Testing for Information Asymmetries in Real Estate Markets

Online Appendix

Pablo Kurlat Johannes Stroebel

A. Theoretical Appendix

This Appendix formally describes the model and equilibrium concept that generate Empirical Predictions 1-4. Before describing the model, it is worth explaining why existing asset pricing frameworks are unsuitable for analyzing the environment we want to consider.

Models in the spirit of Akerlof (1970) assume that all buyers are equally well informed, and assert that assets which buyers cannot distinguish are pooled at the same price. In our framework, we want to allow for differential information of buyers as the ability to tell apart homes of different quality. Only some buyers can tell apart good and bad neighborhoods and houses, so whether or not two assets are distinguishable depends on the identity of the buyer. Therefore, the price pooling assumption cannot be made without further qualification.

Models in the spirit of Grossman and Stiglitz (1980) allow for differential information by different investors, but they make an assumption that is not well suited to model real estate markets. In those models, assets are divisible and clearly identified, and each has its own separate market. Some traders may be less informed about the expected value of the asset, but they can be sure to be trading the same asset as other traders. With indivisibility, which is a natural feature of the housing market, uninformed buyers cannot just free-ride on the informed buyers' information by purchasing the same house. Furthermore, given our notion of relative informedness, only some traders can tell whether two houses have different fundamentals. This implies that an uninformed buyer who buys house A does not know if he is trading the same type of house than a trader who buys house B.

To capture these important aspects of housing markets (indivisibility, and differential information among both buyers and sellers), we model an environment where some buyers and sellers have different information quality and, even though there is no aggregate noise, the difficulty in telling apart different housing types prevents information aggregation. To do this, we extend the analysis in Kurlat (2014) to account for the indivisibility of houses and for heterogeneity among sellers.

A.1 Agents and preferences

There is a unit measure of houses. Current home owners decide whether to offer their homes for sale. If an owner stays in his house, he will derive utility

$$u = [\beta\theta + \eta]\,\varepsilon \tag{A.1}$$

from living in it. θ is distributed according to F_θ, with support in $\left[0, \bar{\theta}\right]$. It is a common shock to all houses in a given neighborhood but takes different values in different neighborhoods. η is distributed according to F_η, with support in $[0, \bar{\eta}]$. It takes a different value for each house. β is the relative weight of neighborhood-level factors in the overall value of the property. We refer to $v \equiv \beta\theta + \eta$ as the total value of a house. The distribution of v, denoted F with density f and support $[0, \bar{v}]$ results from the convolution of $\beta\theta$ and η. Values of ε are distributed according to G, with support in $\mathbb{R}^+$. The variables θ, η and ε are independent. There is a large mass of potential buyers of houses. They all have identical preferences, and their valuation for a house is equal to its total value v, i.e., $\varepsilon = 1$ for all potential buyers.

A.2 Information

Both θ and η are private information of each house's current owner. Buyers can be of two types, informed or uninformed. There is a measure n^U of uninformed buyers. They have no information on either θ or η for any house. There is a measure n^I of informed buyers. They receive a signal from each house, given by:

$$s(v) = \mathbb{I}[v \geq b] \tag{A.2}$$

for a house of total value v, where $b > 0$. This signal allows informed buyers to determine whether any given house is a relatively good house ($v \geq b$) or a relatively bad house ($v < b$).

A.3 Equilibrium

We study a competitive equilibrium, abstracting from transaction costs and search frictions. We propose a notion of competitive equilibrium that respects uninformed buyers' inability to tell houses apart from one another.

Trading takes place as follows: There exists a finite set of prices $P \subseteq [0, \bar{v}]$ at which a house may trade.[1] An auctioneer calls out the prices in P in descending order. At every price, owners decide whether or not to put their house on sale. If an owner does not put his house on sale at price p, then he withdraws from the market and cannot put it on sale at any price $p' < p$; therefore, owners' decisions can be summarized by a reservation price.

At every price, potential buyers must decide whether to buy a house among those on offer. If an uninformed buyer buys a house, he picks one at random from among the houses that are offered. If an informed buyer buys a house, he can be selective and only accept houses for which he observes $s(v) = 1$, i.e., only relatively good houses; he picks a house at random from among the relatively good houses on offer.[2] Letting $S(p, v)$ denote the supply of houses of value v at price p, the expected value of the house bought by uninformed and informed buyers, respectively, will be:

[1] Finiteness is assumed to avoid mathematical complications but is not essential. One can think of prices being rounded to the nearest dollar.

[2] Assuming buyers always receive a random house from among the acceptable ones requires assuming that the relative proportions of different acceptable houses do not depend on which other buyers buy at the same price. A sufficient condition for this is to impose that uninformed buyers pick houses first. Kurlat (2014) shows that if one allows for different possible orderings, this is indeed the one that emerges in equilibrium.

$$\bar{v}^U(p) = \frac{\int_0^{\bar{v}} vS(p,v)\,dv}{\int_0^{\bar{v}} S(p,v)\,dv} \tag{A.3}$$

$$\bar{v}^I(p) = \frac{\int_b^{\bar{v}} vS(p,v)\,dv}{\int_b^{\bar{v}} S(p,v)\,dv} \tag{A.4}$$

as long as the denominator is positive; if the denominator is zero, then the buyer will not find an acceptable house at price p.

Note that, depending on the buyers' decisions, many of the houses that are offered at any price p will remain unsold; some of these will perhaps be sold at lower prices. As in Gale (1996), we do not assume that supply equals demand at any particular price. Instead, the probability of trade is what clears the market at every price.

We denote the reservation price for owners with a house of value v and idiosyncratic shock ε by $p^R(v,\varepsilon)$ and the probability that a house of value v sells at price p by $\rho(p,v)$.

Definition 1. *A competitive equilibrium consists of:*

1. *Reservation prices $p^R(v,\varepsilon)$ for every $\{v,\varepsilon\}$;*
2. *Measures of houses $d^U(p)$ and $d^I(p)$ demanded by uninformed and informed buyers respectively at each price $p \in P \cup \emptyset$, with $\sum_{P\cup\emptyset} d^U(p) = n^U$ and $\sum_{P\cup\emptyset} d^I(p) = n^I$;*[3]
3. *Supply $S(p,v)$ for each $\{p,v\}$.*
4. *Selling probabilities $\rho(p,v)$ for each $\{p,v\}$;*

such that

1. *Reservation prices are set optimally, i.e. $p^R(v,\varepsilon) = v\varepsilon$*
2. *Buyers choose p optimally, i.e. if $d^U(p) > 0$ then p solves*

$$\max_p \bar{v}^U(p) - p \tag{A.5}$$

   and if $d^I(p) > 0$ then p solves

$$\max_p \bar{v}^I(p) - p \tag{A.6}$$

   where it is understood that $\emptyset$ maximizes (A.5) or (A.6) if there is no p such that the objective is strictly positive.

3. *Supply is consistent with reservation prices and selling probabilities*

$$S(p,v) = f(v) \times G\left(\frac{p}{v}\right) \times \prod_{\tilde{p}>p} [1 - \rho(\tilde{p},v)] \tag{A.7}$$

[3] $d^U(\emptyset)$ and $d^I(\emptyset)$ denote the measures of uninformed and informed buyers who choose not to buy a house.

4. The probabilities of selling are consistent with agents' decisions:

$$\rho(p,v) = \frac{d^U(p)}{\int_0^{\bar{v}} S(p,\tilde{v})d\tilde{v}} + \frac{d^I(p)\mathbb{I}[v \geq b]}{\int_b^{\bar{v}} S(p,\tilde{v})d\tilde{v}} \tag{A.8}$$

Equation (A.7) says that the supply of houses of value v at price p will be equal to the total number of owners who (i) have a house of value v, (ii) are willing to sell at price p, i.e., have $\varepsilon \leq \frac{p}{v}$, and (iii) have tried and failed to sell their house at every price higher than p. Equation (A.8) is interpreted as follows: for each type of buyer, the probability of selling a house of value v is the ratio of the demand of that buyer (provided he accepts houses of value v) to the total supply of houses that buyer accepts. Adding up over uninformed and informed buyers results in (A.8).

Under Assumptions 1-3 below, the equilibrium will be such that all trades of houses that look indistinguishable to uninformed buyers take place at the same price p^*. Thus, even though the equilibrium construct allows for trading at many possible prices simultaneously, the on-equilibrium trading behavior is actually quite simple: all houses that are observationally equivalent to uninformed buyers trade at the same price, but informed buyers can pick better houses at that price.

A.4 Equilibrium Characterization

Let p^* be the highest solution to

$$p = \bar{v}^U(p) = \frac{\int_0^{\bar{v}} vS(p,v)\,dv}{\int_0^{\bar{v}} S(p,v)\,dv} \tag{A.9}$$

and make the following assumptions:

Assumption 1. $n^I < \int_b^{\bar{v}} S(p^*,v)\,dv$.

Assumption 2. $n^U > \int_0^{\bar{v}} S(p^*,v)\,dv \times \left(1 - \frac{n^I}{\int_b^{\bar{v}} S(p^*,v)dv}\right)$.

Assumption 3. *For any* $p < p^*$, $\bar{v}^U(p) - p \leq 0$.

Assumption 1 says there are few informed buyers; Assumption 2 says there are many uninformed buyers. Together, they ensure that the marginal buyer is uninformed and makes zero profits. These assumptions are consistent with our empirical measures on informedness, which indicate there are relatively few informed buyers. Assumption 3 ensures that there is a unique equilibrium price. None of these assumptions is essential for the main predictions of the model, but they simplify the characterization of the equilibrium.

Proposition 1. *Under Assumptions 1-3, the equilibrium is given by:*

1. $p^R(v,\varepsilon) = v\varepsilon$*;*

2. *Demand*

$$d^I(p) = \begin{cases} n^I & \textit{if } p = p^* \\ 0 & \textit{otherwise} \end{cases}$$

$$d^U(p) = \begin{cases} \int_0^{\bar{v}} S(p^*, v)\, dv \left(1 - \frac{n^I}{\int_b^{\bar{v}} \int S(p^*, v) dv}\right) & \textit{if } p = p^* \\ n^U - d^U(p^*) & \textit{if } p = \emptyset \\ 0 & \textit{otherwise} \end{cases}$$

3. *Supply*

$$S(p, v) = \begin{cases} f(v)\, G\left(\frac{p^*}{v}\right) & \textit{if } p \geq p^* \\ f(v)\, G\left(\frac{p^*}{v}\right)(1 - \rho(p^*, v)) & \textit{if } p < p^* \end{cases}$$

4. *Probabilities*

$$\rho(p^*, v) = \frac{d^U(p^*)}{\int_0^{\bar{v}} S(p^*, \tilde{v})\, d\tilde{v}} + \frac{n^I \mathbb{I}(v \geq b)}{\int_b^{\bar{v}} S(p^*, \tilde{v}) d\tilde{v}} \tag{A.10}$$

Proof. Reservation prices are set optimally by construction. Since p^* is the highest solution to (A.9), then $\bar{v}^U(p) \leq p$ for all $p > p^*$. Together with Assumption 3, this ensures that both p^* and $\emptyset$ are optimal choices for uninformed buyers, so d^U is consistent with buyer optimization. Assumption 2 ensures that $d^U(\emptyset)$ is positive and Assumption 1 ensures that $d^U(p^*)$ is positive. p^* is an optimal choice for informed buyers because $v^I(p^*) > p^*$ and $S(p, v) = 0$ for any v, p with $v \geq b$ and $p < p^*$. Finally, by construction, $S(p, v)$ and $\rho(p, v)$ are consistent with equilibrium. □

A.5 Predictions

Assume that some owners are informed and observe θ perfectly, while others are less informed and only observe a noisy signal x. Furthermore, assume that whether an owner is informed is independent of the realization of θ. Denote the conditional expectation of θ given x by $\hat{\theta}(x)$. The informed will sell their house if $\varepsilon \leq \frac{p^*}{\beta\theta + \eta}$, while the uninformed will sell theirs if $\varepsilon \leq \frac{p^*}{\beta\hat{\theta}(x) + \eta}$. The following result leads to Prediction 1.a. in the main body of the paper.

Proposition 2. *The proportion of informed agents among sellers is higher in the worst neighborhood ($\theta = 0$) than in the best neighborhood $\left(\theta = \bar{\theta}\right)$.*[4]

Proof. Given θ and η, the fraction of informed owners who choose to sell is

$$\Pr\left[\varepsilon \leq \frac{p^*}{\beta\theta + \eta} | \theta, \eta\right] = G\left(\frac{p^*}{\beta\theta + \eta}\right),$$

so integrating across η, the fraction of informed owners who choose to sell in a neighborhood of quality θ is

$$\Pr[\text{Sell}|\theta, \text{Informed}] = \int_0^{\bar{\eta}} G\left(\frac{p^*}{\beta\theta + \eta}\right) dF_\eta(\eta). \tag{A.11}$$

[4]Proposition 2 does not necessarily imply that the fraction of informed sellers decreases monotonically with θ, but this is true for many common cases; for instance it is true if θ and x are normally distributed.

Similarly, for uninformed sellers,

$$\Pr\left[\text{Sell}|\hat{\theta}(x), \text{Uninformed}\right] = \int_0^{\bar{\eta}} G\left(\frac{p^*}{\beta\hat{\theta}(x)+\eta}\right) dF_\eta(\eta),$$

so integrating across realizations of x:

$$\Pr[\text{Sell}|\theta, \text{Uninformed}] = \int_x \int_0^{\bar{\eta}} G\left(\frac{p^*}{\beta\hat{\theta}(x)+\eta}\right) dF_\eta(\eta)\, dF_{x|\theta}(x). \tag{A.12}$$

For any non-degenerate distribution $F_{x|\theta}$, $0 < \hat{\theta}(x) < \bar{\theta}$ for all x. Equations (A.11) and (A.12) then imply that $\Pr[\text{Sell}|0, \text{Informed}] > \Pr[\text{Sell}|0, \text{Uninformed}]$ and $\Pr\left[\text{Sell}|\bar{\theta}, \text{Informed}\right] < \Pr\left[\text{Sell}|\bar{\theta}, \text{Uninformed}\right]$, which gives the result. □

Suppose now that different houses within a neighborhood have different loadings on neighborhood and idiosyncratic factors; in other words, β_h is different for different houses. Assume that the distribution of β_h within a neighborhood is independent of the realization of θ.[5] The following result leads to Prediction 1.b.[6]

Proposition 3.

1. *The proportion of owners who choose to sell is constant with respect to β_h in the worst neighborhood ($\theta = 0$) and decreasing in β_h in every other ($\theta = \bar{\theta}$).*

2. *The proportion of owners who choose to sell in a neighborhood of quality θ does not change with θ for houses with $\beta_h = 0$ and decreases with θ for houses with $\beta_h > 0$.*

Proof. Owners sell their house if $\varepsilon \leq \frac{p^*}{\beta_h\theta+\eta}$, so the proportion of owners of houses with β_h who sell in a neighborhood of quality θ is:

$$\Pr[\text{Sell}|\theta, \beta_h] = \int_0^{\bar{\eta}} G\left(\frac{p^*}{\beta_h\theta+\eta}\right) dF_\eta(\eta) \tag{A.13}$$

1. Taking the derivative of (A.13) with respect to β_h:

$$\frac{d\Pr[\text{Sell}|\theta, \beta_h]}{d\beta_h} = -\int_0^{\bar{\eta}} g\left(\frac{p^*}{\beta_h\theta+\eta}\right) \frac{p^*}{[\beta_h\theta+\eta]^2}\theta dF_\eta(\eta)$$

   Evaluating this expression for $\theta = 0$ and $\theta > 0$ gives the result.

2. Taking the derivative of (A.13) with respect to θ:

$$\frac{d\Pr[\text{Sell}|\theta, \beta_h]}{d\theta} = -\int_0^{\bar{\eta}} g\left(\frac{p^*}{\beta_h\theta+\eta}\right) \frac{\beta_h p^*}{[\beta_h\theta+\eta]^2} dF_\eta(\eta)$$

[5] In general, changing the weights on θ and η means that the distribution of v, and therefore possibly the equilibrium price, will be different for houses of different β_h. One case where this does not happen is if house-specific and neighborhood-level shocks are drawn from the same distribution (i.e. $F_\theta = F_\eta$) and we only consider local deviations of β_h around $\beta_h = 1$.

[6] With more assumptions about distributions it is possible to make stronger statements. For instance, if $\varepsilon \sim U[a, 1]$, $\eta \sim U[0, \bar{\eta}]$, $\theta < \bar{\eta}$ and $\beta = 1$, then $\frac{d^2 \Pr[Sell|\theta,\beta]}{d\theta d\beta} < 0$.

Evaluating this expression for $\beta_h = 0$ and $\beta_h > 0$ gives the result.

□

Suppose now that owners of houses belong to one of two possible groups, long-tenure and short-tenure owners, with probabilities π_L and π_S. These two groups differ with respect to the conditional distribution of the ε shock, which we denote by G_L and G_S (the unconditional distribution is still G). The following result leads to Prediction 1.c.

Proposition 4. *Suppose that* $\frac{g_S(\varepsilon)}{G_S(\varepsilon)} \geq \frac{g_L(\varepsilon)}{G_L(\varepsilon)}$ *for every* $\varepsilon \geq \frac{p^*}{\bar{v}}$. *Then the proportion of long-tenure sellers in a neighborhood is increasing in* θ.

Proof. Conditional on house quality v, house supply from long-tenure and short-tenure owners, respectively, are

$$S_L(p^*, v) = \pi_L G_L\left(\frac{p^*}{v}\right) f(v) \quad \text{and} \quad S_S(p^*, v) = \pi_S G_S\left(\frac{p^*}{v}\right) f(v)$$

so the proportion of long-tenure owners among sellers of houses of quality v is

$$\pi_{L|Sell}(v) = \frac{\pi_L G_L\left(\frac{p^*}{v}\right)}{\pi_L G_L\left(\frac{p^*}{v}\right) + \pi_S G_S\left(\frac{p^*}{v}\right)}.$$

Taking the derivative with respect to v:

$$\begin{aligned}
\frac{d\pi_{L|Sell}(v)}{dv} &= \frac{-\pi_L g_L\left(\frac{p^*}{v}\right)\frac{p^*}{v^2}\left[\pi_L G_L\left(\frac{p^*}{v}\right) + \pi_S G_S\left(\frac{p^*}{v}\right)\right] + \left[\pi_L g_L\left(\frac{p^*}{v}\right)\frac{p^*}{v^2} + \pi_S g_S\left(\frac{p^*}{v}\right)\frac{p^*}{v^2}\right]\pi_L G_L\left(\frac{p^*}{v}\right)}{\left[\pi_L G_L\left(\frac{p^*}{v}\right) + \pi_S G_S\left(\frac{p^*}{v}\right)\right]^2} \\
&= \frac{-\pi_L g_L\left(\frac{p^*}{v}\right)\frac{p^*}{v^2}\left[\pi_S G_S\left(\frac{p^*}{v}\right)\right] + \left[\pi_S g_S\left(\frac{p^*}{v}\right)\frac{p^*}{v^2}\right]\pi_L G_L\left(\frac{p^*}{v}\right)}{\left[\pi_L G_L\left(\frac{p^*}{v}\right) + \pi_S G_S\left(\frac{p^*}{v}\right)\right]^2} \\
&= \frac{\pi_L \pi_S \frac{p^*}{v^2} G_L\left(\frac{p^*}{v}\right) G_S\left(\frac{p^*}{v}\right)}{\left[\pi_L G_L\left(\frac{p^*}{v}\right) + \pi_S G_S\left(\frac{p^*}{v}\right)\right]^2}\left[\frac{g_S\left(\frac{p^*}{v}\right)}{G_S\left(\frac{p^*}{v}\right)} - \frac{g_L\left(\frac{p^*}{v}\right)}{G_L\left(\frac{p^*}{v}\right)}\right] > 0. \qquad \text{(A.14)}
\end{aligned}$$

The proportion of long-tenure owners among sellers in a neighborhood of quality θ will be

$$\pi_{L|Sell}(\theta) = \int_0^{\bar{\eta}} \pi_{L|Sell}(\beta\theta + \eta)\, dF_\eta(\eta).$$

Taking the derivative with respect to θ and using (A.14):

$$\frac{d\pi_{L|Sell}(\theta)}{d\theta} = \int_0^{\bar{\eta}} \beta \left(\left.\frac{d\pi_{L|Sell}(v)}{dv}\right|_{v=\beta\theta+\eta}\right) dF_\eta(\eta) > 0.$$

□

Consider next the differential appreciation obtained by informed buyers over uninformed ones. Equations (A.3) and (A.4) imply that, unconditionally, this differential is positive. However, the following result underlies Prediction 4.

Proposition 5. *Assume $\beta\bar{\theta} > b$. Then, conditional on buying in a sufficiently good neighborhood, the expected value of houses bought by informed and uninformed buyers is the same.*

Proof. The expected house quality obtained by an informed buyer conditional on buying a house in a neighborhood of quality θ is

$$\bar{v}^I(\theta) = \frac{\int_b^{\bar{v}} vG\left(\frac{p^*}{v}\right) f_{v|\theta}(v)\,dv}{\int_b^{\bar{v}} G\left(\frac{p^*}{v}\right) f_{v|\theta}(v)\,dv}$$

For $\theta > \frac{b}{\beta}$, $f_{v|\theta}(v) = 0$ for all $v < b$ and therefore

$$\bar{v}^I(\theta) = \frac{\int_0^{\bar{v}} vG\left(\frac{p^*}{v}\right) f_{v|\theta}(v)\,dv}{\int_0^{\bar{v}} G\left(\frac{p^*}{v}\right) f_{v|\theta}(v)\,dv} = \bar{v}^U(\theta)$$

□

B. Data Appendix

B.1 Data Cleaning

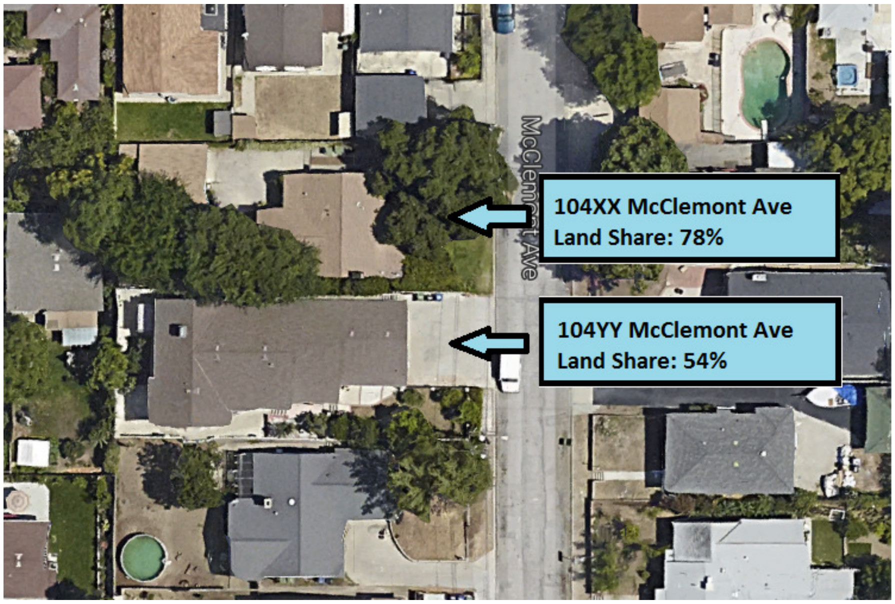
Arms-length Transactions: We identify all deeds that contain information about arms-length transactions in which both buyer and seller act in their best economic interest. This ensures that transaction prices reflect the market value of the property. The procedure to identify arms-length transactions follows Stroebel (2014). We include all deeds that are one of the following: "Grant Deed," "Condominium Deed," "Individual Deed," "Warranty Deed," "Joint Tenancy Deed," "Special Warranty Deed," "Limited Warranty Deed" and "Corporation Deed." This excludes intra-family transfers and foreclosures. We drop all observations that are not a Main Deed or only transfer partial interest in a property. We also drop properties with transaction prices of less than $25,000 and more than $10,000,000.

Death of Owner: We identify those repeat sales pairs for which we observe a death of the owners up to twelve months before the second sale ("forced moves"). The death of an owner is identified if either (i) the seller on a deed is classified as an "estate," "executor," "deceased" or "surviving joint owner," or (ii) if we observe one of the following: "Affidavit of Death of Joint Tenant" or "Executor's Deed."

B.2 Measuring Land Share

We calculate the share of each property's total value made up of land from data in the tax assessment records. In particular, tax assessors report a separate valuation of both the land and "improvements." Improvements include all assessable buildings and structures on the land. Figure A.1 shows an example of two neighboring homes in Los Angeles county. The two houses sit on identically-sized lots. The southern-most house, however, has a larger structure built on it, and thus the share of land in overall value as reported in the tax assessment records is lower.

Figure A.1: Land Share - Example



Note: Figure shows an example of the land share calculated for two properties in Los Angeles county.

B.3 Control Variables

Table A.1 shows summary statistics for the control variables used in the regressions. Most of these controls are not included linearly in the regression, but by splitting them into groups of values represented by dummy variables. This allows for a more flexible functional form. The results are not sensitive to the exact definition of groups.

House Characteristics: Building size is controlled for by adding dummy variables for 10 equally-sized groups capturing the different deciles of the size distribution. To control for the number of bedrooms and bathrooms, we add a dummy variable for each possible value. We construct the age of the property by subtracting the construction year of the house from the year of sale. Controls for building age are included by including four dummies for the age quartiles. We include an "investment property" dummy for properties that are identified as such in the assessor data.

Financing Characteristics: The loan-to-value (LTV) ratio is included by dummy variables for mortgages with an LTV $\leq$ 80%, between 80% and 90%, between 90% and 97%, and $>$ 97%. We also control for the duration of the mortgage, and whether it is a VA or FHA-insured mortgage.

Table A.1: Summary Statistics - Control Variables

	Mean	Standard Deviation	P10	P50	P90
Property Characteristics					
Building Area (sqft)	1,688	886	924	1,465	2,710
Bedrooms (#)	3.07	1.20	2	3	4
Bathrooms (#)	2.31	1.06	1	2	3
Age of Building (years)	36.8	23.7	6	38	72
Condo (binary)	0.26	0.44	0	0	1
Pool (binary)	0.24	0.43	0	0	1
AC (binary)	0.45	0.50	0	0	1
Investment Property (binary)	0.07	0.25	0	0	0
Financing Characteristics					
Loan-To-Value Ratio	0.84	0.14	0.70	0.80	0.99
Mortgage Duration (years)	29.5	3.50	30	30	30
VA Mortgage (binary)	0.01	0.10	0	0	0
FHA Mortgage (binary)	0.17	0.38	0	0	1

Note: The table shows summary statistics for the control variables included in the regressions.

C. Robustness Checks

In this section we present a number of important robustness checks to the empirical results in Section 3. We begin by showing that our measurement of informed sellers is robust to excluding matches to realtors with common last names, as well as to concerns that realtors might be inventory-holding dealers. We also show that our results are not driven by endogenous selection into repeat sales, and that the presence of "flippers" does not confound our measure of long-tenure sellers. Finally, we show that our results persist over subsequent ownership periods, suggesting that not all information is revealed by the time of resale.

C.1 Common names and real estate agent matching

A first concern with the results presented in Section 3.2 is that they might be distorted by the way we deal with common names when we match realtor licenses to sellers in the construction of the measure for "share informed sellers." In particular, as described in Section 2, we match a seller to have a realtor license whenever a realtor license was issued to a person with that name. For common names such as "James Smith" we might, therefore, erroneously identify some individuals to be informed. While such a mismeasurement should lead to attenuation bias but not systematically affect the interpretation of our findings, in this section we test the robustness of our approach to different ways of identifying informed sellers.

In particular, for our construction of the "share informed sellers" measure, we ignore realtor licenses issued to individuals with the 100 and 1,000 most common last names in the 1990 census, as reported by http://names.mongabay.com/data/1000.html. When we drop the 100 most common last names, the average of "share informed sellers" declines from 4.3% to 2.7%, with a conditional within zip code standard deviation of 2.1%. When we drop the 1,000 most common last names, the average of "share informed sellers" is 1.6%, with a conditional standard deviation of 1.5%.

The results when using this measure of "share informed sellers" in regression 2 are presented

Table A.2: Seller Composition and Returns - Robustness in Measuring "Share informed sellers"

	(1)	(2)	(3)	(4)	(5)	(6)
Share Informed Sellers	-4.737***	-11.11***	-15.62***	-4.571***	-8.817***	-11.93***
	(1.663)	(2.035)	(2.220)	(1.436)	(1.680)	(1.968)
Average Seller Land Share				-13.87***	-13.76***	-13.98***
				(1.876)	(1.854)	(1.875)
Share in Zip of Tenure > 3				6.018***	5.907***	5.910***
				(0.578)	(0.568)	(0.570)
Fixed Effects, Property and Financing Controls	✓	✓	✓	✓	✓	✓
Restriction (Drop common last names)		100	1,000		100	1,000
R-squared	0.636	0.637	0.637	0.659	0.659	0.659
$\bar{y}$	12.56	12.56	12.56	13.70	13.70	13.70
N	391,837	391,837	391,837	300,108	300,108	300,108

Note: The table shows results from regression 2. The dependent variable is the annualized capital gain of a property between two sequential arms-length sales. The seller composition variables are measured at the quarter × zip code level. All specifications include sales quarter pair fixed effects, zip code fixed effects, and control for characteristics of the property (property size, property age, property type, number of bedrooms and bathrooms, whether the property has a pool or air conditioning) and characteristics of the financing (mortgage type, mortgage duration and loan-to-value ratio). Standard errors are clustered at the initial quarter × zip code level. Columns (1) - (3) include sales pairs where the first sale was after June 1994, columns (4) - (6) include sales pairs where the first sale was after June 1997. In columns (2) and (5) we exclude "informed seller" matches with the 100 most common last names from the 1990 census; in columns (3) and (6) we exclude matches with the 1,000 most common last names. Significance levels: * (p <0.10), ** (p <0.05) and *** (p <0.01).

in Table A.2. In columns (1) and (4) we repeat results from Table 3, which uses our baseline methodology for matching realtor licenses to sellers. In columns (2) and (4), we drop realtor licenses issued to individuals with the 100 most common last names. The coefficient on "share informed sellers" increases significantly, consistent with this restriction reducing attenuation bias. A one conditional standard deviation increase in the share of informed sellers is associated with a 23 basis point decline in annualized appreciation of houses in that neighborhood. Similarly, when we drop realtor licenses to individuals with the 1,000 most common last names, the coefficient increases further. This evidence suggests that the interpretation of our baseline regressions in Table 3 is not biased in favor of finding evidence for asymmetric information by our approach of matching seller names to realtor licenses.

C.2 Real Estate Agents as Dealers with Inventory?

In many financial markets, dealers provide liquidity for a fee by holding inventory. In housing markets, most real estate agents act as brokers that do not take inventory risk. However, there might be a concern that at least some of the real estate agents whose trading we observe in the data could be acting as dealers, and that their buying-and-selling represents the provision of liquidity. This could confound our interpretation in two different places. First, it might confound the interpretation of the "share informed sellers" measure. If real estate agents are more likely to buy houses to refurbish and then resell them, then the "share informed sellers" might pick up the "share of selling flippers." Importantly, this will not be ruled out by excluding short ownership periods from our return regressions (see Appendix C.4). Second, liquidity-providing real estate agents could affect the interpretation of the "informed buyer" coefficient in Table 9, if the higher capital gains of real estate agents was to compensate them for the provision of liquidity through bearing inventory risk.

There are a number of reasons to think that this might not be a particularly problematic concern. First, as discussed above, realtors generally play the role of brokers, and not that of dealers. Consistent with this, for all properties bought before the year 2000, the time to the next sale for properties bought by real estate agents is only slightly smaller than that for properties not bought by real estate agents (2,042 days vs. 2,098 days).[7] This long average holding period for properties bought by real estate agents is not suggestive of them playing the role of a market-making dealer. Similarly, of all properties owned by real estate agents in 2009, 78.5% were owner-occupied, suggesting that we primarily observe realtors purchasing properties to live in. The owner-occupancy rate for properties not owned by realtors was only slightly higher, at 82.0%.

Nevertheless, in this section we present a number of robustness checks to confirm that our results are not driven by a minority of realtors playing the role of dealers in this market. To do this, in columns (1) and (2) of Table A.3 we only measure the share of informed sellers (realtors) among the group of sellers that have been living in the house for at least three years; any realtor that had initially purchased the house in the role of a liquidity-providing dealer would have resold the property already.[8] The estimated coefficient for the relationship between the share of informed sellers and the subsequent house price appreciation are identical, suggesting that realtor-dealers do not confound our estimates.

Similarly, in columns (3) - (5) we restrict the sample to transaction pairs with at least 3 years between the transactions. Again, the idea is that liquidity-providing realtor-dealers would not generally hold on to properties for such a long period, so we are capturing the outperformance of real estate agents that bought to own. The annualized outperformance of real estate agents in this sample of transaction-pairs is very similar to that in the full sample, as is the interaction with the share of informed sellers, shown in columns (4) and (5). These findings suggest that real estate agents are able to pick properties that outperform, and that this is particularly so in markets with a high share of informed sellers. We conclude that none of these results are driven by realtors acting as dealers, but are related to their superior information about neighborhood characteristics.

[7] We restrict to houses bought before the year 2000 to deal with the right-censoring of holding periods. However, the results are robust to considering other purchase windows.

[8] The results are robust to measuring the realtor shares among sellers who have lived either at least 2 years, or at least 4 years in the property.

Table A.3: Remove "Possible Dealers" from Realtors

	(1)	(2)	(3)	(4)	(5)
Share Informed Sellers (Tenure > 3 years)	-4.602*** (1.559)	-4.344*** (1.375)			-4.803*** (1.535)
Average Land Share		-13.91*** (1.871)			
Share in Zip of Tenure > 3		6.012*** (0.576)			
Real Estate Prof.			0.695*** (0.049)	0.506*** (0.076)	0.514*** (0.074)
Share Informed sellers				-4.836*** (1.634)	
Real Estate Prof. × Share informed sellers				4.377*** (1.553)	
Real Estate Prof. × Share informed sellers (Tenure > 3 years)					4.171*** (1.482)
Fixed Effects, Property and Financing Controls	✓	✓	✓	✓	✓
R-squared	0.657	0.659	0.779	0.779	0.779
$\bar{y}$	13.70	13.70	10.81	10.81	10.81
N	300,108	300,108	177,610	177,610	177,610

Note: The table shows results from regression 2 in columns (1) and (2), and from regression 6 in columns (3) - (5). The dependent variable is the annualized capital gain of a property between two sequential arms-length sales. The sample includes sales pairs where the first sale was after June 1997. The seller composition variables are measured at the quarter × zip code level. In columns (1), (2) and (5) we measure the share of informed sellers only among the subset of sellers that have lived in the house for at least 3 years. In columns (3) - (5) we only include transaction pairs with at least 3 years between the transactions. All specifications include sales quarter pair fixed effects, zip code fixed effects, and control for characteristics of the property (property size, property age, property type, number of bedrooms and bathrooms, whether the property has a pool or air conditioning) and characteristics of the financing (mortgage type, mortgage duration and loan-to-value ratio). Standard errors are clustered at the initial quarter × zip code level. Significance levels: * (p <0.10), ** (p <0.05) and *** (p <0.01).

C.3 Selection into observing repeat sales

One might be worried that the subsample of houses for which we observe a resale is not representative of all homes in a particular neighborhood, and that such a selection might lead us to incorrectly measure the true correlation between seller composition and the average capital gain of homes in the neighborhood.[9] To address such concerns, in Table A.4 we show results from regression 2 similar to Table 3, but restrict the sample to sales pairs where the second sale is precipitated by a plausibly exogenous event. In particular, we only look at those repeat sales pairs where we

[9]For example, this might be a problem if owners who experience high idiosyncratic capital gains on their house are more likely to sell and thus enter our sample more often than other owners in the same neighborhood. If seller composition affected the probability of these idiosyncratic capital gains events, this would lead us to overestimate the correlation between seller composition and average house price movements.

observe the death of the original owners in the 12 months preceding the resale.[10] We argue that such sales are more plausibly prompted by the observed death than by other factors such as the value of the house. The results show that the correlation between neighborhood seller composition and subsequent capital gains is of the same magnitude in the sample of forced moves as it is in the entire sample, and even somewhat larger for the share of informed sellers. This suggests that selection into observing repeat sales does not significantly bias our estimates.

Table A.4: Effect of Seller Composition on Capital Gains - Forced Moves

	(1)	(2)	(3)	(4)
Share Informed Sellers	-9.481***			-7.292*
	(3.452)			(4.287)
Average Seller Land Share		-19.92***		-18.68***
		(3.123)		(3.885)
Share in Zip of Tenure > 3			5.819***	4.537***
			(1.705)	(1.666)
Fixed Effects, Property and Financing Controls	✓	✓	✓	✓
R-squared	0.601	0.602	0.622	0.623
$\bar{y}$	13.97	13.97	15.14	15.14
N	17,613	17,613	13,456	13,456

Note: The table shows results from regression 2 for those transactions where the resale was preceded in the 12 months before by a death of the owner. The dependent variable is the annualized capital gain of a property between two sequential arms-length sales. All specifications include sales quarter pair fixed effects and zip code fixed effects and control for characteristics of the property (property size, property age, property type, number of bedrooms and bathrooms, whether the property has a pool or air conditioning) and characteristics of the financing (mortgage type, mortgage duration and loan-to-value ratio). Standard errors are clustered at the initial quarter × zip code level. Columns (1) - (2) include sales pairs where the first sale was after June 1994, column (3) includes sales pairs where the first sale was after June 1997. Significance levels: * (p <0.10), ** (p <0.05) and *** (p <0.01).

C.4 Presence of "Flippers"

One concern might be that our measure of average seller tenure does not, in fact, pick up owners that are moving out of the neighborhood, but instead captures the share of "flippers" among the sellers, who buy houses to resell them quickly at a profit, often after substantial remodeling or renovation. This could bias our results in either direction, depending on what kinds of neighborhoods tend to attract more flippers. If flippers are more active in overpriced neighborhoods (perhaps because they are trying to time the housing market), then this could drive the correlation we observe in the data: high flipper activity would show up as a larger share of sellers that have a short ownership-tenure and would predict low subsequent appreciation.

To rule out that this is the main driver of the observed correlation, we identify a set of transactors that we classify as flippers, and then repeat the analysis above by only calculating the average tenure among those sellers not identified as flippers. Similar to Bayer, Geissler, and Roberts (2011), we use the fact that the deeds data records the name of buyers and sellers to classify transactors as

[10] These events can be identified in the deeds data as described in Appendix B.1.

flippers. We apply three classification rules. Our first two rules classify an individual as a flipper if someone with that name has engaged in at least 3 transactions over the sample period, with more than 30% (40%) of them being bought and resold within 2 years. Our third rule excludes all transactors that are classified as companies, since some flippers might buy and sell homes through incorporated entities. Table A.5 shows that the results are very robust to only considering the average tenure of sellers that are not classified as flippers.

Table A.5: Remove Possible Flippers from Tenure

	(1)	(2)	(3)	(4)
Share in Zip of Tenure > 3	6.863***	5.374***	5.393***	6.596***
	(0.622)	(0.521)	(0.526)	(0.634)
Restriction	None	> 2 Trans. > 30% within 2y	> 2 Trans. > 40% within 2y	No companies
Fixed Effects, Property and Financing Controls	✓	✓	✓	✓
R-squared	0.658	0.658	0.658	0.658
$\bar{y}$	13.70	13.70	13.70	13.70
N	300,108	300,108	300,108	300,108

Note: The table shows results from regression 2. The dependent variable is the annualized capital gain of a property between two sequential arms-length sales. In column (1) tenure is measured among all sellers, in column (2) we exclude sellers that have more than two transactions and at least 30% of them are resold within 2 years, in column (3) we exclude sellers with at least two transactions of which at least 40% are resold within 2 years. In column (4) we exclude all sales by sellers identified as companies. All specifications include sales quarter pair fixed effects, zip code fixed effects, and control for characteristics of the property (property size, property age, property type, number of bedrooms and bathrooms, whether the property has a pool or air conditioning) and characteristics of the financing (mortgage type, mortgage duration and loan-to-value ratio). Standard errors are clustered at the initial quarter × zip code level. Significance levels: * (p <0.10), ** (p <0.05) and *** (p <0.01).

C.5 Ownership Period of Second Buyer

The framework in Section 1 and the model in Appendix A have only two periods. In reality, information is likely to be revealed gradually over time. If so, then one should expect some of it to be revealed only after the buyer has resold the house; the information would thus affect the appreciation experienced by subsequent owners. To test for this, we need to observe at least three arms-length transactions of the house. We calculate the appreciation between the last two sales, as shown in Figure A.2, and determine to what degree this is predicted by the seller composition at the time of the first sale. In other words, we run regression A.15, where q_1, q_2 and q_3 represent the calendar quarters of the first, second and third sale.

$$CapitalGain_{i,n,q_2,q_3} = \alpha + \beta_1 SellerComposition_{n,q_1} + \chi_{q_1} + \phi_{q_2,q_3} + \xi_n + X_i'\beta_2 + \epsilon_i \tag{A.15}$$

Table A.6 shows the results separately for samples where we allow up to fours years and up to six years between sale one and sale two. For longer time horizons between the first and second

Figure A.2: Measures of Second Buyer's capital gain

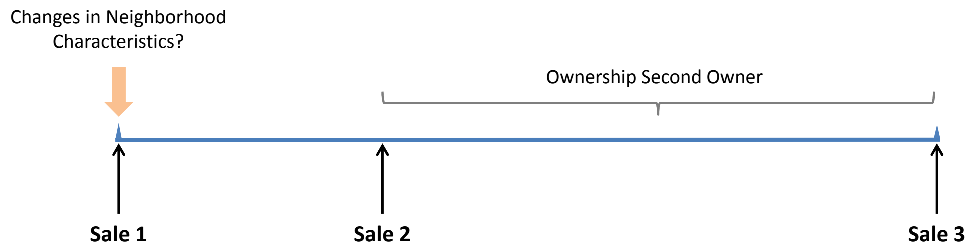


Table A.6: Effect of Seller Composition on Capital Gains during Second Ownership Period

	(1)	(2)	(3)	(4)	(5)	(6)
Share Informed Sellers	-8.645***	-9.248**				
	(2.896)	(3.361)				
Land Share			-21.67***	-19.05***		
			(4.521)	(3.699)		
Share in Zip of Tenure > 3					3.613**	2.669**
					(1.237)	(1.088)
Fixed Effects, Property and Financing Controls	✓	✓	✓	✓	✓	✓
Max. Time between Sales 1 & 2	4 Years	6 Years	4 Years	6 Years	4 Years	6 Years
R-squared	0.557	0.563	0.558	0.563	0.597	0.607
$\bar{y}$	10.99	11.17	10.99	11.17	10.83	10.51
N	58,747	82,996	58,747	82,996	48,161	63,521

Note: The table shows results from regression A.15. The dependent variable is the annualized capital gain of a property between the two repeat sales. The seller composition variables are measured at the quarter × zip code level. All specifications include fixed effects for the sales quarter pair, the quarter of initial sale and the zip code. Standard errors are clustered at the initial quarter × zip code level. All specifications control for characteristics of the property (property size, property age, property type, number of bedrooms and bathrooms, whether the property has a pool or air conditioning) and characteristics of the financing (mortgage type, mortgage duration and loan-to-value ratio). Columns (3) and (6) include sales pairs where the first sale was after June 1997, all other columns include sales pairs where the first sale was after June 1994. Significance levels: * (p <0.10), ** (p <0.05) and *** (p <0.01).

sale more of the initially unobservable neighborhood characteristics will have been revealed, which leaves less scope for initial seller composition to predict additional differential capital gain.

We can see that following an increase in the share of informed sellers, houses in that neighborhood continue to underperform, even during the ownership period of subsequent owners. Similar effects can be seen for changes in our other two measures of seller composition, the average land share of sold homes and the average tenure of sellers.