, 2016). Our model allows researchers to understand the extent to which the resulting increase in regional segmentation of labor search contributes to a decline in the ability of labor flows to facilitate regional risk sharing. In addition, labor market segmentation across industry and occupational groups has increased as a result of the specialization of human capital, with many vacancies only attracting applications from a small set of highly-specialized job seekers. The mechanisms highlighted in this paper show how such changes in segmentation influence the employment effects of immigration, and affect the efficacy of policies such as targeted job training programs (e.g., Peri and Sparber, 2009).1

1Other segmented search markets with an interaction of broad and narrow searchers can also be better understood through the lens of our model. For example, in the dating market, some searchers with very narrow
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 Goodman and Thibodeau, 1998; Leishman, 2001; 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 experiences. 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 and across potential buyers.\footnote{As such, our paper also contributes to the literature that considers house valuation with heterogeneous regions and buyers. For example, Peterba (1991) consider the role of demographics for prices, Bayer, Ferreira and McMillan (2007) look at school quality, and Hurst, Guerrieri and Hartley (2013) study the effects of gentrification. Stroebel (2016) and Kurlat and Stroebel (2015) investigate the market impact of asymmetric information about property and neighborhood characteristics, respectively. Landvoigt, Piazzesi and Schneider (2015) 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 illiquidity discounts due to search frictions and transactions costs, as well as heterogeneity in search preferences within a metro area. Davis and Van Nieuwerburgh (2015) and Piazzesi and Schneider (2017) survey the broader literature on business cycles, asset prices and housing, including studies that do not rely on search frictions.}

We build on a large literature on housing search models (see Han and Strange, 2015, for a recent survey). Recent studies have turned to quantitative evaluations using micro data (see Diaz and Perez, 2013; Head et al., 2014; Ngai and Tenreyro, 2014; Guren and McQuade, 2014; Anenberg and Bayer, 2014; Halket and Pignatti, 2015; Burnside et al., 2016; Guren, 2017). Most of these authors are interested in how search modifies the time series dynamics of house prices and market activity in homogeneous housing markets. In contrast, our focus is on steady states of markets with rich cross sections of houses and buyers. Our theoretical model is based on earlier random matching models such as Wheaton (1990), Krainer (2001), Albrecht, Anderson, Smith and Vroman (2007), Novy-Marx (2009), and Piazzesi and Schneider (2009); the new element is that we allow for matching by heterogeneous buyers across multiple interconnected segments.

The paper in the labor search literature that is closest to our work is by Manning and Petrongolo (2017), who estimate a search and matching model for local labor markets. They divide space into small areas and allow searchers to look in many areas simultaneously, thus generating spillover effects across regions. Their demand side data consists of addresses for a job seeker’s home and the vacancies he applies to. They infer a distribution of preferences that includes a parameter for 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 online housing search behavior, using both spatial and quality information to define the commodity space.
The rest of the paper is structured as follows. Section 2 describes our housing search data, and presents key patterns of housing search behavior. Section 3 establishes stylized facts on housing market and search activity at the segment level, exploring how the slope of the housing Beveridge curve varies with the level of aggregation. Section 4 presents a reduced-form model of a single segment that highlights the key economic forces arising from the interaction of broad and narrow searchers in housing markets. Section 5 constructs and estimates a fully-fledged housing search model with many segments that quantifies the importance of these forces. Section 6 uses estimates from this model to infer illiquidity discounts and to explore the response of prices and quantities in different market segments to local housing market shocks. The final section concludes.

2 Understanding Housing Search

We document housing search behavior in the San Francisco Bay Area using email alerts set on the real estate website trulia.com. The San Francisco Bay Area is a major urban agglomeration in Northern California that includes the cities of San Francisco, San Jose, and Oakland. We analyze data from Alameda, Contra Costa, Marin, San Benito, San Francisco, San Mateo, and Santa Clara counties. In the 2010 Census, these counties had a population of about 6 million people living in 2.2 million housing units. This section first describes the data on email alerts and then highlights the key findings from Appendix A, where we provide a detailed analysis of the raw search data.

2.1 Search Data From 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.

Figure 1: Setting Email Alerts on Trulia.com
The second row in the form provides the option of specifying property 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 bound 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 to receive a daily or weekly email. ³

We observe a random subset of 40,525 “For sale” email alerts set between March 2006 and April 2012. Those alerts were set by 23,597 unique home searchers, identified by their (scrambled) email address. 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.

Representativeness

We do not observe demographic information on the home searchers in our sample. Thus, we cannot provide direct evidence that searchers on trulia.com are representative of the overall pool of home searchers. However, recent surveys conducted by the National Association of Realtors (2013) suggest that the internet is the most important tool in the modern home-buying process. In particular, over 90 percent of home buyers rely on the internet in their search. 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.”

Internet use for home search is also not concentrated among younger or richer 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, and their 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). These statistics suggest that we can learn from online home search about overall home search behavior. Moreover, trulia.com, with approximately 24 million unique monthly visitors during our sample period (71 percent of whom report planning to purchase in the next 6 months), has similar demographics to those of the overall online home search audience (Trulia, 2013).

2.2 Dimensions of Housing Search

To compare the geographic dimension of individuals’ search ranges, we express them in terms of the zip codes that are covered by the pooled email alerts. We observe wide heterogeneity in the geographic breadth considered by various home searchers. While 25% of searchers are narrowly interested in a single zip code, among those individuals who select more than one zip code, the

³Trulia also provides a second way for potential home buyers to set an email alert. After looking at results from regular searches on their website, generally along the same dimensions as those in Figure 1, users can press a single button: “Send me an email whenever houses with these characteristics come on the market.”
10-90 percentile range of the maximum geographic distance between zip codes selected by the same searcher is 2.3 miles to 21.1 miles (see Appendix A.2 for further details).

Roughly two-thirds of the email alerts specify fields in addition to geography. The other fields that are specified regularly are listing price (two-thirds of all alerts) and the number of bathrooms (one-third of all alerts). On the price dimension, among those searchers who set both an upper and a lower bound, the 10th percentile selects a price range of $100k, the median a price range of $300k, and the 90th percentile a price range of $1.1m. Among the third of searchers who specify bathrooms, most select a lower bound of 2.

We want to develop a model that captures the heterogeneity of these search ranges. One possible approach to describe their geographic dimension would be to use contiguous and/or circular subsets of a plane. This approach does not work well for the Bay Area with its complicated topology. Moreover, many searchers look for houses in zip codes that are not adjacent to each other (see Appendices A.2.3 and A.2.4). Our approach in the next section is therefore to define a discrete grid of market segments using zip codes as the basic geographic unit, which we further subdivide along quality and size dimensions. A search range can then be represented as a subset of market segments, allowing us to accommodate non-contiguous and non-circular search patterns.

3 Segment-Level Housing Search and Market Activity

We now describe how 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.

The finest partition of the Bay Area housing stock into different market segments that could be motivated by our search data is obtained by the join of all search ranges in our sample. The preferences of each searcher could then be expressed exactly as a subset of these 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 inventory and time on the market.

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. We start with zip codes as the level of geography, and then subdivide zip codes along price and size boundaries that are common in email alerts that cover this zip code (see Appendix B.1 for details on the algorithm). This process leads us to a final set $H$ of 564 segments. These segments contain houses within a zip code that are of similar quality (based on price) and size (based on the number of bathrooms), with cutoffs that are close to cutoffs regularly used in email alerts.

We then express each search range as a subset of these segments. Here we start from the raw search range, specified along the dimensions geography, quality, and size; we ignore the other dimensions that are rarely specified. We then determine which segments are (approximately) covered by the specified range, using the algorithm described in Appendix B.2. This produces a set $\Theta$ of 11,503 distinct search ranges that can each be represented as a subset of $H$. 
3.1 Segment-level Housing Market Activity

To measure housing market activity at the segment level, we combine three main datasets. We start with the universe of ownership-changing deeds in the San Francisco Bay Area between 1994 and April 2012. The property to which a deed relates is identified by the Assessor Parcel Number (APN). From the deeds data, we obtain the property address, transaction date, transaction price, and the type of deed (e.g., Intra-Family Transfer Deed, Warranty Deed). We use the type of deed to identify armslength transactions (see Appendix B.3). We combine the transaction deeds with the universe of tax assessment records for the year 2009. This dataset includes information on property characteristics such as construction year, size, and the number of bedrooms and bathrooms. Finally, we use a data on all property listings on trulia.com between October 2005 and December 2011. The key variables from this data set 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 our analysis, 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. In Appendix C, we show that segment-level market and search activity are quite stable over time within our sample. To make prices comparable across years, we convert all prices to 2010 dollars using zip-code level repeat sales price indices.

Segment-Level Facts: Notation

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. As before, \( H \) denotes the set of all segments. The measure \( \mu^H \) counts houses, so that \( \mu^H (h) \) is the housing stock in segment \( h \). We normalize the total Bay Area housing stock to a unit mass: \( \sum_{h \in H} \mu^H (h) = 1 \).

The average monthly turnover rate \( V (h) \) in segment \( h \) is defined as the number of armslength transactions divided by the total housing stock. The mean time on the market \( T (h) \) in segment \( h \) is defined as months between listing and sales date, less one month for the typical escrow period. Our measure of inventory in segment \( h \) is \( \mu^S (h) := T (h) V (h) \mu^H (h) \).

We also define the inventory share \( I (h) = \mu^S (h) / \mu^H (h) \) as the fraction of the housing stock that is for sale.

---

\(^4\)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 could construct measures of inventory directly from listings data. The resulting series are noisy because of the incomplete coverage of Trulia listing data, and the need to make assumptions on when the few listings that do not sell are removed. We discuss the trade-offs involved in the choice of how to measure inventory in Appendix B.3.
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 \( \Theta \) as the set of searcher “types.” A searcher of type \( \theta \) scans inventory in the set of segments \( \tilde{H}(\theta) \subset H \). The total housing stock that is of interest to searcher \( \theta \) is

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

Similarly, we define the total inventory \( \nu^S(\theta) \) considered by searcher \( \theta \) as the sum over all inventory \( \mu^S \) for sale in segments in \( \theta \)'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,

\[
\tilde{\Theta}(h) = \{ \theta \in \Theta : h \in \tilde{H}(\theta) \}. \tag{1}
\]

The pattern of clienteles reflects the interconnectedness of segments. One polar case is a perfectly segmented market, in which \( |H| \) types have search ranges that each consist of a single segment, and each segment has a homogeneous clientele of one type who searches only that segment. Another polar case is a perfectly integrated market, where a single type searches all segments 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 \( \theta \) in the data, so that \( \sum_{\theta \in \Theta} \beta(\theta) = 1 \). The distribution of searchers interested in segment \( h \) is then obtained 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 (relative) number of searchers per house in that segment:

\[
\sigma(h) = \frac{1}{\mu^H(h)} \sum_{\theta \in \tilde{\Theta}(h)} \beta(\theta) \frac{\mu^H(h)}{\nu^H(\theta)} = \sum_{\theta \in \tilde{\Theta}(h)} \frac{\beta(\theta)}{\nu^H(\theta)}. \tag{2}
\]

Weighting the contribution of each searcher type \( \theta \) to \( \sigma(h) \) by \( \mu^H(h)/\nu^H(\theta) \) makes broader searchers (who are interested in more housing stock \( \nu^H(\theta) \)) count less towards search activity in \( h \). If every searcher was looking only at one segment, then \( \sigma(h) \) simply reflects the relative number of searchers per house in \( h \), since in that case \( \nu^H(h) = \mu^H(h) \).

So far, all summary statistics have been defined at the segment level. We are also interested in how market and search activity vary at different levels of aggregation. Since \( V \), \( I \) and \( \sigma \) are all defined as ratios relative to housing stock, aggregation uses the 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

\[
\frac{\sum_{h \in G} \mu^H(h) V(h)}{\sum_{h \in G} \mu^H(h)}.
\]
Table 1 presents summary statistics on market and search activity for the San Francisco Bay Area as a whole, as well as by segment. The housing market during our sample period is relatively 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 \( \frac{100}{(0.24 \times 12)} = 35 \) years. The cross-sectional variation in market activity at the segment level is substantial. For example, the 75th percentile of inventory share is 1.5 percent, over twice as much as the 0.6 percent inventory share at the 25th percentile.

<table>
<thead>
<tr>
<th></th>
<th>Inventory Share</th>
<th>Turnover Rate</th>
<th>Search Activity</th>
<th>Mean Price</th>
<th>Housing Stock</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bay Area</td>
<td>1.14</td>
<td>0.24</td>
<td>.</td>
<td>649</td>
<td>2,216,021</td>
</tr>
<tr>
<td>Min</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>72</td>
<td>1,221</td>
</tr>
<tr>
<td>Q25</td>
<td>0.61</td>
<td>0.16</td>
<td>0.51</td>
<td>306</td>
<td>2,333</td>
</tr>
<tr>
<td>Q50</td>
<td>0.94</td>
<td>0.21</td>
<td>0.81</td>
<td>518</td>
<td>3,298</td>
</tr>
<tr>
<td>Q75</td>
<td>1.51</td>
<td>0.29</td>
<td>1.31</td>
<td>790</td>
<td>4,858</td>
</tr>
<tr>
<td>Max</td>
<td>6.66</td>
<td>0.97</td>
<td>7.05</td>
<td>2,491</td>
<td>13,167</td>
</tr>
</tbody>
</table>

Note: Table shows summary statistics for market and search activity, both for the entire San Francisco Bay Area as well as across the 564 housing market segments. We consider inventory share \( I \), volume share \( V \), search activity \( \sigma \) as defined in equation (2), mean price, and total housing stock.

The distribution of \( \sigma(h) \) in equation (2) as a measure of search activity is positively skewed across segments (it is not well-defined for the Bay Area as a whole): 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 only considered by a small number of broad searchers. Other 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.

Table 2 reports cross-sectional correlations of observables for “submarkets” at different levels of aggregation. These submarkets are the 564 segments, 191 zip codes, and 96 cities in our data. The left panels show volatilities and correlation coefficients across submarkets. The right panels consider segment-level variation within submarkets.

Comparison of volatilities shows substantial variation across segments which is below the zip code level. Indeed, the zip code-level movements account for only 46%, 44%, and 64% of the across-segment variance in inventory share \( I \), turnover rate \( V \), and search activity \( \sigma \), respectively.

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 when analyzed across zip codes, and even more negative when analyzed across cities. In addition, more expensive zip codes
Table 2: Cross-Sectional Variation in Market and Search Activity

<table>
<thead>
<tr>
<th>Variation Across Submarkets</th>
<th>Avg. Variation Within Submarkets</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>City</td>
</tr>
<tr>
<td></td>
<td>$I$</td>
</tr>
<tr>
<td>CITY</td>
<td></td>
</tr>
<tr>
<td>vol</td>
<td>0.42</td>
</tr>
<tr>
<td>corr</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>zip code</td>
<td></td>
</tr>
<tr>
<td>vol</td>
<td>0.50</td>
</tr>
<tr>
<td>corr</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>segment</td>
<td></td>
</tr>
<tr>
<td>vol</td>
<td>0.74</td>
</tr>
<tr>
<td>corr</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

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 these units, the right column presents statistics across segments within these units. We consider inventory share, volume share, search activity $\sigma$ and $\log(p)$. We present both volatilities (standard deviations) as well as correlations within and across units.

and cities are searched more. On the other hand, search activity moves together with market activity and against price across segments within zip codes or cities.

The relationship between inventory and measures of search activity is reminiscent of the “Beveridge curve” in studies of the labor market. The canonical Beveridge curve is a relationship between vacant job positions and unemployed workers searching for jobs, usually presented in the time series. Here we have a relationship between vacant homes and individuals searching for homes, but presented in the cross section. The stylized fact is that the housing market Beveridge curve is downward sloping across broad units of aggregation such as cities, while it is on average upward sloping within broad units. Panel A in Figure 2 shows this relationship across cities, and Panel C across all segments within the city of San Francisco. The Beveridge curve is upward sloping in 64 out of the 74 cities with at least two segments, representing 84% of the total Bay Area housing stock. 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.

In contrast, 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. For market activity indicators and prices, the nature of covariation
Figure 2: 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 $\sigma$, and turnover rate $V$, both 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 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).
across and within submarkets is very similar, both qualitatively and quantitatively: inventory share and turnover rate move together and are both negatively correlated with price. To illustrate this, Panel B in Figure 2 shows the positive correlation between inventory and turnover across cities in the San Francisco Bay Area. Panel D shows the positive correlation between inventory and turnover across segments within the city of San Francisco.

**Segment-Level Facts: Search-Breadth**

The measure \( \sigma(h) \) reflects the average search activity in a segment, but it does not tell us whether that activity is due to narrow searchers or due to broader searchers who provide connection to other segments. Indeed, the same \( \sigma(h) \) could arise when there are a few narrow searchers who fully target their search effort to a given segment, or relatively more broad searchers whose search effort in a given segment is diluted because they also consider other segments. To summarize interconnectedness, we now compare segments in terms of the inventory scanned by their typical client. The left panel of Figure 3 plots the share of inventory in segment \( h \) in total Bay Area inventory (on the horizontal axis) against the share of Bay Area inventory scanned by the median client of segment \( h \) (on the vertical axis). Every dot represents a segment, and colors reflect the value on the vertical axis so that segments can be recognized in the map in the right panel.

If the Bay Area were perfectly segmented, then any given segment would only be searched by individuals who scan that particular segment. As a result, all points would line up along the 45-degree line (which is the straight line in the figure). 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 (located north of our current figure). 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 the Bay Area total.

Appendix B.4 provides detailed summary statistics on this measure of search breadth. In the average segment, the median searcher scans 2.1 percent of the total Bay Area inventory in that segment; the across-segment inter-quartile range of the inventory scanned by the median searcher is between 0.94% of Bay Area inventory and 2.53% of Bay Area inventory. Remarkably, the average within-segment inter-quartile range of inventory scanned by different searchers looking in the same segment is nearly as large, at 1.75% of the Bay Area housing stock.

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.
3.2 Search Intensity within Search Range

How do searchers of type \( \theta \) choose among the inventory within their search range \( \tilde{H}(\theta) \)? Are searchers equally interested in all properties or do they prefer properties in some segments to those in others? To address this question, we exploit data on property views by individual home searchers on trulia.com. After defining a search range on trulia.com, a user is presented with a list of properties that are included in the search range (see Appendix Figure A.6). This list provides basic information on each property, such as a picture, its location, the listing price, and the first lines of a description of the property. Home searchers then actively click on those houses that attract their particular interest to view additional property information.

We have obtained these individual property-view data from trulia.com for a random subset of users visiting the site in April 2012. These data contain the set of listings viewed within a “session,” defined as all views by the same user (represented by their IP address) within one day. We interpret viewing a property’s listing details as an expression of particular interest in
that property. In Appendix A.5, we analyze how this signal of particular interest is distributed across listings in the various segments searched by an individual. We find that the rate at which particular interest is expressed for properties in different segments is directly related to the share of total inventory made up by those segments in the individuals’ overall search ranges. These findings suggest that, conditional on the search range, the probability of finding a favorite house in any one particular segment is proportional to the inventory in that segment. This observation motivates one of our key modeling assumptions in the next section.

4 A Stylized Model of a Single Segment

In this section, we describe a reduced-from model of a single segment. We analyze comparative statics to explore how the cross section of observables introduced in Section 3.1 – turnover, inventory and search intensity – is shaped by supply and demand forces. In particular, the interaction of broad and narrow searchers within a city provides an explanation for the facts in Figure 2 – that the cross section of segments within a city looks very different from the cross section of cities.

The equations of the reduced-form model describe flows of buyers and houses in a particular segment. They hold in steady state equilibrium for a large class of continuous-time search models. We specify and estimate one such model in Section 5. The purpose of the discussion below is to illustrate the key mechanisms independently of model details such as price formation and how broad searchers select houses within their search range.

Model Setup and Intuition

Consider a segment with a unit mass of houses. If an inventory of \( I \) houses is for sale and \( B \) agents are looking to buy, transactions occur at the turnover rate \( V = m(B, I) \), where the matching function \( m \) is increasing in both of its arguments and exhibits constant returns to scale. Agents own at most one house. The \((1-I)\) homeowners who are not already sellers, receive moving shocks and put their house on the market at a constant rate \( \eta \). They stay in their house until they sell it, and then look for a new house. In steady state, the rate at which houses come on the market is constant and equals the rate at which houses are sold:

\[
\eta (1 - I) = m(B, I) = V. \tag{3}
\]

The potential clientele of the segment under study consists of two types of agents. There are \( N > 1 \) “narrow types” who only like that segment. Of these narrow types, \( B^N \leq N \) are currently looking to buy in that segment, and the remainder occupy properties in this segment. If all agents are narrow, then the total number of buyers is \( B = B^N = N - 1 \): the only buyers are narrow types who do not currently occupy a house, and all houses are owned by narrow types. We use this special case as a reasonable approximation to think about the cross section of cities.

To compare across segments within a city, we also allow for “broad types” who are interested in
all segments in the city. However, at any instant, broad types can try to buy in only one segment. The number of broad types who look to buy in the segment under study is determined by a nonnegative and increasing function of inventory, $B(I)$. This function captures the sensitivity of broad searchers to local conditions: a segment with higher inventory attracts more broad types wanting to buy there, consistent with the evidence in Section 3.2. The idea is that, all else equal, the more houses are available for sale in a segment, the more likely it is that a broad searcher will find her favorite house in that segment. When we allow for broad searchers, the total number of searchers looking to buy in a segment is therefore given by $B = B^N + B^B(I)$.5

In this simple model, our measure of search activity $\sigma$ maps directly into the number of narrow types looking to buy, $B^N$. To see this, recall that $\sigma$ measures the (weighted) number of home searchers that include a segment in their search range. These searchers could either be broad or narrow types. With only narrow types, the number of searchers who include a given segment in their search range is equal to the number of narrow types in that segment who are looking for a house: $\sigma = B^N$. Since any broad searchers include all segments in their search range, they have the same direct impact on $\sigma$ across all segments. Any differences in $\sigma$ across segments will thus still be due to differences in the number of narrow searchers, $B^N$.

This does not mean that broad searchers do not influence equilibrium search activity $\sigma$ differentially across segments, only that the effect on $\sigma$ of broad searchers can be measured by their effect on $B^N$. Indeed, the relationship between broad searchers and $B^N$ (and therefore $\sigma$) is central to understanding how the interaction of broad and narrow searchers can generate upward-sloping Beveridge curves across segments in a city. Here is the intuition. All narrow types either own or search for a house in their preferred segment. When there are no broad types, narrow types own the entire housing stock, and the remaining $B^N = N - 1$ narrow types are searchers. Now we introduce broad types. These broad types will end up owning some of the local housing stock, leaving fewer units occupied by narrow types. This increases the number of narrow types that search in equilibrium, $B^N$, which increases $\sigma$. The strength of this crowding out depends on how much of the local housing stock is occupied by broad types. Since broad types are more likely to buy a house in high-inventory segments, higher inventory is associated with more crowded-out narrow types, and therefore higher $\sigma$. As we next demonstrate more formally, this is the key force through which broad searchers can contribute to an upward-sloping Beveridge curve where markets with higher inventory have more search activity.

5In our simple model here, the function $B^B(.)$ does not depend directly on $\eta$, but only responds to inventory. Our quantitative model below has the same feature – it assumes that broad searchers flow to segments in their search range in proportion to inventory. In principle, it is also possible to allow $B^B(.)$ to depend on $\eta$: a segment may be less attractive to broad searchers if houses there fall out of favor faster. For example, a common class of models in urban economics assumes that broad searchers are indifferent in equilibrium between buying in the segment under study and receiving fixed utility elsewhere. This type of argument implies a function $B^B(I, \eta) = q(\eta) I$, where $q$ is the tightness of the market – the ratio of buyers to sellers. Tightness is decreasing in $\eta$: there are fewer buyers per seller if homeownership yields utility for a shorter amount of time. Appendix XXX considers what happens if market tightness depends on $\eta$, and shows that under plausible conditions on parameters all results are robust.
Model Steady State

In steady state, the distribution of agents who are active buyers in the segment is constant, and broad and narrow types cycle in and out of homeownership at the same rates. A constant distribution of types implies two key relationships between segment inventory and search activity.

First, to ensure that the buyer and seller shares among broad and narrow types do not change over time, equilibrium flows must be consistent with narrow types only trading with narrow types, and broad types only trading with broad types. The first relationship thus describes local market conditions, ensuring that the number of houses put up for sale by narrow types equals the match rate between narrow buyers and sellers:

\[ \eta (N - B^N - I (N - B^N)) = m (B^N, I (N - B^N)) \]  

(4)

The term \( I(N - B^N) \) measures narrow owners who are currently listing houses for sale (or the total inventory listed by narrow types). The term in parentheses on the left hand side thus captures the measure of narrow homeowners that are not yet selling their house; those individuals put their house on the market at rate \( \eta \). The right hand side captures the rate at which narrow buyers \( B^N \) purchase inventory offered by narrow sellers. The implied relationship between \( B^N \) and \( I \) is negative, as a higher inventory share reduces the number of narrow searchers who can now find properties more quickly. This relationship is independent of the presence of broad searchers.

The second relationship ensures that in equilibrium, the overall buyer-owner ratio in the segment, \( B/1 \), is the same as the buyer-owner ratio for narrow types only. This relationship allows us to capture the interaction of narrow and broad types:

\[ \frac{B^N}{N - B^N} = B = B^N + B^B(I) \]  

(5)

In the presence of broad searchers, the implied relationship between \( B^N \) and \( I \) is positive: higher inventory \( I \) attracts more broad searchers who crowd out ownership from narrow searchers. This increases the equilibrium number of narrow searchers \( B^N \).

Equations (4) and (5) jointly determine \( I \) and \( B^N \), and then \( V \) can be read off from (3). The model thus speaks to the three key observables discussed in Section 3.1, inventory share \( I \), turnover rate \( V \), and the number of searchers per house \( \sigma \), which moves with the number of narrow searchers, \( B^N \). It relates those endogenous variables to segment-specific supply conditions (\( \eta \)) and segment-specific demand conditions (\( B^B(\cdot) \) and \( N \)). The model thus provides a simple tool to understand the cross section of cities and segments in terms of supply and demand forces.

Graphical Analysis

To explore the forces that drive the cross-sectional patterns in the data, the four panels of Figure 4 provide counterparts to the scatter plots in Figure 2. The top row studies the cross section of cities and hence assumes that there are no broad searchers. The bottom row considers
the cross section of segments within a city in the presence of broad searchers. In both rows, the right panel displays the flow equilibrium relationship (3) between inventory and turnover. The left panel illustrates the two forces behind the Beveridge curve captured by equations (4) and (5); on the horizontal axis we plot $\sigma$, but as discussed above, in this simple model this is directly determined by the number of narrow searchers, $B^N$. The solid curves describe an initial equilibrium, whereas the dashed and dotted curves represent comparative statics.

Figure 4: Inventory, Search, and Volume in Stylized Model

(A) Inventory vs. weighted searchers across cities

(B) Inventory vs. turnover across cities

(C) Inventory vs. weighted searchers within city

(D) Inventory and turnover within city

Note: Figure shows inventory $I$, weighted number of searchers $\sigma$, and turnover rate $V$ across and within cities for the stylized model of one segment. In the left panel, the solid downward sloping curve describes local conditions (4). The other solid curve captures the interaction (5) between narrow and broad searchers. The right panels describe the flow equilibrium relationship (3) between inventory and turnover.

In the left panels, the downward sloping solid curve represents the local conditions (4) for narrow types. This relationship is the same in both rows: more inventory leaves fewer narrow searchers in steady state, thereby reducing $\sigma$. The other solid curve describes the interaction (5) between narrow and broad searchers. Without broad searchers, it is vertical: the number of narrow searchers $B^N = N - 1 = \sigma$ is exogenous and pins down inventory. When there are
broad searchers (bottom row), equation (5) generates an upward sloping curve: more inventory attracts more broad searchers who crowd out narrow searchers. This increases the overall number of narrow searchers, and thus $\sigma$. The intersection of the curves determines equilibrium inventory and search activity. Turnover can then be read off the solid curve in the right panel.

The Cross Section of Cities

What forces shape the cross section of cities? In the data, cities with lower inventory have more searchers and lower turnover. In the absence of broad types, these patterns require the co-movement of two forces. First, some cities must be more “popular,” that is, more (narrow) agents $N$ are interested in living there. This is essential in order to generate the observed variation in the number of searchers, $\sigma = B^N = N - 1$, and will involve a rightward shift in the solid green line. Higher $N$ also changes local conditions – it becomes harder to find a house at given inventory. In the top left panel of Figure 4, both curves thus shift to the right; equilibrium in a city with higher $N$ is at the intersection of the dashed lines, with lower inventory and search activity.

Variation in popularity – measured by $N$ – is thus consistent with cities lining up along a downward sloping Beveridge curve. However, it is not, by itself, consistent with the positive co-movement of inventory and turnover that we see in the data. Indeed, the relationship (3) in the right panel is not affected by $N$. Reading across from the new equilibrium in the left panel we thus obtain that more popular cities see lower inventory and more turnover. Intuitively, a larger pool of agents looking for houses not only reduces inventory but also allows for faster matching.

To account for the joint behavior of inventory and turnover in the data, more popular cities must also be “more stable,” that is, houses come on the market at a slower rate $\eta$. The dotted lines in the figure show the effect of a decrease in $\eta$. If agents become unhappy at a slower rate, then for given inventory we have a smaller pool of searchers (the left panel) and lower turnover (the right panel). The increase in stability thus contributes to the downward sloping Beveridge curve, while helping to explain the positive co-movement of inventory and volume.

The Cross Section of Segments within a City

Consider now how inventory, turnover, and search vary within cities. Section 3.1 documented that within most cities the Beveridge curve is upward sloping: segments with more inventory have more searchers per house (see Figure 2). They also have more turnover. Our setup explains this pattern with the response of broad searchers to variation in segment stability. The downward-sloping dotted line in the bottom left panel in Figure 4 is the same as in the top left panel; it shows local conditions in a more stable segment (lower $\eta$). The key difference is that a more stable segment attracts fewer broad searchers because there is less inventory on offer. As a result, fewer more types can own a house at any point in time, and fewer of them are crowded out, leaving fewer narrow searchers. The more stable segment thus has both less inventory and fewer searchers per house, generating an upward sloping Beveridge curve. The bottom right panel in Figure 4 shows that the more stable segment also has less turnover, consistent with the patterns in the data.
Inference, Housing Policy, and the Importance of Search Data

We now illustrate the importance of search data for inference and for assessing the effect of local housing policies. Consider a researcher or policy maker who is trying to predict the effect of building houses in some segment on inventory and the pool of homeowners there. In our setup, the only parameter that reflects the scale of the segment is the number of narrow searchers \( N \). Building houses is thus equivalent to reducing the number of narrow searchers. We study the effect of this comparative static for different calibration strategies the researcher might adopt.

Without direct search data, a common strategy to detect market segments with housing shortages is to use moments on turnover \( V \), inventory \( I \) (or perhaps months supply \( I/V \)) as well as the average time it takes to find a house, which in steady state equals \( B/V \) (e.g., Forbes, 2016; USA Today, 2016). From the flow equilibrium condition (3), this information identifies \( \eta = V/(1-I) \) and also puts restrictions on the matching function.\(^6\) However, even if the matching function is known, there is not enough information to determine the shape of the function \( B(B,I) \). Researchers without access to search data thus need to add judgment calls about the population of searchers who might be buyers in a segment if there were move inventory.

We compare two extreme assumptions that are familiar from the literature. First, if one assumes that all searchers are narrow, the number of narrow types is identified as \( N = (B/V)V + 1 \). Intuitively, if a segment is not connected to other segments, then information on time-to-buy suffices to assess demand. Alternatively, one could assume that all searchers are broad and the number of buyers is proportional to inventory: \( B(B,I) = qI \). In this case, the slope coefficient is identified as \( q = (B/V)V/I \).

The effect of building houses (reducing \( N \)) is quite different in the two cases. With only narrow searchers, the effect of reducing \( N \) can be read off the top row of Figure 4. Adding houses directly reduces demand and hence the number of searchers (shifting the vertical line to the left) while inventory increases. In contrast, if all searchers are broad, then the inventory share and turnover rate do not change. New houses are bought by broad searchers without affecting local conditions.\(^7\) This comparison illustrates the pitfalls of focusing on inventory to guide housing policy. It is tempting to view low inventory in some segment as a “housing shortage” that makes building there particularly important. With narrow searchers and a downward-sloping Beveridge curve, low inventory indeed flags a large number of agents who desire local building. With broad searchers and an upward-sloping Beveridge curve, however, low inventory may simply reflect that few houses come on the market. There is no obvious reason to build in low-inventory segments since broad searchers have no particular preference for those segments.

---

\(^6\)Full recovery of the matching function requires additional assumptions. For example, if \( m \) does not vary across segments, then it can be traced out from the cross section of \( V,I \), and \( B/V \). Alternatively, with a specific functional form such as \( m = \bar{m}\sqrt{BI} \), segment specific shifters \( \bar{m} \) can be recovered as \( \bar{m} = \sqrt{VIT(B/V)} \).

\(^7\)Formally, with only broad searchers, inventory is determined by \( \eta (1-I) = m (B(B,I),I) \). Since the segment under study is small relative to the rest of the economy, changes in its scale do not affect the inventory share, but only the number of broad searchers who enter.
Direct data on the pool of searchers helps distinguish the two cases and hence can improve inference and policy predictions. Indeed, a sample of email alerts would reveal the relative share of narrow searchers $B^N/B$. Based on local conditions (4), this additional information identifies $N$ and hence puts restrictions on the function $B^B(I)$.$^8$ In the absence of such data, the quantitative results below suggest that a model with only narrow (broad) searchers will provide a first approximation of the searcher pool across (within) cities.

5 A Quantitative Model with Multiple Segments

In this section we quantify the effects discussed in the previous section. To this end, we specify a full-fledged search model with multiple segments. At the segment level, equilibrium actions in this model imply the same flow equations as in the simple model from Section 4, for particular buyer flow functions $B^B(I)$ that accommodate the rich clientele structure in our data. The model here also makes explicit how transactions occur and how prices are determined.

Segments and Preferences

We continue the notation used to present the facts in Section 3.1. There are $H$ market segments with mass $\mu^H(h)$ houses in segment $h$, and mass one houses in the entire Bay Area. Agent type $\theta$ is identified by search range $\tilde{H}(\theta) \subset H$. Search ranges are part of preferences, and type $\theta$ can never enjoy a house outside of $\tilde{H}(\theta)$. We use a measure $\mu^\Theta$ on the set of types $\Theta$ to count the number of agents of each type. The total number of agents is $\mu^\Theta = \sum_{\theta \in \Theta} \mu^\Theta(\theta) > 1$. The clientele $\tilde{\Theta}(h)$ of segment $h$ is the set of all agents who are interested in segment $h$, as defined in (1). The inventory and housing stock scanned by type $\theta$ are $v^S(\theta)$ and $\nu^H(\theta)$, respectively.

Agents live forever, discount the future at rate $r$, and receive quasilinear utility from a numeraire good and housing services. Agents only obtain housing services from a “favorite” house. After an agent moves into his favorite house in segment $h$, he obtains housing services $v(h) > 0$. Houses fall out of favor at rate $\eta(h)$ and then no longer provide housing services. Agents can put a house up for sale at no cost. Once it is sold, he can search for a new house, again at no cost. Sellers incur a proportional cost $c$ upon sale of a house. Matching in segment $h$ occurs at the rate $\tilde{m}(B(h), I(h), h)$, where $\tilde{m}$ exhibits constant returns to scale in its first two arguments.$^9$

The specification of how agents decide on a favorite house within their search range $\tilde{H}(\theta)$ is guided by the evidence in Section 3.2: interest in individual segments within a search range is proportional to segment inventory. We thus assume that agents are equally likely to “fall in love”

---

$^8$For example, if $B^B(I)$ is the same for all segments, then it can be traced out from the cross section of $I, V, V/B$ and $B^N/B$. In fact, in this case the supply and demand parameters $\eta, B^B(I)$ and $N$ can be identified without taking a stand on the shape of the matching function.

$^9$The matching function is increasing in the number of buyers and sellers and satisfies $\tilde{m}(0, \mu^S, h) = \tilde{m}(\mu^B, 0, h) = 0$. It is also 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.
with any house for sale in $\tilde{H}(\theta)$. Formally, the number of buyers per house in segment $h$ is:

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

(6)

where, as before, $\beta(\theta)$ is the share of type-$\theta$ buyers in the data. To get the total number of type-$\theta$ buyers, we thus multiply by the total number of searchers of any type, $(\bar{\mu}^\theta - 1)$. For a narrow type $\theta$ who considers only segment $h$, we have $I(h) = \nu^S(\theta)$, so this buyer contributes one-for-one to $B(h)$. To determine the contribution to $B(h)$ of a broad type $\theta$, we multiply the number of these buyers $\beta(\theta) (\bar{\mu}^\theta - 1)$ by $I(h) / \nu^S(\theta)$, which represents the share of total inventory scanned by type $\theta$ that is in segment $h$.

The expression (6) is the counterpart to equation (5) in the simple one-segment reduced-form model from Section 4. In that model, the spirit of our buyer-flow assumption would be captured by the functional form $B^B(I) = \kappa I$ for some exogenous constant $\kappa$. In the current model, each segment differs in their clientele structure. As a result, there are many segment-specific measures of the sensitivity of a segment’s buyer pool to segment inventory, $\sum_{\theta \in \tilde{\Theta}(h)} \beta(\theta) (\bar{\mu}^\theta - 1) / \nu^S(\theta)$, that are all jointly determined in equilibrium.

Matching, Bargaining, and Equilibrium

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 to a real estate agent. The seller then starts to search, whereas the buyer moves into the house and begins to receive utility $v(h)$.

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. 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 no longer receive housing services from it, and (ii) all offers are accepted.

The model endogenously determines inventory shares $I(h)$ and turnover rates $V(h)$ for each segment. It also determines the share of searchers $\beta(\theta)$ for each type, which then allow us to calculate segment-level relative search activity as expressed by equation (2). The cross section of these observables is shaped by three distinct forces. Supply is represented by the rate $\eta(h)$ at which houses fall out of favor. Demand is captured by the clientele structure, that is, the distribution of searcher types $\mu^\theta(\theta)$. Finally, the segment-specific effect on match rates $\hat{m}(.,.,h)$ represents differences in market frictions across segments. Indeed, in the next section we discuss how information on observable endogenous variables $I(h), V(h)$, and $\beta(\theta)$ allows us to quantify of these forces.
5.1 Housing Demand and the Cross Section of Housing Markets

Our quantitative analysis proceeds in three steps. In this subsection, we show how the cross section of turnover, inventory, and search intensity in the Bay Area reflects housing demand – in particular, the presence of broad searchers – as well as the other two exogenous forces capturing supply and market frictions. We also provide further evidence that validates and helps interpret the demand estimates. Section 6 then studies price formation and conducts counterfactuals.

Identification

The analysis in this section requires only the supply and demand parameters; it does not depend on the details of bargaining or the shape of the matching function. The intuition is as in Section 4. With a fixed number of houses and agents, the steady state distribution of agent states (searching for a house, listing one for sale, or owning without listing) is pinned down by house and agent flows, regardless of pricing. Moreover, matching frictions matter for agent flows only via the rates at which buyers find houses in a given segment, defined as \( \alpha(h) = V(h) / B(h) \).

More formally, our identification result in Appendix D establishes a one-to-one mapping between two sets of numbers. The first set consists of the supply and demand parameters \( \eta(h) \) and \( \mu^\Theta(\theta) \), as well as the vector of house finding rates \( \alpha(h) \). The second set consists of objects we observe in the data: 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.\(^{10}\) We use this mapping to back out the vectors of \( \eta(h) \), \( \mu^\Theta(\theta) \), and \( \alpha(h) \).

We refer to the supply parameter \( \eta(h) \) as instability: in a more unstable segment, houses come on the market at a faster rate. We also define popularity as a summary measure of housing demand at the segment level:

\[
\pi(h) := \frac{1}{\mu^H(h)} \sum_{\theta \in \Theta(h)} \mu^\Theta(\theta) \frac{\mu^H(h)}{\nu^H(\theta)}. \tag{7}
\]

A more popular segment is in higher demand as more agents would include any given house in their search ranges; unlike for the definition of \( \sigma(h) \) in equation (2), where we weighted different types of agents by their share among searchers, here we weight different types of agents by their share in the overall population.\(^{11}\) We continue to count individuals who would like many segments less towards the popularity of segment \( h \) than individuals who only like that particular segment.

\(^{10}\)While the equilibrium depends on the total number of buyers of each type, \( \beta(\theta) (\bar{\mu}^\Theta - 1) \), we only observe this up to a constant: the email alert data allow us to infer the relative number of each type, \( \bar{\beta}(\theta) \), but we do not have information on the overall number of buyers, \( \bar{\mu}^\Theta - 1 \). As an additional target moment, we thus 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.

\(^{11}\)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, and depends on the equilibrium share of agents of each type that are searching at any point in time.
Summary Statistics

Table 3 summarizes the estimates of instability \( \eta(h) \), popularity \( \pi(h) \), and the house finding rate \( \alpha(h) \). Panel A provides information on the distribution of the estimates across segments. Panel B reports correlations both among the estimates themselves as well as between estimates 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 3: Quantitative Results from Estimating the Model

**Panel A: Estimated parameters vs. hypothetical segmentation**

<table>
<thead>
<tr>
<th></th>
<th>( 100 \times \eta(h) )</th>
<th>( \pi(h) )</th>
<th>( \alpha(h) )</th>
<th>( \hat{\sigma}(h) - \sigma(h) )</th>
<th>( \hat{\pi}(h) - \pi(h) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.24</td>
<td>1.00</td>
<td>0.19</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>Q25</td>
<td>0.16</td>
<td>0.80</td>
<td>0.07</td>
<td>-0.17</td>
<td>-0.13</td>
</tr>
<tr>
<td>Q50</td>
<td>0.21</td>
<td>1.00</td>
<td>0.13</td>
<td>-0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>Q75</td>
<td>0.29</td>
<td>1.15</td>
<td>0.23</td>
<td>0.12</td>
<td>0.21</td>
</tr>
</tbody>
</table>

**Panel B: Correlation between estimated parameters and data moments**

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

**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.

Instability \( \eta(h) \) tracks turnover almost exactly: its moments in Table 3 are essentially the same as those reported for the turnover rate in Table 1. The result follows from the flow equilibrium
condition for each segment, equation (3), together with the summary statistics in Table 1. 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.

Popularity at the segment level, $\pi(h)$, ranges between 0.2 and 2.4, with an inter-quartile range of 0.80 to 1.15. The fact that popularity is below one for many segments is indicative of the importance of partial integration. Indeed, if segments were either perfectly segmented or perfectly integrated, then the number of (weighted) people interested in each segment would be larger than the number of houses, and popularity would have to be above one.\textsuperscript{12} Popularity also comoves strongly with search activity at all levels of aggregation.

The Cross Section of Cities

Panel B shows how the effects described in Figure 4 quantitatively account for market activity in the cross section of cities. Instability generates comovement of inventory and turnover – in fact, it is almost perfectly correlated with both. At the same time, more popular segments see more searchers per house and lower inventory. Popularity is thus a key force behind the downward sloping Beveridge curve across cities. Its positive impact on turnover is muted by the fact that it is negatively correlated with $\eta$ in the cross section of cities: more popular cities are also more stable, generating the positive comovement of inventory and volume. In contrast, there is no relationship between instability and popularity at the segment level.

In addition to the supply and demand effects that were already present in the simple model in Section 4, the quantitative model also identifies significant differences in matching frictions across cities. Indeed, we find that in cities with low inventory and turnover but many searchers, matching is particularly slow (that is, $\alpha(h)$ is low).\textsuperscript{13} Intuitively, cities with many searchers must be popular. Slow matching explains why this popularity does not translate into high turnover. This effect is also present at the segment level, although to a lesser extent.

The Quantitative Importance of Partial Integration

To develop a summary statistic for the quantitative effect of broad searchers, we compare the actual economy – which exhibits partial integration – with a hypothetical perfectly-segmented benchmark economy. We construct the latter by changing demand parameters so as to remove integration, holding all other parameters fixed. At the same time, we require that the benchmark economy still match the same observed inventory shares and turnover rates as the actual economy. This pins down a unique vector of hypothetical demand parameters $\hat{\mu}\Theta$.

\textsuperscript{12}To see how partial integration can imply $\pi(h) < 1$, consider the following 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.

\textsuperscript{13}To illustrate how the matching technology affects $\alpha(h)$, consider a Cobb-Douglas matching function $m(h) = \bar{m}(h) \mu^B(h)^\delta \mu^S(h)^{1-\delta}$, which implies $\log \bar{m}(h) = \delta \log \alpha(h) + (1 - \delta) \log (V(h)/I(h))$.\textsuperscript{26}
The benchmark economy can be viewed as the model estimated by an econometrician who observes $I(h)$, $V(h)$, and buyer match rates by segment, but who does not have information on integration and proceeds by assuming that the economy is perfectly segmented. This misspecification will lead the econometrician to infer the wrong demand parameters, and hence incorrect measures of popularity $\hat{\pi}(h)$. We thus use the distribution of the difference $\hat{\pi}(h) - \pi(h)$ to assess the error from ignoring partial integration.\(^{14}\)

Table 3 shows that the error incurred by an econometrician who ignores partial integration is large. Quantiles for the difference in popularities $\hat{\pi}(h) - \pi(h)$ are reported in the rightmost column in Panel A. The interquartile range of this difference is of the same order of magnitude as the estimated parameter $\pi(h)$ itself, reported on the top left. The reason is that the range of estimated 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.

Intuitively, the econometrician infers too little dispersion in popularity across segments because he ignores the endogenous response of broad searchers to market conditions. A model with narrow searchers must explain observed market and search activity entirely via narrow searchers’ demand, as opposed to broad searchers chasing inventory. In the actual economy, some unpopular and unstable segments attract very few narrow searchers, but nevertheless see a lot of market activity due to attention from broad searchers who are drawn to the segment’s high inventory. Ignoring partial integration therefore overstates the popularity of such high-inventory segments. As discussed in Section 4, overestimating the popularity of a segment generates misleading predictions on the quantity-effects of additional construction, changes in zoning regulation, and other place-based policies. We discuss below how it will also bias estimates of the effects of such policies on transaction prices.

**Broad Searchers and the Beveridge Curve within Cities**

We now quantify the contribution of broad searchers to the Beveridge curve within cities. The perfectly segmented benchmark economy predicts a different cross section of search activity, $\hat{\sigma}(h)$, and hence a different Beveridge curve. We can therefore use the difference in search activity $\hat{\sigma}(h) - \sigma(h)$ to summarize the effect of partial integration on the Beveridge curve. Indeed, if integration were not important, then the actual and hypothetical Beveridge curves would coincide.

The key difference between the benchmark and actual Beveridge curves is that the slope of the latter reflects the response of broad searchers to inventory. In high inventory segments, broad searchers crowd out narrow searchers which increases the overall number of searcher per house. As a result, the actual Beveridge curve is steeper than the benchmark curve in $(\sigma,I)$-plane and the difference $\hat{\sigma}(h) - \sigma(h)$ should be positively correlated with inventory in integrated areas.

\(^{14}\)The other parameters are not materially affected by misspecification. Indeed, we have seen that the parameter $\eta(h)$ closely tracks turnover – this result follows from the flow equations regardless of the clientele pattern. Moreover, given the equilibrium conditions and the fact that the matching function remains unchanged, the benchmark economy predicts the same number of buyers and buyer match rates $\alpha(h)$ as the actual economy.
Table 3 shows that the contribution of broad searchers to the slope of the Beveridge within San Francisco is large. A correlation coefficient of 0.67 between $I(h)$ and $\hat{\sigma}(h) - \sigma(h)$ means that a one standard-deviation increase in inventory increases $\hat{\sigma}(h) - \sigma(h)$ by 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. In contrast, we observe a smaller effect across cities or segments. For example, across cities, the absence of broad searchers adds only about 10 percent worth of search activity per unit of inventory. This finding is consistent with the results in Section 3.1 that showed less integration across cities.

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 these segments are part of an integrated area, then broad searchers are attracted to the higher inventory and crowd out narrow searchers, generating an upward sloping Beveridge curve.

5.2 Further Evidence on Housing Demand

Our estimation infers housing demand from data on market and search activity. To interpret the estimates, we now relate popularity – our segment level summary statistic of demand – to observable characteristics of segments. We also relate search breadth – our key individual level statistic – to demographics. Finally we compare model-implied household flows to data on moves.

*Popularity and Geography*

What makes a segment popular? One characteristic is quality: Table 3 shows a correlation coefficient between log price and popularity of 0.36. More expensive segments thus have a larger average clientele interested in living there. To learn about how other characteristics shape segment popularity, we construct the housing-stock-weighted average popularity of all segments in a zip code. We then regress this measure on characteristics observable at the zip code level, in particular, school quality, the availability of restaurants and bars, crime levels, and weather conditions. The results in Table 4 show that the availability of bars and restaurants is the most important driver of popularity; other characteristics matter but are quantitatively less important. A one

\[\text{We measure school quality as the student-weighted average Academic Performance Index (API) across all schools in a zip code, as reported by the California Department of Education. To measure the availability of restaurants, we divide the number of establishments with SIC code 58 (Eating & Drinking Places) by the number of housing units. Crime levels are measured on a scale of 0-100, as provided by Bestplaces.net. To measure weather, we calculate the total number of inches of rain, the total number of cooling degree days, and the total number of heating degree days, as reported by Melissa Data. Heating Degree Days (HDD) and Cooling Degree Days (CDD) are measures of how far, and for how long, temperatures deviate from 70 Fahrenheit. For example, every day the temperature is at 65, counts as 5 HDD. Every day the temperature is 80 counts as 10 CDD.}\]
A one standard deviation increase in school quality increases popularity by 0.18 standard deviations, whereas a one standard deviation increase in violent crime reduces popularity by 0.14 standard deviations. Zip code weather also matters: more rain, more extreme hot weather, and more extreme cold weather all reduce popularity.

Table 4: Drivers of Zip Code Level Popularity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Effect of a 1 s.d. increase on zip code popularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>School Quality (Average API / 100)</td>
<td>0.053**</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td>(0.026)</td>
<td></td>
</tr>
<tr>
<td>Restaurants &amp; Bars (Per 100 housing units)</td>
<td>0.154***</td>
<td>0.185</td>
</tr>
<tr>
<td></td>
<td>(0.028)</td>
<td></td>
</tr>
<tr>
<td>Violent Crime (Scale 0-100)</td>
<td>-0.002</td>
<td>-0.038</td>
</tr>
<tr>
<td></td>
<td>(0.001)*</td>
<td></td>
</tr>
<tr>
<td>Rain (Inches)</td>
<td>-0.001**</td>
<td>-0.018</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>Cooling Degree Days (k days)</td>
<td>-0.115**</td>
<td>-0.034</td>
</tr>
<tr>
<td></td>
<td>(0.051)</td>
<td></td>
</tr>
<tr>
<td>Heating Degree Days (k days)</td>
<td>-0.054*</td>
<td>-0.027</td>
</tr>
<tr>
<td></td>
<td>(0.029)</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Table shows results from an OLS regression of zip code level popularity estimated from the model on observable characteristics. N = 183. Robust standard errors in parentheses. Significance levels: * (p<0.10), ** (p<0.05), *** (p<0.01).

**Search Breadth and Demographics**

Who are the broad searchers? While we do not directly observe demographics for the individuals setting email alerts in our data, our model estimates imply a distribution of search breadth for households living in equilibrium in each segment. We can thus compare the average search breadth of people living in a zip code, measured as the share of Bay Area inventory scanned, with demographic information from the 5-year estimates of the 2012 American Community Survey.\(^\text{16}\)

We focus on individuals’ age, income, and the presence of children. Figure 5 presents binned scatter plots to show the relationship between these measures of demographics and search breadth. In Panel A, we group our 191 zip codes by the median age of the inhabitants. There is a strong negative relationship between age and search breadth. People living in zip codes with median age of 30 search five times as much inventory, on average, as people living in zip codes with median age 50. Panel B shows that the average search breadth of households living in zip codes with many children is higher, and Panel C shows that people living in zip codes with higher median income

\(^\text{16}\)The average search breadth of people living in a zip code in equilibrium can be different from that of people searching in a zip code, which we explored in Figure 3.
have smaller search ranges. All these relationships are true both unconditionally, and conditional on the other two demographic measures. Overall, this shows that younger households, poorer households, and households with children appear to have broader search ranges.

Figure 5: Search Breadth and Demographics

(A) By Median Age  
(B) By Child Share  
(C) By Median Income

Note: Figure shows binned scatter plots of the average search breadth of people living in a zip code in equilibrium, separately by zip code demographics. We show search breadth by median age (Panel A), in by the share of population with children (Panel B), and by median zip code income (Panel C).

Moving Patterns in Model vs. Data

While our estimation has targeted only moments of the cross section of market activity, the model also has implications for flows of households from segments. We now confront those implications with data on actual household moves. This exercise provides a joint over-identifying restrictions test of the structure of the model – in particular, our assumption on broad searchers flowing to segments in proportion to inventory – as well as the quality of our search data. To measure the flow of households between segments, we use a sample of all individuals who moved to a new Bay Area address between May 2012 and October 2012. The data come from Acxiom, a marketing analytics company that compiles this list of movers from “Change of Address” notices filed with the USPS. Our sample contains 96,170 individuals moving to a new address in one of the 191 zip codes in our sample; for these individuals, we have information on their new and their previous address. We restrict attention to movers who had previously also lived in one of the 191 Bay Area zip codes, about 75% of our sample. We then compute the share of movers between each of the 191 × 191 (directed) zip code pairs, both in the model and the data. To obtain the model moments, we aggregate model implied segment-to-segment flows to the zip code level.

Figure 6 shows a scatter plot of the mover share in the data and the model at the (directed) zip code-pair level. The association is strong: the correlation coefficient is 83%. The result is not driven by households moving within the same zip code: when we exclude such moves, the correlation drops only slightly, to 75%. We conclude that moving patterns provide additional support for our quantitative account of Bay Area housing market dynamics.
Figure 6: Moving Activity: Model vs. Data

Note: This figure shows the share of total moves predicted by the model (horizontal axis) vs. the moving share in the data (vertical axis) for each (directed) pair of zip codes.

6 Prices and Spillovers

In this final section, we explore how the forces in our model relate to equilibrium prices across segments. We also explore how the clientele structure in the Bay Area shapes the response of different segments to housing market shocks, such as the influx of new narrow searchers as the result of the gentrification of neighborhoods.

6.1 Equilibrium Prices and Illiquidity 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 illiquidity discount reflects both frictions. Let $V^F(h; \theta)$ be the utility of a type-$\theta$ 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-$\theta$ 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.

\[
rV^S(h; \theta) = \frac{m(h)}{\mu^S(h)} (E \left[p(h, \theta) (1 - c) | h \right] - V^S(h; \theta)),
\]
where the expectation uses the equilibrium distribution of buyers of type $\theta$. 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

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

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:

$$p (h) = \frac{v (h)}{r} - \frac{\eta (h)}{r + m (h)/\mu^S (h) + \eta (h)} \left( \frac{v (h)}{r} + c p (h) \frac{m (h)}{\mu^S (h)} \right). \quad (8)$$

The first term on the right-hand side is the present value of the permanent flow of housing services. This frictionless price obtains if houses never fall out of favor ($\eta = 0$) or if sales involve no transaction costs ($c = 0$) and no search costs if matching is infinitely fast ($m/\mu^S \to \infty$). Indeed, if a new buyer can be found instantaneously and at no monetary cost, then the price captures the permanent flow of services. More generally, the house price reflects an illiquidity discount – the second term – that capitalizes search and transaction costs. The illiquidity discount is larger if houses fall out of favor more often (higher $\eta$) 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 illiquidity 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 (8) 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

$$p (h) \approx \frac{v (h) (1 - I (h))}{r + c V (h)} = \frac{v (h)}{r} (1 - I (h)) \frac{r}{r + c V (h)}. \quad (9)$$

Frictions thus modify the frictionless price $v/r$ by first reducing housing services proportionately by $I$ and then increasing the discount rate to $r + c V$. 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 1, 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.

**Illiquidity 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 illiquidity discount in the pricing equation (8), 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 (8).

**Figure 7: Illiquidity Discounts**

![Figure 7: Illiquidity Discounts](image)

**Note:** Left panel: mean segment price versus segment illiquidity discount in percent of frictionless price. Color coding reflects illiquidity discount. Right panel: map with dots for each segment in same color as in left panel. Dots for segments within the same zip code are 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 illiquidity discounts, stated as a percentage of frictionless price. The right-hand panel shows the geographic distribution of the illiquidity discounts. There are two notable results here. First, illiquidity discounts are large. The median discount is 14 percent and the 90th percentile is at 24 percent. From Table 1 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 illiquidity discounts differ widely by segment, often within the same zip code. Table 2 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 static 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, when 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 value 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 there: depending on the segment within the zip code, between 12 and 35 percent of buyers were searching only in 94015, with the largest narrow buyer share in the most expensive segment.

The second neighborhood is San Francisco’s Outer Mission, zip code 94112. For comparability with 94015, we select only the cheapest three segments in this zip code; we thereby obtain about the same total housing stock and average price. However, the population of searchers is quite different: there are 2,000 searchers looking at an average inventory of 84 houses. Moreover, the share of broad searchers is large: at most 3 percent of buyers search exclusively in zip code 94112.

We now recompute equilibrium under the assumption that 500 narrow 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 by 500 agents. This counterfactual captures long-run effects on market conditions 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 η(h) as before. The time to sell a house thus moves proportionally with inventory. Moreover, price effects follow from equation (9): 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,
Figure 8: Inventory Response to an Increase in Popularity; Joint Searchers

(A) More searchers in Daly City (94015)

(B) More searchers in San Francisco Outer Mission (94112)

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.
which obtains in both cases within the zip codes that become more popular, shaded in gray.

Clientele patterns imply dramatically different behaviors 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 Daly City and 0.82 for the three cheapest segments in the Outer Mission: 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 for Daley City and the Outer Mission, respectively. An econometrician using a perfectly segmented economy to assess the inflow of new searchers would therefore incorrectly expect more competition for scarce housing in the Outer Mission compared to Daly City.

7 Conclusion

Most search markets feature competition from broad and narrow searchers. We show that understanding the exact structure of these search clienteles is important for understanding the forces behind equilibrium market outcomes such as the shape of the Beveridge curve, as well as the response of different market segments to shocks. In the absence of comprehensive data on buyer search patterns, previous studies of housing search behavior had to make stark assumptions on the segmentation structure of the markets under study. We show how data from online search behavior can allow researchers to overcome this measurement challenge. We expect that similar data from online services such as Facebook, LinkedIn, Tinder, ZipRecruiter, and Indeed, will allow researchers to measure the clientele structure in other search markets, from dating to job search. As highlighted in this paper, this will improve our understanding of both the cross sectional patterns across submarkets, as well as the response of these markets to shocks.
References


USA Today, “Housing shortage eases in some markets,” 2016.

A Description of Search Behavior

In this Appendix we provide more descriptive statistics of the housing search behavior inferred from our email alerts data. In Appendix A.1, we analyze the frequency with which the major dimensions of search, geography, price, and the number of bathrooms are selected. In Appendix A.2, we establish stylized facts on the geographic breadth of housing search. In Appendix A.3, we discuss the key characteristics of housing search along the price and size dimensions. In Appendix A.4, we provide information about how the size and price dimension correlate with the geographic breadth of the search range. For example, we document that searchers that are more specific about the price range and home size cover larger geographic areas.

A.1 Major Dimensions of Search

As discussed in the paper, all email alerts are required to provide the geographic dimension of the potential home buyer’s search range. Roughly a third of the alerts do not specify any fields in addition to geography. The other fields that are used regularly include listing price and the number of bathrooms. Table A.1 shows the distribution of the dimensions that are specified across the email alerts in our sample. Just under a third of email alerts specify both price and the number of bathrooms, while another third specify just a price range. The remaining 5 percent of email alerts 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 email alerts specify square footage while 2.7 percent of alerts 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.

Table A.1: Distribution of Email Alert Parameters

<table>
<thead>
<tr>
<th></th>
<th>Price not specified</th>
<th>Price specified</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baths not specified</td>
<td>13,019</td>
<td>13,777</td>
<td>26,796</td>
</tr>
<tr>
<td>Baths specified</td>
<td>1,848</td>
<td>11,881</td>
<td>13,729</td>
</tr>
<tr>
<td>Total</td>
<td>14,867</td>
<td>25,658</td>
<td>40,525</td>
</tr>
</tbody>
</table>

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

A.2 Search by Geography

We next describe the geographic dimensions of housing search inferred from the Trulia email alerts. We begin by outlining how we deal with alerts that report the geography dimensions at different levels of aggregation. We then provide summary statistics on the distribution of distances covered by the various email alerts.
A.2.1 Assigning Zip Codes to Email Alerts

Each email alert defines the geographic dimension of housing search 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 in terms of neighborhoods. Some searchers include geographies in terms of cities, zip codes, and neighborhoods in the same alert. In order to compare email alerts that specify geography at different levels of aggregation, we translate every alert into the set of zip codes that are (approximately) covered. This requires dealing with alerts that specify geography at a level that might not perfectly overlap with zip codes. For alerts that select listings at the city level, we include all zip codes that are at least partially within the range of that city (i.e., for a searcher who is looking in Mountain View, we assign the alert 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 (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, a list of zip codes that are covered by each email alert. The alerts cover 191 unique Bay Area zip codes.

A.2.2 Distance

To summarize how search ranges reflect geographic considerations, we construct various measures of size of the area considered. Since the unit of observation we are interested in is the searcher, not the email alert, we pool all zip codes that are covered in at least one email alert by a particular searcher. 26 percent of searchers consider only a single zip code. For the remaining searchers, we measure the average and maximum 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>
<thead>
<tr>
<th>Population-Weighted Zip Code Centroids</th>
<th>Min</th>
<th>Bottom Decile</th>
<th>Median</th>
<th>Top Decile</th>
<th>Max</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Geographic Distance</td>
<td>0.5</td>
<td>2.3</td>
<td>6.8</td>
<td>21.1</td>
<td>103.3</td>
<td>9.7</td>
</tr>
<tr>
<td>Mean Geographic Distance</td>
<td>0.5</td>
<td>1.8</td>
<td>3.2</td>
<td>8.9</td>
<td>74.0</td>
<td>4.7</td>
</tr>
<tr>
<td>Max Car Travel Time</td>
<td>4.0</td>
<td>9.5</td>
<td>20.5</td>
<td>38.5</td>
<td>143.5</td>
<td>22.8</td>
</tr>
<tr>
<td>Mean Car Travel Time</td>
<td>3.8</td>
<td>8.9</td>
<td>13.1</td>
<td>19.7</td>
<td>132.5</td>
<td>14.0</td>
</tr>
<tr>
<td>Max Public Transport Time</td>
<td>10.5</td>
<td>40.5</td>
<td>79.0</td>
<td>375.0</td>
<td>573.5</td>
<td>140.1</td>
</tr>
<tr>
<td>Mean Public Transport Time</td>
<td>9.3</td>
<td>27.3</td>
<td>48.0</td>
<td>120.0</td>
<td>375.0</td>
<td>169.9</td>
</tr>
</tbody>
</table>

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.
Table A.2 reports these measures of the geographic breadth of housing search. Geographic distance is measured in miles, and is direct “as the crow flies.” For the average searcher, the maximum distance between two zip codes included in a search range is 9.7 miles. There is significant heterogeneity in the geographic breadth of the search ranges. Indeed, a searcher at the 90th percentile of the distribution has a maximum distance between covered zip codes of 21.1 miles, while a searcher at the 10th percentile has a maximum geographic distance of 2.3 miles. These would usually be searchers that select two neighboring zip codes.

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.\textsuperscript{17} 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; again, there is sizable heterogeneity in this measure; the across-searcher 10-90 percentile range of maximum travel times is 9.5 minutes to 38.5 minutes for travel by car, and 40.5 minutes to 375 minutes for travel by public transport.

In Appendix C we show that these 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 represent time-invariant measures of household preferences.

### A.2.3 Contiguity

To guide our modeling of clientele heterogeneity, we next explore whether there is a simple and parsimonious organizing principle for observed geographic 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. We begin by describing how we construct measures of contiguity, before providing summary statistics on how many search queries cover contiguous geographies.

**Dealing with the SF Bay**

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.1 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

\textsuperscript{17}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.
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 email alerts, 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.1: Bridge Adjustments - Contiguity Analysis

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

Examples of Contiguous and Non-Contiguous Search Sets

In the following, we provide examples of contiguous and non-contiguous search sets. In Figure A.2, we show four actual contiguous search sets from our data. 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. 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
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.

right panel) was generated by a searcher that selected a significant number of South Bay cities.\footnote{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, and Woodside.} 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.

In Figure A.3, 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 sets 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.
Figure A.3: 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.

Summary Statistics on Contiguity

Table A.3 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 email alerts that select contiguous geographies. While only 18 percent of searchers have non-contiguous search ranges, they tend to come from 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 maximum number of contiguous areas covered by an email alert. 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.3: Contiguity Analysis – Summary Statistics

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

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

### A.2.4 Circularity

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 A.4 illustrates this procedure.

About 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 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.

Overall, we conclude that real-world housing search behavior cannot be well approximate using a simple and parsimonious search specification, either in terms of selecting contiguous geographies, or in terms of taking a "circular" search approach. This conclusion motivates our modeling approach in the paper, which is highly flexible and allows us to capture non-contiguous and non-circular search patterns.
A.3 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.5 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.5 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
Figure A.5: 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.
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.5 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 studios, 1 bedroom
apartments, and very small houses.

A.4 Tradeoffs between Search Dimensions

The three major search dimensions we have identified (geography, price, and size) are not necessar-
ily 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.4 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 searchers who do not specify a price range 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.

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

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

A.5 Search within Search Range

In Section 3.2, we provided evidence for our key modeling assumption that the probability that
a searcher would find her favorite home in any one of her considered segments is proportional to
the share of that segment’s inventory in total inventory across all considered segments. In this
appendix, we provide some additional evidence for this assumption.
Figure A.6: Example of Search Return List

Note: Example of list of properties for sale returned by trulia.com for a search query looking at all houses in zip code 10012.
Unfortunately, we cannot map individual home searchers to their final purchases. Therefore, to address this question, we exploit data on property views by individual home searchers on trulia.com. After defining a search range on trulia.com, a user is presented with a list of properties that are included in that search range (see Appendix Figure A.6). This list provides basic information on each property, such as a picture, its location, the listing price, and the first lines of a description of the property. Home searchers then actively click on those houses that attract their particular interest to view additional property information.

We have obtained these individual property-view data from trulia.com for a random subset of users visiting the site in April 2012. These data contain the set of listings viewed within a “session,” defined as all views by the same user within one day. We use these data to test whether users express interest by clicking on properties across different segments in the same proportion as those segments’ inventory in the searcher’s overall range. While we are unable to observe a user’s search range in these particular data, we focus on those sessions that contain views of properties located in at least two of our 576 Bay Area segments. We observe information from 6,242 such sessions. About 25% of sessions include two property views, 19% include three property views, and 15% include more than 10 property views. In about half of these sessions, users view properties in just two segments, in 20% of sessions users view properties in 3 segments, and in 5% of sessions users view properties in more than 7 segments. For each session and segment, we calculate the share of views of houses in that segment, relative to the total number of houses viewed in the session (“view share”). We also calculate the share of inventory in that segment, relative to the total inventory in all segments in which houses were viewed in that session (“inventory share”).

Appendix Figure A.7 shows binned scatter plots documenting the relationship between view share and inventory share. The unit of observation is a session-segment. In Panel A we focus on the 2,704 sessions where the user viewed at least five individual properties. We split the resulting 12,225 session-segments into 20 equally sized buckets ordered by inventory share. On the horizontal axis we plot, for each bucket, the average inventory share, and on the vertical axis we plot the average view share. There is a strong positive relationship. The rate at which particular interest is expressed for properties in different segments is increasing in the share of total inventory made up by those segments in the overall search range. For those segments with the lowest inventory share, we find view shares to be somewhat above inventory shares, while for segments with a very high inventory share we find view shares to be somewhat below inventory shares. This effect is mechanical, and a result of observing relatively few property views per session on average, combined with the fact that we infer a searcher’s search range from the properties viewed. To see this, take a session in which 5 properties were viewed. For us to know that a segment was included in that session’s search range, we need to observe at least one property from that segment being viewed. Therefore, any segment that we know is included will at least have 20% view share, even if its inventory share was only 5%. As we investigate sessions with more properties viewed, this bias gets smaller, as shown in Panels B to D of Appendix Figure A.7.
Figure A.7: “View Share” vs. “Inventory Share”

(A) At least 5 property views per session (N = 2,704)  
(B) At least 10 property views per session (N = 1,043)  
(C) At least 15 property views per session (N = 500)  
(D) At least 20 property views per session (N = 279)

Note: Figure shows binned scatter plots at the segment-session level; different panels vary the minimum number of property views required for a session to be included. On the horizontal axis is the share of inventory of a segment, relative to the total inventory in all segments viewed in that session. On the vertical axis is the share of properties in that segment viewed, relative to the total number of properties viewed in that session.
B Construction of Segments, and Segment-Level Activity

In this Appendix, we discuss how we construct the housing market segments based on the email alerts in our data. We also highlight how we then assign segments to email alerts, and how we measure segment-level housing market activity.

B.1 Segment Construction

This section describes the process of arriving at the 564 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 describe how we further split these zip codes into segments based on a quality (price) and size dimension. Importantly, we 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 contain information on the universe of the housing stock. However, as a result of Proposition 13, assessed property values in California do not correspond to true market value, and it is thus not appropriate to divide the total zip code housing stock into different price segments based on this assessed value. To measure the housing stock in different price segments, we thus use data from the 5-year estimates from the 2011 American Community Survey, which reports the total number of owner-occupied housing units per zip code for a number of price bins. We use these data to construct the total number of housing units in each of the following price bins: $<200k, 200k−300k, 300k−400k, 400k−500k, 500k−750k, 750k−1m, >1m$. 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 adjust all prices and price boundaries to 2010 house prices.

Not all zip codes have an equal distribution of houses in each price (quality) bin. For example, Palo Alto has few homes valued at less than $200,000, while Fremont has few million-dollar homes. Since we want to avoid cutting a zip code into too many quality segments with essentially

---

19 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.

20 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 in each price bin reported in the ACS into a housing stock for different 2010-"quality" segments, we first construct zip code-specific annual repeat sales house 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.
no housing stock 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 ACS price bin cut-offs should constitute segment cut-offs, we use information from the email alerts. This proceeds in two steps: First we change the price parameters set in the email alerts to account for the fact that we observe alerts 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 alert was set and 2010. 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) email alert price restrictions to the closest cut-off. We select the set of segment-specific price cut-offs that minimizes the average of this value across all alerts that cover a particular zip code. This ensures, for example, that if many email alerts covering a zip code include a high cut-off such as $1 million (either as an upper bound, or as a lower bound), $1 million is likely to also be a segment boundary.

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 rental 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 prices 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.

The other search dimension regularly specified in the email alerts is the number of bathrooms.

---

21 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 an alert set in 2006 that restricts price to be between $500,000 and $800,000 will search for homes in the same quality segment as an alert set in 2010 that restricts price to a $250,000 - $400,000 range.

22 For example, imagine testing how good the the boundaries 100k, 300k and 1m fit for a particular zip code. An alert with an upper bound of 500k has the closest absolute distance to a cut-off of \(\min\{500 - 100, 500 - 300, 500 - 1000\}\) = 200. An alert with an upper bound of 750k has the closest absolute distance to a cut-off of 250. An alert 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 alert the smallest absolute distance of an alert limit to a cut-off, and then find the average across all email alerts.

23 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.
as a measure of the size of a house conditional on its location and quality. Since Section A.3 shows 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 fewer 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 zip code by price segment into groups by bathrooms, 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 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 segment-specific moments such as the average time on market with some precision, we want 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 price segment, see whether it has a stock of less than 1,200, and if not merge it with the next higher price segment. This procedure generates 564 segments.

Figure B.1: Segment Overview

![Segment Overview](image)

Note: 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.

Figure B.1 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 B.1 shows the distribution of housing stock across segments.
On average, segments have a housing stock of 3,929, with a median value of 3,298. The largest segment has a housing stock of 13,167.

A First Look at the Segments

The left panel of Figure B.2 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 six 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 large heterogeneity within zip codes: indeed, 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 B.2 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 B.2: Segments in San Francisco and the Bay Area

Note: The left panel shows a map of downtown San Francisco as the shaded area in addition to areas south of downtown. The right panel shows a map of the entire Bay Area. The color bar indicates the price of the segment in thousands of Dollars.
B.2 Assigning Segments to Email Alerts

We next describe how we assign which segments are covered by each email alert. In Appendix A.2, we already discussed how we determine which zip codes are covered by which alert, and how we deal with alerts that specify geography at a different level of aggregation. We next discuss how we incorporate the price and size dimensions of housing search to determine which segments in a zip code are covered by each alert. The challenge is that price ranges selected will usually not overlap perfectly with the price cutoffs of the individual segments. For those alerts that specify a price dimension, we assign an alert to cover a particular segment in one of three cases:

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

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

3. For alerts that partially cover a non-open ended segment, we determine the share of the segment price range covered by the alert. For example, an alert with price range $200k - $500k covers 20% of the segment 0-$250k, and 50% of the segment $250k - $750k. We assign all alerts that include at least 50% of the price range of a segment to cover that segment.

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

Pooling across Queries

The housing market segments constructed above allow us to pool across all email alerts set by the same individual. In particular, we add all segments that are covered by at least one email alert of an individual to that individual’s search set. After pooling all segments covered by the same searcher in this way, we arrive at a total of 11,503 unique search profiles, set by the 23,597 unique users in our data. Figure B.3 shows the distribution of how many different searchers are represented by the different search profiles. A total of 9,264 search profiles represent only a single searcher, 940 search profiles represent two searchers. 438 search profiles represent at least seven searchers, while the most common search profile represents 788 unique searchers.

B.3 Construction of Segment-Level Market Activity

Segment Price. 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
Figure B.3: Number of Searchers per Unique Profile

Note: Figure shows how many unique individuals are represented by each of the 11,503 individual search profiles.

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, 2016, for details on this process of identifying arms-length transactions).

Volume Share. We then calculate the total number of transactions per segment between 2008 and 2011, and and use this to construct annual volume averages. To allocate houses to particular segments, we adjust transaction prices for houses sold in years other than 2010 by the same house price index we used to adjust listing price boundaries (see Appendix B.1). We then measure “volume share” by dividing the annual transaction volume by the segment housing stock.

Time on Market. 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. This match is done via the (standardized) addresses across the two data sets. Panel A of Figure B.4 shows the time series of the share of listings that we are unable to match to deeds data, starting in 2008, the first year of our estimation sample. On average, for properties listed between January 2008 and July 2011, we can match between 85% and 90% of all listings to subsequent transactions. This number is relatively constant across segments. There are three reasons for why we may not be able to match all listings to transactions:

---

24We measure time on market as the period between the first listing of the property and the final transaction; this combines across listing spells of properties that are repeatedly listed and delisted by real estate agents in order to avoid the appearance of a stale listing. We subtract one month to allow for the typical escrow period.
1. The listed property sells, but due to a different formatting of the address in the listing and the deed (or an incomplete address in the listing), we cannot match listing and sale.

2. The property got withdrawn from the market without being sold.

3. The property is still for sale by the end of our transaction sample (April 2012).

The increase in the share of listings without sales from the middle of 2011 onward is likely due to the last reason. Across all listings that we can match to a final transaction, the 90th percentile of time-on-market is 343 days, the 95th percentile is 502 days (the median is 84 days, and the mean is 144 days). This means that a significant number of properties listed toward the end of our sample will not sell by the end of our deeds data window (April 2012).

We find segment-level measures of time on market by averaging the time on market across all transactions between 2008 and 2011.\textsuperscript{25} We also constructed an alternative measure where we restricted the sample to be the time on market for all properties that were listed between 2008 and July 2011; for this calculation we excluded the last 6 months in order to avoid the problem of the censoring of time on market numbers for properties that are listed towards the end of our sample. Both measures provide very similar measures of segment-level time on market, and a nearly identical ranking of segments across this dimension.

**Figure B.4: Listings-Transaction Match Rates**

\begin{figure}[h]
\centering
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{FigureA}
\caption{Listings without sales}
\end{subfigure}
\begin{subfigure}{0.4\textwidth}
\centering
\includegraphics[width=\textwidth]{FigureB}
\caption{Sales without Listings}
\end{subfigure}
\end{figure}

**Note:** Panel A shows the number of listings for which we do not eventually find a deed to match, by month of listing. Panel B shows the share of transactions for which we also observe a listing, by month of sale. Both Panels cover the period 2008-2011, for which we observe Trulia email alerts.

\textsuperscript{25}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 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. In addition, in our baseline estimates we exclude the few properties where we observe a time of more than 900 days between listing and sale; this does not have a significant effect on the final measures of segment-level time on market.
**Inventory Share.** As discussed in the paper, in our baseline estimates we measure construct the inventory share with the steady-state equation $I = T \times V$ instead of calculating it using the actual listings we observe. There is a trade-off in this choice. In particular, as discussed in the paper, the downside of our approach is that it does not include in our measure of inventory properties that are listed and subsequently delisted. However, as we show above, this number is relatively small: it is bounded above by 10% to 15% of all listings, and manual examination of un-matched listings suggests that the vast majority of our inability to match properties to listings is due to incomplete addresses in listings. In addition, as discussed above, the share of listings without sales is relatively similar across segments; therefore, abstracting from withdrawn listings is unlikely to affect the cross-sectional patterns that we focus on in our analysis.

We also tried an alternative approach to constructing the inventory share based on actual listings that we observe in the data. Under this approach, a property gets added to inventory the first time a listing appears, and gets removed once the property sells. For every month, we then have a set of properties that are on the market in that month, and we can form segment-level averages by calculating the share of the housing stock that is for sale across all the months between January 2008 and December 2011. Since we observe listings starting in October 2005, we do not require a “burn in period” at the beginning of the sample. One issue with this approach is that the coverage of Trulia listings data during our time period is not complete. Indeed, Panel B of Figure B.4 shows the number of transactions that can be matched to a previous is increasing throughout our sample, from about 40% at the beginning, to about 60% towards the end of our sample. Much of this improvement is due to the increasing coverage of the Trulia data. To not significantly bias the results under this alternative approach, one has to make two adjustments to the data. First, one has to decide how to treat properties that never sell; in our baseline approach, we removed them from the sample if they stayed on the market for more than 270 days. Second, we have to adjust for the incomplete coverage of the Trulia deeds data, which might differ by segment. To do this, we scale up the actual inventory share by the fraction of deeds with no listings.

While the measures using both approaches produce similar cross-sectional patterns in the data, and similar cross-sectional rankings across segments, we prefer calculating the inventory share from the more precisely measured time on market; it has less measurement error, and, for the reasons discussed above, any bias introduced by ignoring delistings is likely to be small.

**B.4 Segment-Level Search Breadth**

In this Appendix, we present more details on the search-breadth of the various searchers covering each segment. The first column of Table B.1 summarizes the distribution of inventory scanned by the median searcher. In the average segment, the median searcher scans 2.1 percent of the total Bay Area inventory. The table also clarifies that most dots in the left panel of Figure 3 are clustered in the bottom left; the 75th percentile of the distribution is at only at 2.5 percent of total inventory. The second column in Table B.1 shows the distribution of the within-segment
interquartile range 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 narrow and broad searchers.

Table B.1: Variation in Scanned Inventory by Clienteles

<table>
<thead>
<tr>
<th>SEGMENT:</th>
<th>total inv. scanned (in percent)</th>
<th>ZIP CODE: share of search ranges (in percent)</th>
<th>CITY: share of search types (in percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>median IQ range</td>
<td>one</td>
<td>many</td>
</tr>
<tr>
<td>Mean</td>
<td>2.10</td>
<td>1.75</td>
<td>5.9</td>
</tr>
<tr>
<td>Q25</td>
<td>0.94</td>
<td>1.04</td>
<td>0</td>
</tr>
<tr>
<td>Q50</td>
<td>1.55</td>
<td>1.56</td>
<td>1.2</td>
</tr>
<tr>
<td>Q75</td>
<td>2.53</td>
<td>2.17</td>
<td>6.5</td>
</tr>
</tbody>
</table>

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 B.1 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 segment 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 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.
C Time-Series Stability of Patterns

Our empirical analysis focuses on the period 2008 to 2011, a period for which we are able to observe both search behavior and housing market activity (while we observe some email alerts from before 2008, the vast majority come from after that period). One natural question is thus the extent to which our conclusions generalize beyond the period under investigation. In the following sections we show that, as far as is possible to say with our data, both the search patterns as well as the ranking of segments in terms of market-level activity are relatively stable over time. This increases our confidence that the patterns we investigate in this paper are not just an artefact of the particular period studied.

C.1 Stability of Search Patterns

Our model interprets the observed search ranges as a time-invariant feature of buyer preferences. For example, our assumption in the comparative statics exercises in Section 6.2 is that the search breadth would not adjust based on the amount of inventory within the range. It is then interesting to analyze empirically whether 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 explore important summary statistics on the geographic breadth of each email alert between 2008 and 2011, split by the year in which the email alert was set. The results are presented in Figure C.1. We consider 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 as defined in Appendix A.2.4 (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, that does not respond to the relative supply available across the search range.

A second test 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 C.2, 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 C.2 show averages of the same summary statistics on the search parameters as Figure C.1, split by the month of the year when the email alert 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 C.1: 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 the year when the email alert was set. 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).

C.2 Stability of Market Activity

A related interesting question is whether the market-level outcomes are particular to our period of study. Unfortunately, we do not have the data to analyze the key patterns outside of our sample period; in particular, the listings data necessary to construct inventory are only reliably observed during that period. However, to provide some evidence that our results are not just a feature of a period of declining house prices, we exploit the fact that our sample period includes two rather distinct housing market episodes: a period of declining house prices, between January 2008 and March 2009, and a period of relatively flat or even increasing house prices between April 2009 and December 2011 (see Panel A of Figure C.3). This allows us to test whether the across-segment patterns we observe are similar in these two periods.

Panel B, C, and D of Figure C.3 show the across-episode correlation of segment-level house prices, volume share, and inventory share. There is a high correlation between these measures of
Figure C.2: 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.
Figure C.3: Stability of Segment Moments

(A) SF House Prices

(B) Segment Prices

(C) Segment Volume

(D) Segment Inventory

Note: Panel A shows the Case-Shiller House Price Index for San Francisco. Our sample period is delineated with vertical solid lines. It is split into a period of declining house prices and stable house prices by a dashed vertical line. Panels B, C, and D show segment-level scatter plots of key housing market moments (prices, volume, and inventory, respectively) across the periods of declining and stable house prices.

market activity across the “declining market” period and the “stable market” period: zip codes with high volume and high inventory during the 2008 bust also have high volume and inventory during the subsequent stable price period. Indeed, the Spearman’s rank correlation coefficient for these two variables is 0.70 and 0.75, respectively. It is much higher, at 0.96, for the more-precisely measured average house prices per segment.

We can also analyze the correlation between different moments across segments in both the bust period and the stable period, and compare it to the pooled correlation presented in Table 4 of the paper. For example, inventory and volume had an across-segment correlation of 0.93 in the pooled period. This measure was 0.81 and 0.83 in the bust and the stable period, respectively. Similarly, the correlation between inventory and price levels in the pooled sample was -0.63, while it was -0.29 and -0.43 in the bust and the stable period, respectively. The lower correlation in
either of the sub-periods highlights the additional noise introduced by splitting the sample, and reinforces our choice to analyze the pooled sample in our baseline analysis.

Therefore, while we cannot rule out that the observed relationship between inventory, volume, and search behavior might look different during a housing boom period, it is reassuring that there are no significant differences in the cross sectional relationship between these variables during periods of strongly declining prices and periods of stable prices.
D 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.

Analogous to equation (3), the frequency of moving shocks \( \eta(h) \) can be written directly as a function of inventory and turnover:

\[
\eta(h) (1 - I(h)) = V(h).
\] (D.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 (6), it can be expressed in terms of observables (up to a constant) as

\[
\frac{1}{\alpha(h)} = \sum_{\theta \in H(h)} \frac{I(h) \beta(\theta) (\mu^\Theta - 1)}{\mu^H(\theta)} \frac{1}{V(h)}.
\] (D.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 \( \mu^\Theta \). 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 \( \theta \) is the same as the average inventory-weighted match rate across segments. Indeed, let \( \bar{\mu}^S \) denote total inventory and consider the identity

\[
(\bar{\mu}^S)^{-1} \sum_{\theta \in \Theta} \frac{\nu^S(\theta)}{\bar{\mu}^B(\theta)} \sum_{h \in H(\theta)} \frac{\mu^S(h)}{\nu^S(\theta)} \frac{m(h)}{\mu^B(h)} = (\mu^S)^{-1} \sum_{h \in H(\theta)} \frac{\mu^S(h)}{\mu^B(h)} m(h).
\]

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 \( \theta \) which depends on the relative inventory available in \( \theta \)’s search range as well as the share of \( \theta \) in each segment’s buyer pool.

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

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

The adding up constraint says that the owner-buyer ratio for type \( \theta \) agents should be the inventory weighted average of owner-buyer ratios at the segment level. We will therefore infer the presence of more types \( \theta \) not only if we observe more buyers of type \( \theta \) (higher \( \beta(\theta) \) and hence higher \( \tilde{\mu}^B(\theta) \)), but also if type \( \theta \)'s search range has on average relatively more owners relative to buyers. In the latter case, more types \( \theta \) 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 (??). For example, consider the Cobb-Douglas case with a multiplicative segment-specific parameter \( \bar{m}(h) \) that governs the speed of matching

\[
\bar{m}(\mu^B(h), \mu^S(h), h) = \bar{m}(h) \mu^B(h)^{\delta} \mu^S(h)^{1-\delta}.
\]

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