WestVirginiaUniversity
THE RESEARCH REPOSITORY @ WVU

Economics Faculty Working Papers Series

Economics

2018

Does the Asset Pricing Premium Reflect Asymmetric or Incomplete Information?

Crocker H. Liu
Cornell University, chl62@cornell.edu

Adam Nowak
West Virginia University, Adam.Nowak@mail.wvu.edu

Patrick S. Smith
San Diego State University, patrick.smith@sdsu.edu

Digital Commons Citation

Liu, Crocker H.; Nowak, Adam; and Smith, Patrick S., "Does the Asset Pricing Premium Reflect Asymmetric or Incomplete Information?" (2018). *Economics Faculty Working Papers Series*. 5.
https://researchrepository.wvu.edu/econ_working-papers/5

Department of Economics

Working Paper Series

Does the Asset Pricing Premium Reflect Asymmetric or Incomplete Information

Crocker H. Liu

Adam D. Nowak

Patrick S. Smith

Working Paper No. 18-06

Does the Asset Pricing Premium Reflect Asymmetric or Incomplete Information?

Crocker H. Liu [*] Adam D. Nowak [†] Patrick S. Smith [‡]

April 26, 2018

Abstract

We develop a framework for using text as data in asset pricing models. We use the framework to test whether real estate agents exploit their informational advantage to sell properties they own for a premium. Consistent with the previous literature, baseline estimates that exclude textual information indicate agents sell their own house at a 3 to 4 percent premium in both Phoenix, AZ and Atlanta, GA. However, this premium dissipates when textual information is included. The results suggest that the baseline estimates suffer from an omitted variable bias, which previous studies incorrectly ascribe to market distortions associated with asymmetric information.

JEL Codes: D82 - Asymmetric and Private Information; Mechanism Design, G14 - Information and Market Efficiency; Event Studies; Insider Trading, R00 (General Urban, Rural, and Real Estate Economics)

Key Words: Asset pricing, Asymmetric Information, Omitted variable bias, Textual analysis

[*] Cornell University, SC Johnson College of Business; Email: chl62@cornell.edu

[†] West Virginia University, College of Business & Economics; Email: adam.d.nowak@gmail.com

[‡] San Diego State University, Fowler College of Business; Email: patrick.smith@sdsu.edu

[1] We thank Doreen Longfellow and the Phoenix Board of Realtors for providing access to the Arizona MLS data and Brian Chew for providing access to the Georgia MLS data. We also thank Bob Connolly, Tian Liu, Jenny Schuetz, and Andrey Ukhov for helpful comments. Errors are the responsibility of the authors.

I, Crocker H. Liu, have nothing to disclose.

Crocker H. Liu

I, Adam Nowak, have nothing to disclose

I, Patrick S. Smith, have nothing to disclose.

Patrick Smith

1 Introduction

Information encoded in text can provide a rich complement to more traditional forms of data. However, its high-dimensional nature precludes statistical inference using standard econometric techniques (Gentzkow et al., 2017). The purpose of this study is to develop a framework for using text as data in asset pricing models. We use a LASSO-type double-selection method to select relevant words and phrases (tokens) from unstructured text and then include these tokens directly in the asset pricing model.[1] We show that including the textual information expands the researcher's information set to better reflect the information set of the market participants and, in doing so, has a significant effect on pricing estimates.

Although the framework we present is applicable to any asset class, we demonstrate its efficacy in the pricing of residential real estate. More specifically, we test for asymmetric information among market participants in residential real estate. Direct tests of asymmetric information present identification challenges given the difficulty in observing and measuring the heterogeneity of information among market participants (Kelly and Ljungqvist, 2012). The few real estate studies that do test for asymmetric information attempt to address the identification problems using indirect information variables that identify market participants who likely have an informational advantage. For example, Garmaise and Moskowitz (2003) use professionally brokered transactions, Rutherford et al. (2005) and Levitt and Syverson (2008) use agent-owned transactions, and Kurlat and Stroebel (2015) use three measures of seller composition. The use of indirect information variables is necessary because researchers' have incomplete information which, by definition, makes it impossible to measure the heterogeneity of information among market participants. A natural concern is that the coefficient on the indirect information variable is biased when the indirect information variables are correlated with omitted variables.

The extant literature focuses on differential information between market participants even though the studies cannot determine whether an informational advantage (i) actually

[1] LASSO is an acronym for least absolute shrinkage and selection operator.

exists, (ii) can be exploited for a financial gain, or (iii) the estimated mispricing effect is related to an omitted variable bias. Recognizing these limitations, we focus on incorporating textual information into the asset pricing model to more closely align the information set in the asset pricing model with the information set of the market participants involved in the transaction. We show that textual information offers a rich complement to structured data and that not including it in the asset pricing model results in an omitted variable bias.

Textual analysis is not a new concept in finance and economics.[2] However, several factors have prevented its widespread use. Most notably, numeric representations of text are not readily available to the researcher. To overcome this obstacle researchers typically specify, ex-ante, a dictionary of words that are associated with a given topic or sentiment and then map the text into a numeric index using the dictionary. However, as Loughran and Mcdonald (2011) note dictionaries are not universally applicable because words that imply a positive tone in general can imply a neutral or negative tone in a specific context. For example, *charming* and *cozy* are positive words in the widely used Harvard IV-4 Dictionary, but they are euphemisms for smaller, old-fashioned houses in a real estate context.

We create a real estate specific dictionary using the remarks section of the multiple listing service (MLS). The listing agent, who is the only professional to enter and evaluate the interior of the house, uses the remarks section to provide a description of the property that complements the information reported in the standard MLS data fields.[3] Given its limited length, listing agents use the remarks section to highlight important information such as the condition and quality of the property, motivation (if any) of the seller, purchase

[2] Tetlock (2007) and Loughran and Mcdonald (2011) are early examples of textual analysis in the finance literature. The bulk of the literature focuses on the impact of textual information on equity valuations. However, more recent studies have used textual information to examine topics such as financial constraints (Hoberg and Maksimovic 2014; Bodnaruk et al. 2015; Buehlmaier and Whited 2018). See Loughran and McDonald (2016) for a recent survey of textual analysis in both finance and accounting.

[3] Even appraisers who play a critical role in lenders' underwriting decisions rely on MLS listing information. For example, Young (2012) states that "today's appraisers are required to rate property conditions of both subject properties and comparables using a numerical scale from C1 to C6. Where do they get the information needed to make these ratings? Typically from the information that is provided in the MLS listing by the listing agent, including photos, remarks, and descriptions of physical features found in the various fields for listing input. As appraisers rely on the information found in the MLS, the more descriptive and accurate that information is, the better appraisal reports can be."

incentives, and/or neighborhood amenities that are not clearly conveyed in other areas of the listing. We show that the remarks include both positive and negative information about the time-varying and time-invariant features of the house and neighborhood. We also show that the remarks include information that is correlated with the indirect information variables (e.g. agent-owned transactions) used in previous research, so estimates of the mispricing effects of asymmetric information are likely biased in the extant literature.[4]

Creating a real estate specific dictionary is complicated by the fact that the remarks section of the MLS contains tens of thousands of tokens, thereby rendering conventional variable selection methods computationally infeasible. Nowak and Smith (2017) overcome this issue using a LASSO-type single-selection method to create a dictionary that strongly predicts property prices. Single-selection variable selection methods, including the LASSO, select variables based on their ability to predict a single dependent variable; because of this, variables that have a low predictive power but are correlated with the variable of interest may be omitted from the pricing model, resulting in an omitted variable bias. In contrast, double-selection variable selection methods identify variables that are strong predictors of the dependent variable and/or the variable of interest. The double-selection procedure we use, which relies heavily on Belloni et al. (2014), is computationally feasible and allows for valid statistical inference on the coefficient of interest.

The double-selection method selects over 1,000 tokens for our real estate specific dictionary. In contrast, Levitt and Syverson (2008) use a dictionary of 61 words and phrases. This difference in the number of tokens reflects the value of machine learning and other high-dimensional methods rather than the limitation of pre-defined dictionaries. This is not too surprising as King et al. (2017) find that humans perform poorly when creating dictionaries from scratch, yet perform well when associating words.

After creating the real estate specific dictionary, we use it to examine agent-owned trans-

[4] A textual description of the property was not available in every study. Garmaise and Moskowitz (2003) used commercial real estate data and Kurlat and Stroebel (2015) used tax assessor data, so they did not have access to the textual information used in this study. Rutherford et al. (2005) and Levitt and Syverson (2008) used MLS data, so they had access to the textual information.

actions in a residential real estate setting. Agent-owned transactions offer a rare example of a clean identification strategy. When real estate agents list a property they own on the MLS, they are required by law to notify potential buyers that the owner of the property holds a real estate license.[5] The seminal studies by Rutherford et al. (2005) and Levitt and Syverson (2008) argue that real estate agents have better information than their clients and exploit their informational advantage to sell their own house for a higher price than a comparable client-owned house. However, both studies also note that agent-owned houses may systematically differ from those of their clients and report their findings with a caveat that the agent-owned estimates may suffer from an omitted variable bias. For example, Rutherford et al. (2005) note "another possible explanation is that owner-agents initially buy higher quality properties" while Levitt and Syverson (2008) state "a particular concern. . . is that agents live in houses that are especially attractive along dimensions that are difficult to observe or quantify." Although we disagree that these dimensions are difficult to observe, we certainly agree that this information is difficult to quantify. In any event, we address these concerns by including salient textual information from the remarks in the asset pricing model.

Sans textual information, we estimate a 3 to 4 percent premium for agent-owned transactions in Atlanta, Georgia and Phoenix, Arizona. These naive estimates are similar to those reported in Rutherford et al. (2005) and Levitt and Syverson (2008). However, after we incorporate the textual information the premium drops to 1.7 percent in Phoenix and is no longer statistically significant in Atlanta. We also find that the agent-owned estimate is statistically insignificant in Phoenix during the pre-boom (2000-2003) and bust (2007-2009) subperiods using a restricted subsample that limits the variability in location and physical characteristics.

Our empirical results have several important implications. Contrary to previous studies,

[5] For example, Rule 520-1-.09 (8) of Georgia's Administrative Code states that "A licensee shall not advertise to sell, buy, exchange, rent, or lease real estate in a manner indicating that the offer to sell, buy, exchange, rent, or lease such real estate is being made by a private party not licensed by the Commission."

we find that agents do not necessarily sell their own house for more than comparable client-owned houses. This suggests that agents do not use their informational advantage to exploit the principals they represent. We also find that including the textual information from the remarks in the asset pricing model explains a large portion of the naive agent-owned premium that previous studies attributed to asymmetric information. These findings highlight the difficulty in testing for information asymmetry using incomplete information.

The plan of our analysis is as follows. Section 2 describes the framework and methodology that we employ while Section 3 describes the data employed in the empirical analysis. Section 4 presents our results inclusive of robustness tests and Section 5 concludes.

2 Theory and Estimation

2.1 Omitted Variable Bias in Asset Pricing

Implicitly, a home buyer and seller negotiate a sales price based on a set of property attributes that are common knowledge and observable to all parties, X. We use the term *property-observable* when referring to this set of attributes. X includes both objective (e.g. square feet living area, number of bathrooms, and age) and subjective (e.g. condition, quality and character) attributes of the house and neighborhood. Home buyers learn about X by visiting the property, reviewing property improvement and tax records, hiring a professional home inspector, and consulting with their real estate agent.[6] In this sense, X is fully revealed to both the buyer and seller, so there is no information asymmetry between the market participants.[7]

[6] Sales price is also determined by buyer-specific or seller-specific attributes, including expectations and motivation, that are unobserved by the other party. Of course, the home buyer has insight into the home seller's motivation if the information is included in the agent's remark (e.g. an agent may note that the seller is "highly motivated").

[7] Relative to real estate agents, home buyers are infrequent, uninformed market participants who transact in the housing market, on average, once every ten years. Thus, we recognize that the seller likely has an information advantage if they are a real estate agent (i.e. agent-owned transactions). However, because we have incomplete information we cannot estimate the magnitude of the agents' information advantage or test whether the advantage can be exploited for financial gain. For these reasons we focus on incorporating the

The term observable often takes on a different meaning in academic research. First, researchers may use the term observable when referring to the set of attributes available in the data set, X^D. We use the term *data-observable* when referring to X^D. Because researchers work with incomplete information, the term data-observable generally refers to a subset of the property's property-observable attributes, $X^D \subseteq X$. For example, a property might have an unpleasant view that is not recorded anywhere in the data set. However, this view is easily observed when visiting the property.

Second, researchers may use the term observable when referring to the set of attributes used as explanatory variables in a hedonic model, X^M. We use the term *model-observable* when referring to X^M. The model-observable variables are often a subset of the aforementioned data-observable attributes, $X^M \subseteq X^D \subseteq X$. For example, although the public remarks section is available in most MLS data sets and is data-observable, the information contained in the remarks is generally not included in the hedonic pricing model.

In this study, we focus on two types of omitted variables: data-omitted and model-omitted. Data-omitted variables refer to the set of property attributes that are property-observable, but not data-observable. The data-omitted variables can be written as $X \setminus X^D$. In contrast, model-omitted variables refer to the set of property attributes that are (i) data-observable, but not model-observable and (ii) contribute to the price of the property in a non-trivial manner. When every variable in X^D contributes to the price of the property, the model-omitted variables can be written as $X^D \setminus X^M$.

Data-omitted variable bias can result from measurement difficulties. Two of the most difficult to measure attributes in real estate are the condition and the quality of the house. Condition refers to a time-varying measure of the house's maintenance and upkeep, and quality refers to a time-invariant measure of the workmanship and materials used in its construction. Researchers have long recognized that condition and quality are likely correlated with variables of interest. When this occurs, the coefficient estimate for the variable of inter-

textual information that is available to all market participants in the asset pricing model.

est suffers from an omitted variable bias. Thus, if condition and quality are correlated with agent-owned houses, the estimated price effect for agent-owned transactions will be upward (downward) biased if agent-owned houses are, on average, in better (worse) condition and/or of higher (lower) quality.

Data-omitted variable bias can also result from data collection limitations. For example, it is easy to measure the number of fireplaces in a house, identify if the kitchen was remodeled, or determine if the property has premium landscaping. However, the cost of measuring and recording these attributes can be prohibitively expensive when the number of properties is large, the attributes are time-varying, and resources are finite. For this reason, many data sets maintained at the county or municipality level include a limited set of property attributes, many of which are time-invariant or nearly time-invariant.

Unlike county or municipality authorities, real estate agents visit and observe the property in person prior to listing it for sale in the MLS. As a result, data sets that real estate professionals maintain include a more comprehensive set of time-varying and time-invariant property-observable attributes. For example, real estate agents indicate whether recent capital improvements, such as a remodeled kitchen, have been made to the property. If the researcher does not observe these capital improvements (data-omitted) and the capital improvements are correlated with agent-owned properties, then the agent-owned coefficient estimate suffers a data-omitted variable bias. If the researcher does observe the capital improvements but does not include this information in the model (model-omitted), then the agent-owned coefficient will suffer a model-omitted variable bias. In either event, if agent-owned properties are more likely to have capital improvements, the agent-owned coefficient estimate is upward biased.

To eliminate or at least mitigate model-omitted variable bias, we incorporate textual information provided by the listing agent in the remarks section of the MLS. The information we incorporate is present in most, if not all, MLS data sets and is therefore data-observable. The unstructured nature of the text does not immediately lend itself to a form that can be readily incorporated in a hedonic pricing model; for this reason, information in the remarks is almost always model-omitted. The following sections describe a practical and flexible approach to incorporate textual information into a hedonic pricing model.

2.2 Textual Analysis and Agent Remarks

Considerable variation in sales price often exists among properties that are identical in terms of listing period, location, bedrooms and bathrooms. Take, for example, the transactions in Table 1 which includes two three bedroom, two bathroom houses that are approximately the same size, located in the same census tract, and sold within two weeks of each other. Despite their structural and locational similarities, their sales price varies considerably. Notice that information in the remarks can be used to explain some of the within group sales price variation. For example, the more expensive house has a *fully remodeled* bath. In contrast, the less expensive house is an *opportunity for investment* that *needs cosmetic work*. This simple example highlights the property-observable textual information contained in the remarks and the need to include it in the asset pricing model.

Although the remarks contain textual information about the property, it is not immediately apparent how the researcher can use the text as data in a hedonic pricing model. One approach is to use one or more pre-specified topic dictionaries to represent the remarks numerically. Topic dictionaries for positivity, negativity, and readability include words commonly associated with their respective subject matter. An index for a given text can be computed as a weighted average of the counts of words present in the text that belong to one or more topic dictionaries. For example, a simple sentiment index can be calculated as the difference in the fraction of positive and negative words in a text. The index can then be included in standard regression models. When using this approach, it is important to use a dictionary that is suitable to the task at hand. That is, generic or off-the-shelf dictionaries may provide misleading results when applied to a finance or real estate setting (Loughran and Mcdonald, 2011).

In practice, creating a topic dictionary from scratch requires significant up-front costs and is likely to omit some relevant words. Noting this, we are interested in creating a *sufficient dictionary*. A dictionary is sufficient if it includes a set of words and phrases that are sufficient for valid statistical inference on the coefficient of interest. Words or phrases not included in the sufficient dictionary may be associated with the variable of interest, but the additional information they convey is not necessary for valid statistical inference on the coefficient of interest.

Incorporating textual data into a regression model presents several challenges. First, incorporating information in the remarks requires a numeric representation of the information. Using a common approach in the textual analysis literature, we create indicator variables based on the presence of a given *token* in the remarks. Tokens refer to single words (*unigrams*), two-word phrases (*bigrams*), or a combination of words and phrases (*flex-grams*).[8] Creating indicator variables in this way treats the remarks as an unordered collection of tokens which is commonly referred to as the *bag-of-words* approach. Of course, the use of bigrams and flex-grams allows the researcher to treat sets of consecutive words as a single token.

Second, remarks are entered manually so they may contain errors or idiosyncrasies. To mitigate the effects of misspellings, we use an open-source spell checking software to identify and correct spelling mistakes. We also run a depluralization process that converts every word to its singular form. In unreported results, we also consider a stemming algorithm. The results reported are not sensitive to the inclusion of plurals or unstemmed words. Recognizing that some of the textual information in the remarks overlaps, we also remove tokens that are redundant for the variable of interest.[9]

[8] The flex-gram tokenization approach uses a collection of n-grams for various n. Intuitively, flex-grams identify which words or phrases are constituent parts of larger phrases.

[9] We use the **hunspell** package to check for misspellings. Documentation is available at https://cran.r-project.org/web/packages/hunspell/hunspell.pdf. We use Porter's word stemming algorithm available in the **SnowballC** package. Documentation is available at https://cran.r-project.org/web/packages/SnowballC/SnowballC.pdf. We provide a step-by-step overview of the tokenization process and removal of redundant tokens in Section B of the internet appendix.

Third, the number of unique tokens in the data can be large. Although limited to descriptions of single-family houses, the remarks include more than 50,000 unique words when the Phoenix and Atlanta data sets are combined. Since not all tokens are relevant for asset pricing, it is standard practice in the textual analysis literature to eliminate tokens based on information content and frequency. As such, we remove a set of frequent words that convey negligible information known as *stop words* (i.e. *a, an, the, for, and, but*, etc.). Although we use recent techniques developed for high-dimensional data, it is still necessary to restrict the number of tokens to a manageable dimension. As such, we drop all but the 2,000 most frequent tokens. The remaining set of tokens define our *candidate* set of tokens. We also experimented with sets of 3,000 or 5,000 candidate tokens and found the resulting dictionaries and parameter estimates are comparable to the dictionary and parameter estimates when using 2,000 candidate tokens. The variable selection process, which we describe in the next section, is performed on the candidate set to create our dictionary.

2.3 Variable Selection with High-Dimensional Data

The procedure for creating the dictionary is quite general and can be applied in any setting where textual information may be used to augment conventional regression methods. The resulting dictionary is sufficient for valid asymptotic inference on a parameter of interest, τ, associated with a binary variable of interest, d, here an indicator for agent-owned transactions. In the context of a hedonic pricing model, the parameter τ is understood to be the expected difference in log price between agent-owned transactions and non-agent-owned transactions.

Beginning with the candidate set of $K = 2,000$ tokens, a dictionary can be identified by performing variable selection on the K tokens. Conventional likelihood based procedures such as AIC or BIC can be used to identify $S_p \subseteq \{1, ..., K\}$ as the set of $Q_p \ll K$ tokens that are strong predictors of p. However, these methods have several drawbacks when K is large. First, identifying which tokens are strong predictors using AIC or BIC is computationally infeasible as these methods require more than $2^{K=2,000}$ separate least-squares coefficient estimates.[10] Second, although AIC and BIC select variables that are strong predictors of p, a model-omitted variable bias associated with τ may still remain when the complement of tokens not in S_p are strong predictors of d.

Noting this computational infeasibility, one approach is to use the least absolute shrinkage and selection operator known as LASSO to select tokens (Tibshirani, 1996). The LASSO is a penalized regression where an ℓ_1 penalty is placed on the coefficients. The shape of this penalty yields a coefficient estimate with many elements equal to 0. Vectors with many elements equal to 0 are known as *sparse* vectors. By setting some coefficient estimates equal to 0, the LASSO performs both variable selection and coefficient estimation. By identifying which tokens are strong predictors of p, a dictionary from a set of candidate tokens is built using LASSO. More importantly, the penalized regression is a convex optimization problem that is computationally feasible even when $K = 2,000$ candidate tokens are considered.

However, using LASSO to select Q_p tokens based on their ability to predict p does not explicitly control for a model-omitted variable bias associated with τ. Noting this, Belloni et al. (2014) describe a *double-selection* procedure to identify additional tokens that can mitigate model-omitted variable bias. The procedure identifies an additional set $S_d \subseteq \{1, ..., K\}$ of Q_d tokens that may not be strong predictors of p but are strong predictors of d. The union $S_2 = S_p \cup S_d$ is a set of Q_2 tokens that are either strong predictors of p or strong predictors of d. These Q_2 tokens can then be used as controls in a second stage regression known as *post double-selection* estimation (Belloni et al., 2014).

Of course, variable selection methods are prone to variable selection errors in which S_2 may not include all predictors of p or d. However, the double-selection procedure described above identifies strong predictors of either p or d. Any tokens not in S_2 are at most mildly associated with p or d, and their omission does not lead to a significant model-omitted variable bias. In this way, the dictionary of tokens in S_2 is sufficient for valid asymptotic

[10] For comparison, there are approximately 10^{14} cells in the human body and 10^{80} atoms in the universe.

inference on τ; alternatively, the dictionary is robust to variable selection errors. However, because S_2 may omit some tokens that are weak predictors of p or d, the dictionary defined by S_2 is not a complete dictionary but rather a sufficient dictionary.

2.4 Asset Pricing Model and Double-Selection

The price of house n at time t, p_{nt}, can be written as

$$p_{nt} = x_{nt}\beta + d_{nt}\tau + \mu_n + \psi_{nt} + v_{nt} \tag{1}$$

In Equation 1, x_{nt} is a vector of time-invariant and time-varying variables selected by the researcher, β is a vector of implicit prices, d_{nt} is an indicator variable for agent-owned sales transactions, τ is the price effect associated with agent-owned transactons, $\mu_n(\psi_{nt})$ is a time-varying (time-invariant) effect, and v_{nt} is a zero-mean error term uncorrelated with any variables x_{nt}, d_{nt}, μ_n, or ψ_{nt}. The μ_n and ψ_{nt} effects include both the data-omitted and model-omitted variables. In the analysis below, x_{nt} also includes dummy variables for the time of sale (quarter by year), location (zip code or census tract), and time-location interactions.

Equation 1 can be estimated using least-squares. The estimate $\hat{\tau}$ is upward biased if $0 < \mathbb{E}[d_{nt}(\mu_n + \psi_{nt})]$. For example, when $\mu_n + \psi_{nt}$ includes only condition and quality effects, $\hat{\tau}$ will be upward biased if agent-owned properties are more likely to be of higher quality or in better condition.

To mitigate this bias, we augment Equation 1 with indicator variables for the set of candidate tokens in the MLS remarks under the assumption that the tokens can be used to approximate $\mu_n + \psi_{nt}$. Following Nowak and Smith (2017), the tokens are interpreted as proxies for the relevant property-observable attributes that are not included in x_{nt} in Equation 1. This approximation can be described as

$$w_{nt}\theta = \sum_{k=1}^{K} \mathbf{1}(token_k \in remarks_{nt})\theta_k \tag{2}$$

$$p_{nt} = x_{nt}\beta + d_{nt}\tau + w_{nt}\theta + e_{nt} \tag{3}$$

$$e_{nt} = \underbrace{\mu_n + \psi_{nt} - w_{nt}\theta}_{r_{nt}} + v_{nt} \tag{4}$$

In Equation 2, w_{nt} is a vector of indicator variables for the presence of the K candidate tokens and θ is a vector of the implicit prices for each token. For each token, $\mathbf{1}(token_k) = 1$ if token k is in the remarks for property n sold at time t, $remarks_{nt}$, and $\mathbf{1}(token_k) = 0$, otherwise. Equation 3 states that price can also be written as a function of x_{nt}, d_{nt}, and $remarks_{nt}$ plus an error term. Equation 4 states that the error in Equation 3 is the sum of the original error term in Equation 1, v_{nt}, plus an approximation error, r_{nt}.

The approximation r_{nt} reflects the inability of the tokens to perfectly capture $\mu_n + \psi_{nt}$. However, we assume that with enough tokens, we have chosen w_{nt} such that r_{nt} is uncorrelated with x_{nt} or d_{nt}. We place an ℓ_1 penalty on the implicit prices for the tokens and identify a set of Q_2 tokens by solving the following system of equations

$$(\hat{\beta}_p', \hat{\tau}_p, \hat{\theta}_p')' = \arg\min_{\beta,\tau,\theta} \sum (p_{nt} - x_{nt}\beta - d_{nt}\tau - w_{nt}\theta)^2 + \lambda_p \sum_k |\theta_k \phi_{p,k}| \tag{5}$$

$$(\hat{\beta}_d', \hat{\theta}_d')' = \arg\min_{\beta,\theta} \sum (d_{nt} - x_{nt}\beta - w_{nt}\theta)^2 + \lambda_d \sum_k |\theta_k \phi_{d,k}| \tag{6}$$

Define the index of the $\hat{Q}_p$ non-zero coefficients in $\hat{\theta}_p$ as $\hat{S}_p \in \{1, ..., K\}$ and similarly for $\hat{Q}_d$ and $\hat{S}_d$. The objective function in Equation 5 is a penalized hedonic pricing model and the objective function in Equation 6 is a penalized linear probability model.[11] More importantly,

[11] An ℓ_1 penalized logit likelihood was also considered and yielded a $\hat{S}_d$ similar to that in Equation 6. Moreover, the ℓ_1 penalized logit model is well-defined even when the data is linearly separable (Hastie et al., 2015).

the objective functions in Equations 5 and 6 are convex and have solutions that can be found using numerical methods even when K is large.

The $0 \leq \lambda_p$ and $0 \leq \lambda_d$ are tuning parameters that control the size of the penalty. When $\lambda_p = 0$, there is no penalty on θ_k and the solution can be found by least-squares. As λ_p increases, the penalty on θ_k increases and $\hat{\theta}_p$ is shrunk towards 0. The $0 < \phi_{p,k}$ are token-specific penalties that control for heteroskedasticity in e_{nt} (Belloni et al., 2012). Similarly, λ_d and $\phi_{d,q}$ control the penalty in Equation 6. An ℓ_1 penalty is placed on θ_k in both equations.[12] The shape of this penalty results in a sparse solution where many coefficients in $\hat{\theta}_p$ and $\hat{\theta}_d$ will be exactly equal to 0.

Given the objective function in Equation 5, $\hat{S}_p$ is a set of $\hat{Q}_p$ tokens that best predict house prices. Similar to other single-selection methods, there may be variables in $\{1, ..., K\} \setminus \hat{S}_p$ that are correlated with d_{nt} but are poor predictors of house price. Thus, solving Equation 5 alone may not control for omitted variable bias as the variables in S_p may not adequately control for the omitted variable bias associated with d_{nt}. Belloni et al. (2014) demonstrate that the union $\hat{S}_2 = \hat{S}_p \cup \hat{S}_d$ yields a set of $\hat{Q}_2$ tokens that can be used to control for the omitted variable bias associated with d_{nt}. Thus, constructing $\hat{S}_2$ requires two variable selection procedures, which is commonly referred to as a *double-selection*.

The post-LASSO estimator uses the tokens in $\hat{S}_p$ as explanatory variables in an additional regression (Belloni et al., 2013). Similarly, the post double-selection estimator uses only the tokens in $\hat{S}_2$ as explanatory variables in an additional hedonic regression that includes the variable of interest. The post double-selection estimator solves the following

$$(\hat{\beta}_2', \hat{\tau}_2, \hat{\theta}_2')' = \arg\min_{\beta,\tau,\theta} \sum (p_{nt} - x_{nt}\beta - d_{nt}\tau - w_{2,nt}\theta)^2 \tag{7}$$

$$w_{2,nt}\theta = \sum_{q \in \hat{S}_2} \mathbf{1}(token_q \in remarks_{nt})\theta_q \tag{8}$$

[12] The ℓ_1 length of a $K \times 1$ vector x is $\|x\|_1 = \sum_k |x_k|$

In Equation 7, $w_{2,nt}$ is a $\hat{Q}_2 \times 1$ vector of indicator variables for the tokens indicated by $\hat{S}_2$, and $\hat{\theta}_2$ is a $\hat{Q}_2 \times 1$ vector of implicit prices for these tokens. $\hat{S}_2$ is the set of tokens that can be used to (i) predict house prices and/or (ii) identify agent-owned transactions in the data. The subset of tokens that are in $\hat{S}_2$ but not in $\hat{S}_p$, $\hat{S}_2 \setminus \hat{S}_p$, represent the tokens that can mitigate the omitted variable bias associated with agent-owned transactions but, because they do not have large predictive power for house prices, are not included when using a single-selection method.

Theorem 2 in Belloni et al. (2014) provides conditions in which $\hat{\tau}_2$ has an asymptotically normal distribution. These conditions require that with high probability $\hat{\theta}_p$ and $\hat{\theta}_d$ are sparse and provide good approximations to the true conditional expected values of p_{nt} and d_{nt}. As emphasized in Belloni et al. (2014), valid asymptotic inference on $\hat{\tau}$ can take place in the presence of imperfect variable selection. That is, valid asymptotic inference is still possible when $\hat{S}_2$ both omits some relevant tokens and includes some irrelevant tokens.

3 Data

We examine agent-owned transactions using MLS data from Atlanta, Georgia and Phoenix, Arizona. The Georgia Multiple Listing Service (GAMLS) provided the data for Atlanta and the Arizona Multiple Listing Service (ARMLS) provided the data for Phoenix. The GAMLS data covers the five counties (Clayton, Cobb, DeKalb, Fulton and Gwinnett) that form the core of metro-Atlanta. The GAMLS data set includes single-family detached houses that were sold using the services of a real estate agent between January 2000 and September 2016.[13] The Phoenix data includes all transactions in Maricopa County and Pinal County. The two counties cover the city of Phoenix and several surrounding cities including Glendale, Mesa, Scottsdale, and Tempe. The ARMLS data set includes single-family detached houses that were sold using the services of a real estate agent between January 2000 and December

[13] The agent-owned variable was not populated in the GAMLS data until 2007, so the empirical analysis for Atlanta includes every transaction between January 2007 and September 2016.

2013.

Both MLS data sets contain extremely detailed information including the house's address, physical characteristics (e.g. square feet living area, bedrooms, and bathrooms), listing information (e.g. agent-owned, vacant, and rental), transaction details (e.g. time-on-market and sales price), and a text description (i.e. public remark) that the real estate agent uses to market the house. The GAMLS data does not consistently report the square feet of living area or lot size, so we match the properties to county tax assessor records obtained from CoreLogic. We use the agents' description of the property to create the remarks variable, w_{nt}, in Equation 2.

Prior to running the empirical analysis, we impose a number of restrictions on both MLS data sets. We geocode the data using the property address listed in the MLS to obtain location controls (census tract and zip code) for the empirical analysis. Records with property addresses that did not geocode properly are dropped. Using the geocoded address we create a unique identifier that allows us to link listing and sales activity on a given property over time. We remove records for which data on variables of interest are missing or contain invalid values. We also remove houses that sold more than twice within a three year period (i.e. flipped houses) or were part of a distressed sales transaction (i.e. short sale, foreclosure, or REO). To eliminate outliers and minimize data errors, we filter the data on several physical characteristics. A complete list of the filters is reported in the appendix in Section A.1. The filters are comparable to those employed in Levitt and Syverson (2008). The results we report are not sensitive to the filters employed. Summary statistics for the filtered Atlanta and Phoenix data sets are displayed in Table 2.

4 Results

4.1 Variable Selection

Since the results are similar regardless of the token set employed, we only report the results for the unigram token set unless otherwise noted. Not every token is included in the empirical analysis; only tokens selected by the double-selection procedure are included as control variables. For practical purposes, we start with a candidate set of the 2,000 most frequent tokens.

The token selection procedures in Equation 5 and Equation 6 include additively separable time (quarter by year) and location (zip code or census tract) fixed effects. The bulk of the analysis uses zip code fixed effects for four reasons. First, the use of zip code fixed effects facilitates the comparison with the extant literature on agent-owned transactions. Second, our main conclusions are unaffected when including more granular fixed effects for either census tract or census block group. Third, census tract fixed effects have been shown to overfit in-sample (Nowak and Smith, 2017). Fourth, we find that zip code fixed effects are computationally tractable.[14]

Several property characteristics are also included in the token selection procedure in Equation 5 and Equation 6. Age enters into the hedonic functions linearly and indicator variables are included for lot size, living area, bedrooms, and bathrooms. We find these specifications allow for possibly important nonlinear relationships in the true hedonic price function while also providing easily interpreted coefficients. Least-squares coefficient estimates for the indicator variables are presented in the appendix in Tables A1 and A2.

One of the primary contributions of this study is the inclusion of w_2 in Equation 7 to control for the omitted variable bias that may be present when estimating τ using Equation 1.

[14] We used the `hdm` package in `R` to estimate the heteroskedastic LASSO as in Belloni et al. (2014). Computation time for the heteroskedastic LASSO using a Macbook Pro with 8GB of memory and a 2.7 Ghz Intel Core i5 was approximately 30 minutes using 2,000 candidate tokens and zip code fixed effects. Computation time on the same machine with additively separable census tract fixed effects was more than 3 hours. Computation using multiplicatively separable census tract fixed effects was infeasible on this same machine.

The vector w_2 includes indicator variables for the tokens in $\widehat{S}_2$ that represent the observable attributes that were omitted from previous studies.[15] The tokens are selected by minimizing Equation 5 and Equation 6. Table 3 examines the variables selected for Atlanta, GA in Panel A and Phoenix, AZ in Panel B. Using the 2,000 most frequent unigram tokens we create a 2,000 by 1 vector of indicator variables. The top part of each panel presents the correlation between the vectors and the bottom part (i.e. last row) of each panel lists the total number of non-zero variables that were selected.

The process selects 827 of the 2,000 candidate tokens that explain price in $\widehat{S}_p^{tract}$ and 386 tokens that explain the agent-owned indicator in $\widehat{S}_d^{tract}$ when using additively separable time and census tract fixed effects for the Atlanta data set. In total, there are 1,064 unique tokens from both of these sets in $\widehat{S}_2^{tract}$. When using additively separable time and zip code fixed effects the process selects 1,057 unique tokens in $\widehat{S}_2^{zip}$. The results show that the two token sets are highly correlated (83.4%) especially in their selection of the token set in $\widehat{S}_d$ (99.0%). Similar results are reported for Phoenix, AZ in Panel B. In unreported results we create the token sets for every subperiod in Atlanta and Phoenix. The agent-owned subperiod estimates remain the same whether we use the token set for the entire study period or the subperiods. Thus, we only report estimates using the token set for the entire study period.

By including the indicator variables for the tokens in the least-squares estimating equation, we are able to estimate implicit prices for each token in $\widehat{S}_2$. However, similar to others in the machine learning literature, such as Mullainathan and Spiess (2017), we refrain from a strict interpretation of these coefficient estimates as the true price associated with a given token. If anything, we favor an interpretation similar to the inverse regression approach in Taddy (2013) where the likelihood of the appearance of any given token in the remarks is determined by the true condition and quality of the property. More importantly, removing the phrase *fixer upper* from a description while not making any repairs to the property is unlikely to increase its sales price.

[15] The tokens represent data-omitted variables in studies that used county tax assessor data and model-omitted variables in studies that used MLS data.

For informational purposes, Figure 1 plots the ten largest positive and negative tokens for both Atlanta and Phoenix. The results offer some interesting insights into the variable selection process at the zip code and census tract level. For example, five of the ten positive unigram tokens selected using zip code fixed effects are neighborhoods in Atlanta (Collier, Ashford, Ormewood, Grant and Walton). In contrast, only one of the ten positive tokens selected using census tract zip code effects is a neighborhood (EAV which is short for East Atlanta Village). This suggests that zip code fixed effects may not properly control for unobserved variation in the characteristics of the house due to its location. The remaining tokens are relatively intuitive. A house with a *dock* sells for a premium. Whereas, a house that needs a *fix* or is a *fixer upper* sells for less. Although it is the listing agent's job to present the house in the best light possible, Figure 1 shows that listing agents include both positive and negative information in the public remarks section of the MLS. This is important, because it allows us to control for the condition and quality of the house - thereby isolating the pricing differential on agent-owned and non-agent-owned houses.

As mentioned above, $\hat{S}_2$ includes strong predictors of both p and d. Figure 2 presents $\hat{\theta}_p$ and $\hat{\theta}_d$ for bigrams in the Phoenix data.[16] For clarity, only twenty of the strongest predictors in either $\hat{S}_p$ or $\hat{S}_d$ are presented. Properties with *granite-slab* or *travertine-floors* sell for a higher price and are more likely to be agent-owned. In contrast, properties sold via an *estate-sale* or that *need-tlc* (tender loving care) sell for a lower price and are less likely to agent-owned. This finding aligns closely with Campbell et al. (2011) who find that death-related discounts (estate-sale) reflect poor maintenance (need-tlc). Figure 2 also indicates that tokens in $\hat{S}_p$ are not guaranteed to be in $\hat{S}_d$. For example, remarks that have a *lake-view*, are located on a golf course *fairway*, or have a view of the *city-lights* sell for a significant premium but are not likely to be agent-owned.[17]

By construction, tokens in $\hat{S}_p$ explain large variations in transaction price, and many

[16] Additional insight into the bigram and flex-gram tokens that are selected is provided in Section B.5 of the internet appendix.

[17] *th-fairway* is an artifact of the cleaning procedure where all numbers are removed from the remarks.

of these tokens appear to be associated with time-invariant features of the property. The predominance of *new* in the tokens in $\hat{S}_d$ but not $\hat{S}_p$ suggests real estate agents are more likely to upgrade their property before sale. However, because these tokens are not in $\hat{S}_p$, these upgrades are modest in nature and not significant sources of variation in price, relative to the property attributes mentioned in the previous paragraph. Moreover, because these upgrades are time-varying and do not require a permit, they most likely are not included in county tax assessor data sets.

Finally, it is important to note that the tokens in Figure 2 refer to objective, verifiable features of the property. The only possible exception being a *spectacular-view*. This is important because previous studies that used a pre-specified dictionary included words related to the subjective measure of property quality in the dictionary. For example, the dictionary in Levitt and Syverson (2008) includes *amazing, beautiful*, and *breathtaking* while the dictionary in Goodwin et al. (2014) is comprised of 17 positive words such as *beautiful, cozy, gorgeous*, and *lovely*. Furthermore, because these features are verifiable and posted alongside photos of the house, it is difficult for real estate agents to favorably misrepresent the house in the remarks. Lastly, many of the words in Figure 2 would not appear in dictionaries of positive or negative words that are not specific to real estate. This further emphasizes the need for researchers to use a dictionary specific to the research question at hand.

4.2 Agent-Owned Estimates

When comparing agent-owned and non-agent-owned houses it's important to note that the tenure status of the two seller types may systematically differ. This is especially true if the agent-owned sample includes real estate agents who own several rental properties or flip houses.[18] In Table 4 we address this concern by including two indicator variables that control for tenure status. The tenure status variables identify whether the house was listed as vacant or a rental. Rentals generally sell for a discount because they are of lower quality and have

[18] We filter out houses that sold more than two times within a three year period. However, we still suspect that rehabbers and landlords with numerous rental properties are more likely to have a real estate license.

more wear and tear relative to owner-occupied houses (Wang et al., 1991). Vacant properties also sell for a discount that is generally attributed to (i) empty houses not showing as well or (ii) motivated sellers who have less bargaining power (Turnbull and Zahirovic-Herbert, 2011). Rutherford et al. (2005) include similar tenure status controls, but Levitt and Syverson (2008) do not. In subsequent analysis we remove houses that are listed as vacant or rental to allow for a cleaner, more direct comparison of agent-owned and non-agent-owned sales transactions.

Baseline results are presented for Atlanta and Phoenix in Panels A and B of Table 4, respectively. The first three columns include time by zip code fixed effects alongside the aforementioned standard property attributes. The fixed effects in columns 1 to 3 are comparable to the time by city fixed effects employed in Levitt and Syverson (2008). In addition to including the tenure status controls, we also interact them with the agent-owned indicator variable to isolate the agent-owned premium for occupied housing. Although Rutherford et al. (2005) include the tenure status controls, they do not interact them with the agent-owned variable. Consistent with the extant literature, our initial estimate for the agent-owned premium is 3.3% in Atlanta and 4.0% in Phoenix. The estimates in column 1 control for differences in the standard house characteristics, but do not include the textual information available in the remarks section of the MLS. Thus, they likely suffer from an omitted variable bias.

In the absence of the textual information, the agent-owned estimate in Column 1 is both economically and statistically significant for Atlanta. After we include the textual information from the remarks in columns 2 ($\widehat{S}_2^{zip}$) and 3 (L&S), the agent-owned estimate is no longer statistically significant.[19] Also note that the $\widehat{S}_2^{zip}$ token set reduces the vacant (rental) estimate from -12.1% (-3.8%) in column 1 to -7.2% (-1.7%) in column 2. In contrast, the L&S token set has a smaller impact on the vacant estimate and actually increases the rental discount from -3.8% to -4.9%. This is probably due to the fact that the L&S token

[19] The L&S tokens represent the token set described in the appendix of Levitt and Syverson (2008).

set includes mostly "positive" tokens.[20] Similar results are reported in Panel B for Phoenix. The agent-owned estimate drops from 4.0% to 1.7% when the textual information from the remarks is included. In contrast to Atlanta, the agent-owned estimate remains statistically significant for occupied housing.

Columns 4 through 6 of Table 4 include time by census tract fixed effects alongside the standard property attributes and tenure status controls. In the absence of information from the remarks, we naively estimate an agent-owned premium in column 4 of 3.3% for Atlanta and 3.5% for Phoenix. When the tokens in $\widehat{S}_2^{tract}$ are included in column 5, the agent-owned premium declines to 1.5% in Phoenix and is no longer statistically significant in Atlanta. The final column of Table 4 includes the L&S token set alongside the controls in column 4. The agent-owned estimate for Atlanta is 2.1% and is statistically significant at the 10% level. Once again, the agent-owned estimate for Phoenix using the L&S token set (2.7%) is noticeably larger than the estimate using the $\widehat{S}_2^{tract}$ token set (1.5%).

4.2.1 Alternative Token Sets

In Table 5 we filter out (i) transactions in which the house was listed as vacant, (ii) transactions in which the house was listed as a rental, and (iii) agent-owned transactions in which the listing agent had more than three agent-owned transactions during the entire study period. The resulting subsample is homogeneous in terms of tenure status, thereby allowing us to isolate and compare occupied agent-owned and non-agent-owned housing transactions.

Every column in Table 5 includes the standard property controls and time by zip code fixed effects. Column 1 displays agent-owned estimates for the occupied housing subsample in the absence of tokens. Consistent with previous research, we estimate an agent-owned premium of 2.9% in Atlanta and 3.5% in Phoenix. Column 2 includes the 1,057 (1,167) unigram tokens in $\widehat{S}_2^{zip}$ for Atlanta (Phoenix). Given the large number of tokens we include in column 2, it is reasonable to ask if there is any information in the tokens we did not select.

[20] We examine the extent to which the positive and negative tokens in the MLS remarks address the omitted variable bias associated with agent-owned sales transactions in Section C.1 of the internet appendix.

To answer this, we include 943 (833) of the 2,000 candidate tokens that are in the complement of $\widehat{S}_2^{zip}$ as regressors alongside the standard attributes in column 3 of Panel A (Panel B). In doing so, we ask the question "do the tokens not selected by the variable selection procedure contain important information?" The short answer is yes. The agent-owned estimate drops from 2.9% to 2.4% for Atlanta and 3.5% to 2.9% for Phoenix. Although not reported in Table 5 the explanatory power of the complement token set in column 3 is weaker than the $\widehat{S}_2^{zip}$ token set in column 2. For example, in Phoenix the complement token set increases R^2 from 91.2% to 91.7%. Whereas, the $\widehat{S}_2^{zip}$ token set increases the R^2 to 93.7%.

A natural criticism is that by including many regressors we are overfitting the data in-sample and reporting a misleading agent-owned estimate. We assuage this critique using a permutation of the remarks. Specifically, we permute the remarks by randomly sampling the remarks without replacement and treat these remarks as the true remarks. We then create token indicators based on the $\widehat{S}_2^{zip}$ token set that was created using the non-permuted remarks. Results for this experiment are reported in column 4. The agent-owned premium reported in column 4 is nearly identical to the estimate reported in column 1. Thus, it does not appear as though the estimates in column 2 are the result of overfitting. Instead, they are the product of the approach's ability to accurately identify the set of tokens that indicate the true condition and quality of the underlying property.

As noted earlier, the bulk of the analysis reported in this study uses the unigram token set. For comparison purposes we report agent-owned estimates using bigrams in column 5 and flex-grams in column 6. The bigram token set differs from the unigram token set in that it uses two-word phrases instead of single words. Similarly, the flex-gram token set includes commonly used single and multi-word phrases. The flex-gram creation process is described in Section B.2 of the internet appendix. Regardless of the token set employed the agent-owned estimate is no longer statistically significant in Panel A and the magnitude of the estimate decreases considerably in Panel B.[21]

[21] We also ran the analysis using only the tokens in $\widehat{S}_p$. The results are similar, but magnitude of the agent-owned coefficient differs slightly.

Market specific subperiod estimates are also provided in Table 5. The subperiod delineations are selected using a home price index specific to each market. Additional information on the subperiod selection process is provided in Section A.3 of the appendix. The subperiod estimates follow the same pattern as those reported for the entire study period, although they increase in magnitude during the boom, bust and recovery. The results suggest that at least a portion of the agent-owned premium reported in previous studies is attributable to agents purchasing properties that differ in terms of quality, condition, and/or features relative to the average property in the market. This is in contrast to the incentive problems discussed in Rutherford et al. (2005) and Levitt and Syverson (2008).

We recognize that the study periods in Rutherford et al. (2005) and Levitt and Syverson (2008) predate the rise of the internet when potential home buyers had to go to a brokerage office to view the houses available for sale in a preprinted MLS book. During this time period, real estate agents were the "gatekeepers" of the MLS listings, so they almost certainly had an information advantage. The proliferation of real estate websites such as Redfin, Trulia and Zillow has partially leveled the playing field and made more information available online. However, these websites have not changed an agent's incentive to sell their house for a higher price or the fact that most home buyers are still at an informational disadvantage relative to real estate agents. The subperiod estimates in this study lend support this conjecture as the magnitude of the naive agent-owned estimates monotonically increase in both markets from the beginning to the end of the study periods.

4.2.2 Restricted Samples

In this section we further restrict the occupied housing subsample using the joint constraints employed in Rutherford et al. (2005). The first constraint, whose estimates are displayed in column 2 (without tokens) and column 6 (with tokens) of Table 6, restricts the sample to only properties listed in the same census block group as an agent-owned transaction. The first constraint helps limit the variability in location and physical characteristics. The second constraint, whose estimates are displayed in column 3 (without tokens) and column 7 (with tokens), restricts the sample to only include properties listed by an agent that has at least one agent-owned sales transaction during the study period. The second constraint creates a sample in which agency issues should be more easily identified if they exist. Column 4 (without tokens) and column 8 (with tokens) apply both constraints simultaneously to further assess the sensitivity of the agent-owned estimates.

Panel A displays the results for Atlanta for the entire study period and two market subperiods. After imposing the restrictions on the sample, the agent-owned estimates are no longer significant. The results validate our earlier findings that agent-owned houses did not sell for a premium relative to non-agent-owned houses in Atlanta from 2007-2016. The results highlight the fact that our approach effectively incorporated the textual information from the public remarks section of the MLS to control for the variability in location, physical characteristics, and potential agency issues. The results also highlight the need to include the information provided in the textual description of an asset in pricing models when the asset trades in a heterogeneous market.

Panel B displays the agent-owned estimates using the restricted samples for Phoenix. Although the constraints reduce the magnitude of the agent-owned estimates, they remain significant in every subperiod. However, once the constraints and the textual information are included in columns 6 to 8, the agent-owned estimates are no longer significant during the pre-boom and bust subperiods. Although the agent-owned estimates remain significant during the boom and recovery subperiods, the magnitude of the premium drops considerably. The results highlight the value of the informational content in the textual description of the property and its ability to address a portion of the omitted variable bias that is present in previous studies.

4.2.3 Repeat Sales

A repeat sales methodology is often used to address omitted variable bias concerns in the literature. The approach has been used to examine, among other things, information asymmetry in real estate markets (Kurlat and Stroebel 2015; Stroebel 2016), school quality's effect on house prices (Figlio and Lucas 2004; Ries and Somerville 2010), the performance of real estate auctions relative to negotiated sales (Mayer, 1998), and investments in alternative assets such as art (Goetzmann 1993; Korteweg et al. 2015). In this study we show that the omitted variable bias is not resolved when a repeat sales methodology is employed. Differencing Equation 1 gives us:

$$\Delta p_{nt} = p_{nt} - p_{ns} = \Delta x_{nt}^{m} \beta + \Delta d_{nt} \tau + \Delta \psi_{nt} + \Delta \upsilon_{nt} \tag{9}$$

Similar to the hedonic model, an unbiased estimate of the agent-owned premium requires the assumption that there is no correlation between the change in a property's condition before and after an agent-owned transaction, $E[\Delta d_{nt} \Delta \psi_{nt}] = 0$.[22] Unlike the hedonic model, an unbiased agent-owned premium in the repeat-sales estimator does not require any assumptions about the correlation between agent-owned transactions and quality because the approach assumes quality remains constant over time. In any event, if agent-owned properties have superior maintenance (i.e. they are in excellent condition), then the agent-owned estimates will still be biased.

The repeat sales specifications in Table 7 are comparable to the baseline specifications in Table 4, except for the removal and replacement of the standard house attributes with house fixed effects. The inclusion of house fixed effects requires a repeat sales sample in which houses that sold once during the study period are dropped. The results in Table 7 are similar to the previously reported results. The naive agent-owned estimate is significant and of a greater magnitude than the agent-owned estimates that incorporate the textual information

[22] In addition, we also require $E[\Delta x_{nt}^{m} \Delta \psi_{nt}] = 0$ but this is not the focus of the paper.

from the public remarks. The results suggest that using a repeat sales estimation approach does not adequately address omitted variable bias concerns.

4.2.4 Time-on-market

Up to this point we have focused solely on sales price. In this section we estimate time-on-market (TOM) for agent-owned houses. Rutherford et al. (2005) find that agent-owned houses sell for a premium, but do not stay on the market for a longer period of time. Levitt and Syverson (2008), in contrast, argue that real estate agents have an incentive to convince their clients to sell their houses too cheaply *and* too quickly. They find that agent-owned houses stay on the market 9.5 days longer.

Table 8 displays TOM estimates for Atlanta in Panel A and Phoenix in Panel B. The TOM estimates are for occupied housing only and represent the additional number of days that the house was listed on the market.[23] The estimates for Atlanta are insignificant across the entire study period regardless of the functional form and tokens employed. The TOM estimates are insignificant in Phoenix except for the bust subperiod. Similar to Rutherford et al. (2005), we find that agent-owned houses are generally not on the market longer than non-agent-owned houses.[24]

4.3 Robustness Check

As a robustness check and to demonstrate the generalizability of our approach, we also examine vacant house price discounts. Rutherford et al. (2005) include an indicator variable for vacant houses that identifies when "the owner has already moved and thus needs to sell." Although it is not the focus of their study, Rutherford et al. (2005) estimate that vacant houses sell for a 6% to 7% discount. Studies whose primary focus is the estimation of vacancy discounts, such as Turnbull and Zahirovic-Herbert (2011), report similar estimates.[25]

[23] Similar results are found when log(TOM) is the dependent variable.

[24] We also examine the co-determination of sales price and time-on-market in Section C.4 of the internet appendix.

[25] Levitt and Syverson (2008) do not identify vacant houses in their list of standard attributes or keywords.

Although we do not doubt the sign and significance of the results reported in previous studies, we suspect that the magnitude of the results may be overestimated due to an omitted variable bias. Turnbull and Zahirovic-Herbert (2011) raise a similar concern noting that "vacancy might also signal the presence of an unobservable factor that reduces buyer willingness to pay for the house. The notion here is that vacant houses have undesirable characteristics that are observed by sellers and buyers but are not reported in the data (condition, architecture, etc.)." The undesirable characteristics not only reduce the buyer's willingness to pay, but also contribute to the property being vacant in the first place. Thus, if the undesirable characteristics are not properly controlled for the magnitude of the vacancy discount will be biased. We examine whether we can control for the "undesirable characteristics" that Turnbull and Zahirovic-Herbert (2011) mention using the textual information in the remarks section of the MLS. To do so, we rerun the double-selection LASSO procedure with an indicator variable for vacant houses in lieu of the indicator variable for agent-owned transactions in Equation 6.

Table 9 displays the results for Atlanta in columns 1 to 4 and Phoenix in columns 5 to 8. Agent-owned and rental property transactions are not included to allow for a more direct comparison of vacant versus occupied price differentials. Every column includes the standard property attributes and time by zip code (census tract) fixed effects are included in columns 1, 2, 5, and 6 (3, 4, 7 and 8). The vacant estimates are provided for the entire study period and several subperiods.[26] Odd columns in Table 9 display naive estimates and even columns display estimates that incorporate the textual information from the remarks.

Similar to the agent-owned estimates we find that the magnitude of the vacant estimates decrease when the textual information from the public remarks is incorporated. In Atlanta, the vacant discount estimate drops 41% from 8.7% to 5.1% using zip code controls. The results in Phoenix mimic Atlanta. The vacant discount estimate drops 48% from 4.8% to

[26] When estimating agent-owned premiums, Atlanta's entire study period is 2007 to 2016 because the agent-owned variable was not available in the data set prior to 2007. However, when estimating the vacant discounts, Atlanta's entire study period is 2000 to 2016 because the vacant variable is populated and available for the entire study period.

2.5%. The results in Table 9 highlight the effect the real estate market cycle has on vacancy discounts. The magnitude of the vacancy discounts is much lower during the pre-boom and boom subperiods relative to the bust and recovery subperiods. Regardless of the subperiod, the inclusion of the textual information from the agent remarks reduces the vacant discount estimate, thereby demonstrating our approach's ability to address a portion of the data-omitted variable bias inherent in the naive estimates.

5 Conclusion

Although researchers often have access to textual information about an asset, the high-dimensional nature of text precludes the use of conventional econometric techniques. We provide a data-driven framework for incorporating text into asset pricing models using a double-selection LASSO procedure that removes the researcher's qualitative judgment as part of the analytical procedure. More specifically, we create a real estate specific dictionary that (i) can be used to predict house prices and/or agent-owned transactions and (ii) is sufficient for valid asymptotic inference on the true parameter of interest (e.g. agent-owned transactions). The framework we describe is incredibly flexible and can be used to incorporate textual information into asset pricing models beyond real estate.[27]

Empirically, we show that the textual information in the remarks section of a MLS listing can be used to mitigate the data-omitted variable bias associated with agent-owned transactions. Using MLS data from Atlanta, Georgia and Phoenix, Arizona we replicate the naive agent-owned premium estimates in Rutherford et al. (2005) and Levitt and Syverson (2008) of 3 to 4 percent. However, after we incorporate the textual information from the remarks section of the MLS, the agent-owned premium drops by over 50 percent in Phoenix and is no longer statistically significant in Atlanta. Using a restricted subsample we also find that the agent-owned estimate is statistically insignificant in Phoenix during the pre-boom

[27] For example, the framework we develop is applicable to any asset class in which an Akerlof (1970) lemons type market can arise such as used cars or unbranded hotels.

(2000-2003) and bust (2007-2009) subperiods.

Rutherford et al. (2005) and Levitt and Syverson (2008) interpret the naive agent-owned estimates as agents having an information advantage that they can exploit when selling their own house. We show that the naive estimates suffer from a model-omitted variable bias that can be mitigated by including textual information from the remarks section of the MLS. After we include the textual information, we find that real estate agents do not necessarily sell their own house for more than comparable client-owned houses. Thus, we conclude that real estate agents do not use their information advantage to their clients' detriment (i.e. an agency problem does not exist).

We also show that using a repeat sales estimator approach does not resolve the model-omitted variable bias. This is because the repeat-sales approach assumes that the condition and quality of the house and neighborhood remain constant over time. We show that the inclusion of the textual information in the repeat sales model helps control for the time-varying condition of the house and neighborhood, thereby allowing us to isolate the pricing differential for agent-owned properties. The results highlight the importance of including textual information in asset pricing models, especially when the assets trade in a heterogeneous market.

References

Akerlof, G. A. (1970). The market for" lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, pages 488–500.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A., Chernozhukov, V., et al. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Bodnaruk, A., Loughran, T., and McDonald, B. (2015). Using 10-k text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4):623–646.

Buehlmaier, M. M. and Whited, T. M. (2018). Are financial constraints priced? evidence from textual analysis. *The Review of Financial Studies*, forthcoming.

Campbell, J. Y., Giglio, S., and Pathak, P. (2011). Forced sales and house prices. *The American Economic Review*, 101(5):2108–2131.

Figlio, D. N. and Lucas, M. E. (2004). What's in a grade? school report cards and the housing market. *The American Economic Review*, 94(3):591–604.

Garmaise, M. J. and Moskowitz, T. J. (2003). Confronting information asymmetries: Evidence from real estate markets. *The Review of Financial Studies*, 17(2):405–437.

Gentzkow, M., Kelly, B. T., and Taddy, M. (2017). Text as data. Technical report, National Bureau of Economic Research.

Goetzmann, W. N. (1993). Accounting for taste: Art and the financial markets over three centuries. *The American Economic Review*, 83(5):1370–1376.

Goodwin, K., Waller, B., and Weeks, H. S. (2014). The impact of broker vernacular in residential real estate. *Journal of Housing Research*, 23(2):143–161.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Hoberg, G. and Maksimovic, V. (2014). Redefining financial constraints: A text-based analysis. *The Review of Financial Studies*, 28(5):1312–1352.

Kelly, B. and Ljungqvist, A. (2012). Testing asymmetric-information asset pricing models. *The Review of Financial Studies*, 25(5):1366–1413.

King, G., Lam, P., and Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*.

Korteweg, A., Kräussl, R., and Verwijmeren, P. (2015). Does it pay to invest in art? a selection-corrected returns perspective. *The Review of Financial Studies*, 29(4):1007–1038.

Kurlat, P. and Stroebel, J. (2015). Testing for information asymmetries in real estate markets. *The Review of Financial Studies*, 28(8):2429–2461.

Levitt, S. D. and Syverson, C. (2008). Market distortions when agents are better informed: The value of information in real estate transactions. *The Review of Economics and Statistics*, 90(4):599–611.

Loughran, T. and Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Mayer, C. J. (1998). Assessing the performance of real estate auctions. *Real Estate Economics*, 26(1):41–66.

Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nowak, A. and Smith, P. (2017). Textual analysis in real estate. *Journal of Applied Econometrics*, 32(4):896–918.

Ries, J. and Somerville, T. (2010). School quality and residential property values: evidence from vancouver rezoning. *The Review of Economics and Statistics*, 92(4):928–944.

Rutherford, R. C., Springer, T. M., and Yavas, A. (2005). Conflicts between principals and agents: evidence from residential brokerage. *Journal of Financial Economics*, 76(3):627–665.

Stroebel, J. (2016). Asymmetric information about collateral values. *The Journal of Finance*, 71(3):1071–1112.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance*, 62(3):1139–1168.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Turnbull, G. K. and Zahirovic-Herbert, V. (2011). Why do vacant houses sell for less: holding costs, bargaining power or stigma? *Real Estate Economics*, 39(1):19–43.

Wang, K., Grissom, T. V., Webb, J. R., and Spellman, L. (1991). The impact of rental properties on the value of single-family residences. *Journal of Urban Economics*, 30(2):152–166.

Young, M. (2012). Property Condition. *Southern Nevada Realtor Magazine.*

Tables and Figures

Table 1: Sample of MLS Listings

Sample Listing 1

Tract	Beds	Baths	Sqft	Date	Price	$\hat{u}$	$\hat{u}_2$
04013217900	3	2	1,491	3/28/2013	\$167,000	0.134	0.054

MLS Remarks: **flagstone** in kitchen, dining and den. new kitchen cabinets with island. wood floors in family room, hallway and master suite. **fully-remodeled** baths, **new tile** and plumbing fixtures. all new copper plumbing throughout. a/c and master cool evap. great pool with redwood deck. block fence. el dorado park just down the street!

Sample Listing 2

Tract	Beds	Baths	Sqft	Date	Price	$\hat{u}$	$\hat{u}_2$
04013217900	3	2	1,497	3/14/2013	\$139,900	-0.146	-0.004

MLS Remarks: great **opportunity** for **investment** or **affordable** housing-3 bedroom, 2 bath + bonus room (ideal office/playroom)-this solidly built brick hallcraft **home-needs** cosmetic work but has good roof, a/c, large yard, rv gate & great **potential** washer, dryer, refrigerator & dishwasher are included in 'as is' condition

Note: Table 1 displays two transactions in the data with the original remarks in the MLS listing. Both listings are 3 bedroom, 2 bathroom houses with approximately 1,500 square feet of living area that sold within two weeks of each other and are located in the same census tract. $\hat{u}$ is the baseline residual from a hedonic model using census tract by time fixed effects and property controls. $\hat{u}_2$ is the residual when the textual information from the remarks, $\hat{S}_2$, is included as indicator variables in the hedonic model. The flex-grams (i.e. words and phrases) in $\hat{S}_2$ are indicated with bold text.

Table 2: Descriptive Statistics

Panel A: Atlanta (2007-2016)						
	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Price (000s)	50.00	136.10	225.31	192.74	283.00	2,600.00
Sfla (000s)	0.57	1.64	2.20	2.12	2.67	6.00
Age	2.00	14.00	29.81	24.00	42.00	196.00
Bedrooms	1.00	3.00	3.63	4.00	4.00	6.00
Bathrooms	1.00	2.00	2.49	2.50	3.00	3.50
Agent-owned	0.00	0.00	0.02	0.00	0.00	1.00
Rental	0.00	0.00	0.03	0.00	0.00	1.00
Vacant	0.00	0.00	0.35	0.00	1.00	1.00
Panel B: Phoenix (2000-2013)						
	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Price (000s)	50.00	142.50	247.48	204.00	293.00	3,000.00
Sfla (000s)	0.50	1.44	1.92	1.77	2.24	6.00
Age	2.00	7.00	19.86	16.00	29.00	122.00
Bedrooms	1.00	3.00	3.28	3.00	4.00	6.00
Bathrooms	1.00	2.00	2.16	2.00	2.50	3.50
Agent-owned	0.00	0.00	0.06	0.00	0.00	1.00
Rental	0.00	0.00	0.03	0.00	0.00	1.00
Vacant	0.00	0.00	0.37	0.00	1.00	1.00

Note: Panel A includes transactions in Atlanta, GA from 2007 to 2016. Panel B includes transactions in Phoenix, AZ from 2000 to 2013. The descriptive statistics are based on the authors' calculations.

Table 3: Similarity of Tokens Selected

Panel A: Atlanta (2007-2016)						
	tract, Price	tract, Agent-owned	zip, Price	zip, Agent-owned	tract, Both	zip, Both
tract, Price	1					
tract, Owner Agent	-0.027	1				
zip, Price	0.797	-0.023	1			
zip, Owner Agent	-0.030	0.990	-0.028	1		
tract, Both	0.788	0.459	0.621	0.456	1	
zip, Both	0.618	0.462	0.786	0.462	0.834	1
Non-Zero	827	386	818	386	1,064	1,057

Panel B: Phoenix (2000-2013)						
	tract, Price	tract, Agent-owned	zip, Price	zip, Agent-owned	tract, Both	zip, Both
tract, Price	1					
tract, Owner Agent	0.079	1				
zip, Price	0.782	0.088	1			
zip, Owner Agent	0.074	0.983	0.078	1		
tract, Both	0.789	0.492	0.619	0.481	1	
zip, Both	0.612	0.484	0.789	0.49	0.818	1
Non-Zero	927	502	932	503	1,162	1,167

Note: Table 3 examines the tokens selected using the double-selection variable selection method. We create indicator variables for each of the 2,000 candidate tokens where each indicator variable is equal to 1 if the token is non-zero in the variable selection procedure. Using these 2,000 indicator variables, we create a $2,000 \times 1$ vector of the indicator variables. The top part of each panel presents the correlation between these vectors. The last row of each panel lists the total number of non-zero variables for each vector. [tract, Price] and [zip, Price] ([tract, Agent-owned] and [zip, Agent-owned]) are the vector of indicator variables when Price (Agent-owned indicator) is the dependent variable in the variable selection procedure. [tract, Both] ([zip, Both]) is the element-wise maximum of [tract, Price] and [tract, Agent-owned] ([zip, Price] and [zip, Agent-owned]) and correspond to those variables in $\widehat{S}_2^{tract}$ ($\widehat{S}_2^{zip}$). Panel A includes the token sets for transactions in Atlanta, GA from 2007 to 2016 and Panel B includes the token sets for transactions in Phoenix, AZ from 2000 to 2013.

Table 4: Baseline agent-owned estimates by tenure

Panel A: Atlanta (2007-2016)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.033^{**}	0.012	0.021	0.033^{**}	0.015	0.021^{*}
	(0.015)	(0.012)	(0.013)	(0.013)	(0.011)	(0.012)
Vacant	-0.121^{***}	-0.072^{***}	-0.099^{***}	-0.103^{***}	-0.063^{***}	-0.085^{***}
	(0.017)	(0.008)	(0.013)	(0.014)	(0.007)	(0.011)
Vacant x Agent-owned	0.055^{*}	0.015	0.013	0.063^{**}	0.021	0.023
	(0.033)	(0.020)	(0.030)	(0.025)	(0.014)	(0.021)
Rental	-0.038^{*}	-0.017	-0.049^{***}	-0.013	-0.005	-0.026^{**}
	(0.019)	(0.010)	(0.016)	(0.015)	(0.009)	(0.012)
Rental x Agent-owned	0.020	-0.002	0.014	0.030	0.006	0.024
	(0.031)	(0.026)	(0.029)	(0.024)	(0.020)	(0.023)
N	106,048	106,048	106,048	106,048	106,048	106,048
K	1,110	2,167	1,161	5,642	6,706	5,693
R^2	0.809	0.862	0.827	0.862	0.898	0.876
Panel B: Phoenix (2000-2013)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.040^{***}	0.017^{***}	0.031^{***}	0.035^{***}	0.015^{***}	0.027^{***}
	(0.005)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)
Vacant	-0.048^{***}	-0.025^{***}	-0.040^{***}	-0.041^{***}	-0.023^{***}	-0.035^{***}
	(0.007)	(0.004)	(0.007)	(0.005)	(0.003)	(0.004)
Vacant x Agent-owned	-0.025^{***}	-0.016^{***}	-0.029^{***}	-0.013^{***}	-0.009^{**}	-0.017^{***}
	(0.008)	(0.006)	(0.008)	(0.005)	(0.004)	(0.005)
Rental	-0.069^{***}	-0.030^{***}	-0.057^{***}	-0.063^{***}	-0.029^{***}	-0.052^{***}
	(0.011)	(0.005)	(0.010)	(0.008)	(0.004)	(0.007)
Rental x Agent-owned	-0.026^{**}	-0.004	-0.020^{*}	-0.021^{**}	-0.006	-0.018^{**}
	(0.012)	(0.010)	(0.012)	(0.009)	(0.007)	(0.008)
N	275,049	275,049	275,049	275,049	275,049	275,049
K	1,838	3,005	1,890	11,571	12,733	11,623
R^2	0.901	0.929	0.908	0.929	0.947	0.934
Controls	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓			
Time x Tract FE				✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	L&S		$\hat{S}_2^{tract}$	L&S

*p<0.1; **p<0.05; ***p<0.01

Note: Agent-owned estimates are reported for Atlanta in Panel A and Phoenix in Panel B. Every column incorporates a set of controls that include the age of the house and indicator variables for bedrooms, bathrooms, living area, and lot size. Columns 1 to 3 include multiplicatively separable time and zip code fixed effects. Columns 4 to 6 include multiplicatively separable time and census tract fixed effects. Columns 1 and 4 do not include any tokens. Columns 2 and 5 include unigram tokens in the set $\hat{S}_2$ that differ only in the use of zip or tract fixed effects during the variable selection process. Columns 3 and 6 use the token set described in Levitt and Syverson (2008).

Table 5: Agent-owned estimates using alternative tokens by subperiod

Panel A: Atlanta						
	(1)	(2)	(3)	(4)	(5)	(6)
Entire	0.029**	0.014	0.024*	0.029**	0.017	0.017
Bust	0.021**	0.009	0.017	0.022**	0.011	0.014
Recovery	0.034***	0.016	0.027	0.034***	0.021	0.019
Panel B: Phoenix						
	(1)	(2)	(3)	(4)	(5)	(6)
Entire	0.035***	0.016***	0.029***	0.036***	0.019***	0.019***
Pre-boom	0.024***	0.010**	0.021***	0.024***	0.013***	0.013**
Boom	0.037***	0.016***	0.029***	0.036***	0.021***	0.020***
Bust	0.044***	0.020**	0.035***	0.044***	0.020**	0.021**
Recovery	0.049***	0.027***	0.041***	0.049***	0.031***	0.030***
Controls	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓	✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	$\hat{S}_2^{zip^c}$	$\hat{S}_2^{zip^p}$	$\hat{S}_2^{zip^{bi}}$	$\hat{S}_2^{zip^f}$

*p<0.1; **p<0.05; ***p<0.01

Note: Transactions in which the house was flagged as either vacant or a rental are not included, so the estimates reported are for occupied housing only. Agent-owned transactions by listing agents with more than three agent-owned transactions over the entire study period are also removed. Estimates for Atlanta are displayed in Panel A and Phoenix in Panel B. Every column incorporates a set of controls that include the age of the house and indicator variables for bedrooms, bathrooms, living area, and lot size. Multiplicatively separable time and zip code fixed effects are employed in every column. Column 1 does not include any tokens and column 2 includes the unigram token set $\hat{S}_2^{zip}$. Column 3 uses the set of 2,000 most frequent unigram tokens not in $\hat{S}_2^{zip}$. Column 4 uses the $\hat{S}_2^{zip}$ unigram token set, but permutes the remarks. Column 5 uses the bigram token set and column 6 uses the flex-gram token set. Agent-owned estimates are provided for the entire study period and several market specific subperiods.

Table 6: Agent-owned estimates using restricted subsamples

Panel A: Atlanta

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Entire	0.029**	0.021*	0.018	0.010	0.014	0.009	0.015	0.012
Bust	0.021**	0.015	0.012	0.010	0.009	0.001	−0.002	−0.006
Recovery	0.034***	0.024	0.021	0.016	0.017	0.010	0.019	0.017

Panel B: Phoenix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Entire	0.035***	0.029***	0.033***	0.024***	0.016***	0.011***	0.017***	0.011***
Pre-boom	0.024***	0.018***	0.023***	0.017***	0.010**	0.006	0.009**	0.005
Boom	0.037***	0.031***	0.033***	0.025***	0.016***	0.010***	0.017***	0.012***
Bust	0.044***	0.038***	0.041***	0.030***	0.020**	0.014*	0.023***	0.016
Recovery	0.049***	0.037***	0.045***	0.034***	0.027***	0.019***	0.037***	0.031***
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓	✓	✓	✓	✓	✓
BG Restrict		✓		✓		✓		✓
AO Restrict			✓	✓			✓	✓
Tokens					$\hat{S}_2^{zip}$	$\hat{S}_2^{zip}$	$\hat{S}_2^{zip}$	$\hat{S}_2^{zip}$

*p<0.1; **p<0.05; ***p<0.01

Note: Transactions in which the house was listed as either vacant or a rental are not included, so the estimates reported are for occupied housing only. Agent-owned transactions by listing agents with more than three agent-owned transactions over the entire study period are also removed. Market specific subperiod estimates are reported for Atlanta in Panel A and Phoenix in Panel B. Every column incorporates a set of controls that include the age of the house and indicator variables for bedrooms, bathrooms, living area, lot size, and multiplicatively separable time and zip code fixed effects. Columns 1 to 4 do not include any tokens. Columns 5 to 8 include the unigram token set $\hat{S}_2^{zip}$. Columns 1 and 5 include the entire occupied housing transaction sample (i.e. vacant and rental properties are not included). Columns 2 and 6 restrict the occupied housing sample to only include transactions located in a census block group that had at least one agent-owned transaction during the study period. Columns 3 and 7 restrict the occupied housing sample to only include transactions by listing agents that had at least one agent-owned sales transaction during the study period. Columns 4 and 8 restrict the occupied housing sample to include transactions that were both (i) by a listing agent with at least one agent-owned transacation and (ii) located in a census block group that had at least one agent-owned transaction.

Table 7: Repeat sales agent-owned estimates

Panel A: Atlanta (2007-2016)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.063***	0.017	0.034	0.063**	0.038	0.044
	(0.022)	(0.029)	(0.021)	(0.031)	(0.042)	(0.033)
N	13,404	13,404	13,404	13,404	13,404	13,404
K	7,496	8,551	7,547	9,761	10,822	9,812
R^2	0.969	0.981	0.973	0.984	0.991	0.986
Panel B: Phoenix (2000-2013)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.029***	0.013***	0.018***	0.032***	0.017***	0.022***
	(0.005)	(0.005)	(0.004)	(0.004)	(0.005)	(0.004)
N	58,810	58,810	58,810	58,810	58,810	58,810
K	30,343	31,510	30,395	36,917	38,079	36,969
R^2	0.982	0.986	0.983	0.988	0.991	0.989
House FE	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓			
Time x Tract FE				✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	L&S		$\hat{S}_2^{tract}$	L&S

*p<0.1; **p<0.05; ***p<0.01

Note: The repeat sales estimates only include houses that sold at least twice during the study period. Transactions in which the house was flagged as either vacant or a rental are not included. Agent-owned transactions by listing agents with more than three agent-owned transactions over the entire study period are also removed. Panel A includes transactions in Atlanta, GA from 2007 to 2016 and Panel B includes transactions in Phoenix, AZ from 2000 to 2013. Every column includes house fixed effects. Multiplicatively separable time and zip code fixed effects are employed in columns 1 to 3 and multiplicatively separable time and census tract fixed effects are employed in columns 4 to 6. Columns 1 and 4 do not include any tokens. Columns 2 and 5 include unigram tokens in the set $\hat{S}_2$ that differ only in the use of zip or tract fixed effects during the variable selection process. Columns 3 and 6 use the token set described in Levitt and Syverson (2008).

Table 8: Agent-owned time-on-market estimates

Panel A: Atlanta						
	(1)	(2)	(3)	(4)	(5)	(6)
Entire	1.95	0.10	2.04	2.82	0.89	2.90
Bust	7.04	6.40	6.91	8.94	7.14	8.79
Recovery	−1.12	−2.98	−0.93	−0.77	−2.80	−0.59
Panel B: Phoenix						
	(1)	(2)	(3)	(4)	(5)	(6)
Entire	2.14	1.53	1.97	2.04	1.46	1.84
Pre-boom	0.79	−0.00	0.71	0.43	−0.33	0.32
Boom	−1.32	−2.20	−1.48	−1.37	−2.22	−1.55
Bust	9.89*	11.22*	9.77*	10.60	11.01*	10.47*
Recovery	5.26	4.47	4.67	5.55	4.87	4.97
Controls	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓			
Time x Tract FE				✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	L&S		$\hat{S}_2^{zip}$	L&S

*p<0.1; **p<0.05; ***p<0.01

Note: Time-on-market (TOM) estimates for Atlanta and Phoenix are displayed in Panel A and B, respectively. Transactions in which the house was flagged as either vacant or a rental are not included, so the estimates reported are for occupied housing only. Agent-owned transactions by listing agents with more than three agent-owned transactions over the entire study period are also removed. Every column incorporates a set of controls that include the age of the house and indicator variables for bedrooms, bathrooms, living area, and lot size. Columns 1 to 3 include multiplicatively separable time and zip code fixed effects and columns 4 to 6 include multiplicatively separable time and census tract code fixed effects. Columns 1 and 4 do not use any tokens. Columns 2 and 5 use the unigram token set $\hat{S}_2^{zip}$. Column 3 and 6 use the token set described in Levitt and Syverson (2008). The agent-owned TOM estimates are provided for the entire study period and the subperiods described in Section A.3.

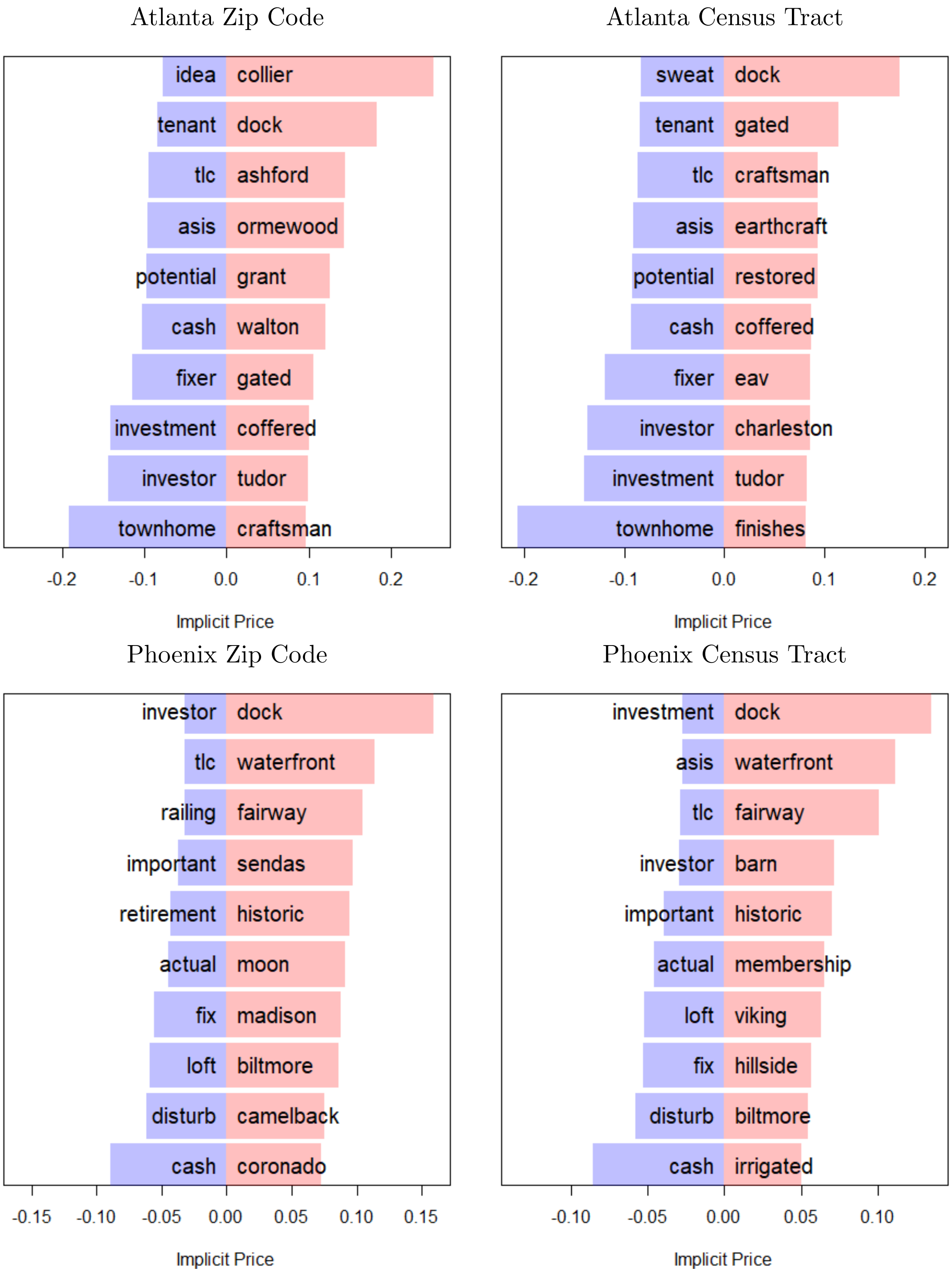
Table 9: Vacant discount estimates

	Atlanta				Phoenix			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Entire	-0.087^{***}	-0.051^{***}	-0.075^{***}	-0.046^{***}	-0.048^{***}	-0.025^{***}	-0.042^{***}	-0.023^{***}
Pre-boom	-0.033^{***}	-0.016^{***}	-0.032^{***}	-0.016^{***}	-0.032^{***}	-0.015^{***}	-0.032^{***}	-0.016^{***}
Boom	-0.055^{***}	-0.029^{***}	-0.049^{***}	-0.028^{***}	-0.032^{***}	-0.015^{***}	-0.029^{***}	-0.015^{***}
Bust	-0.156^{***}	-0.082^{***}	-0.134^{***}	-0.074^{***}	-0.073^{***}	-0.038^{***}	-0.063^{***}	-0.034^{***}
Recovery	-0.107^{***}	-0.069^{***}	-0.090^{***}	-0.059^{***}	-0.074^{***}	-0.039^{***}	-0.059^{***}	-0.033^{***}
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓			✓	✓		
Time x Tract FE			✓	✓			✓	✓
Tokens		$\hat{S}_2^{zip}$		$\hat{S}_2^{tract}$		$\hat{S}_2^{zip}$		$\hat{S}_2^{tract}$

*p<0.1; **p<0.05; ***p<0.01

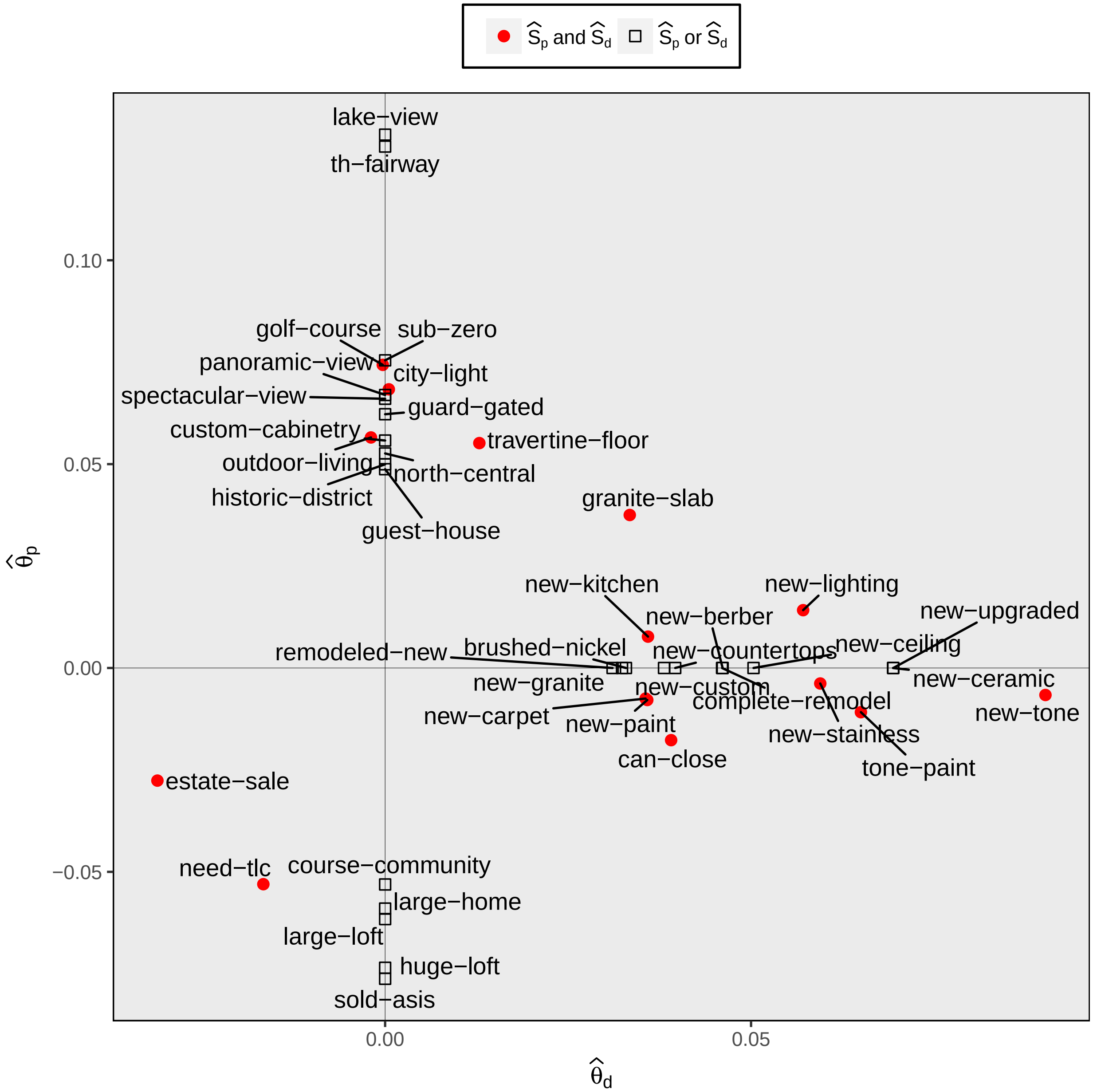
Note: Vacant discount estimates are displayed for Atlanta, GA in columns 1 to 4 and Phoenix, AZ in columns 5 to 8. Transactions in which the house was flagged as agent-owned or a rental are not included. Every column incorporates a set of controls that include the age of the house and indicator variables for bedrooms, bathrooms, living area, and lot size. Columns 1, 2, 5 and 6 include multiplicatively separable time and zip code fixed effects. Columns 3, 4, 7 and 8 include multiplicatively separable time and census tract fixed effects. Odd columns do not include tokens. Even columns include unigram tokens in the set $\hat{S}_2$ that differ only in the use of zip code or census tract fixed effects during the variable selection process. The first row displays vacant coefficient estimates for the entire study period in Atlanta (2000-2016) and Phoenix (2000-2013). The pre-boom (2000-2003) and boom (2004-2006) periods are aligned for Atlanta and Phoenix. However, Atlanta's bust (2007-2011) and recovery (2012-2016) subperiods differ from the bust (2006-2009) and recovery (2010-2013) subperiods in Phoenix. Additional information about the subperiods is provided in the appendix in Section A.3.

Figure 1: Implicit Prices for Unigram Tokens



Note: The top ten positive and negative unigram tokens are displayed for Atlanta and Phoenix. The tokens with the largest magnitudes were selected from the zip code and census tract token sets.

Figure 2: $\hat{\theta}_p$ and $\hat{\theta}_d$ for Bigram Tokens



Note: Figure 2 plots $\hat{\theta}_p$ and $\hat{\theta}_d$ from the ℓ_1 penalized hedonic and linear probability models in Equations 5 and 6. $\hat{\theta}_p$ represents the coefficients for the tokens in $\hat{S}_p$ when transaction price is the dependent variable in Equation 5 and $\hat{\theta}_d$ represents the coefficients for the tokens in $\hat{S}_d$ when the agent-owned indicator variable is the dependent variable in Equation 6. Tokens that are in both $\hat{S}_p$ and $\hat{S}_d$ are displayed as red circles. Tokens in $\hat{S}_p$ ($\hat{S}_d$) but not $\hat{S}_d$ ($\hat{S}_p$) are displayed as unfilled squares. For clarity, only the 20 largest $\hat{\theta}_p$ in absolute value and the 20 largest $\hat{\theta}_d$ in absolute value are displayed.

Appendices

Contents

A Data Overview

A.1 Data filters

We drop houses that sold twice within a three year period, had a public remark with a length less than ten characters, or were involved in a distressed sales transaction. To eliminate outliers, we also drop transactions that do not meet the following criteria:

1. \$50,000 $\leq$ sale price $\leq$ \$3,000,000
2. 500 $\leq$ square feet of living area $\leq$ 6,000
3. 1 $\leq$ bedrooms $\leq$ 6
4. 1 $\leq$ bathrooms $\leq$ 3.5
5. age $\geq$ 2
6. acres $\leq$ 5
7. time-on-market $>$ 0

A.2 Standard housing attributes

Table A1: Atlanta agent-owned controls

	Mean	Basic	Tokens
	(1)	(2)	(3)
Age (Years)	29.806	-0.001	-0.001
Rental	0.031	-0.046	-0.019
Vacant	0.353	-0.125	-0.074
Acres: .5 - 1	0.108	0.024	0.027
Acres: 1 - 2	0.031	0.103	0.091
Acres: 2 - 5	0.010	0.227	0.222
Baths: 1.5	0.021	0.067	0.048
Baths: 2	0.286	0.245	0.189
Baths: 2.5	0.361	0.302	0.237
Baths: 3	0.167	0.327	0.265
Baths: 3.5	0.127	0.450	0.356
Beds: 2	0.034	-0.050	-0.040
Beds: 3	0.424	-0.058	-0.017
Beds: 4	0.428	-0.018	0.026
Beds: 5	0.105	-0.001	0.047
Beds: 6	0.008	-0.057	0.018
Sfla: 500 - 1000	0.015	-0.079	-0.069
Sfla: 1500 - 2000	0.263	0.134	0.112
Sfla: 2000 - 2500	0.242	0.272	0.225
Sfla: 2500 - 3000	0.177	0.409	0.335
Sfla: 3000 - 3500	0.088	0.510	0.418
Sfla: 3500 - 4000	0.035	0.598	0.494
Sfla: 4000 - 4500	0.013	0.655	0.550
Sfla: 4500 - 5000	0.004	0.733	0.607
Sfla: 5000 - 5500	0.002	0.815	0.703
Sfla: 5500 - 6000	0.000	0.898	0.797

Note: Table A1 displays descriptive statistics and implicit prices for the control variables based on transactions in Atlanta, GA from 2007 to 2016. The age of the house is the only continuous variable. The remaining variables are dummies for the number of bedrooms, bathrooms, living area, lot size, vacant, and rental. All implicit prices are relative to a 1 bed, 1 bath owner-occupied property with less than or equal to half an acre of land and 1,000 to 1,500 square feet of living area. Here we use 500 square feet living area (sfla) bins to save space when approximating the sfla coefficients. The empirical analysis uses 100 sfla bins. Column 2 presents the coefficient estimates without tokens and column 3 presents coefficient estimates when the tokens are included in the regression.

Table A2: Phoenix agent-owned controls

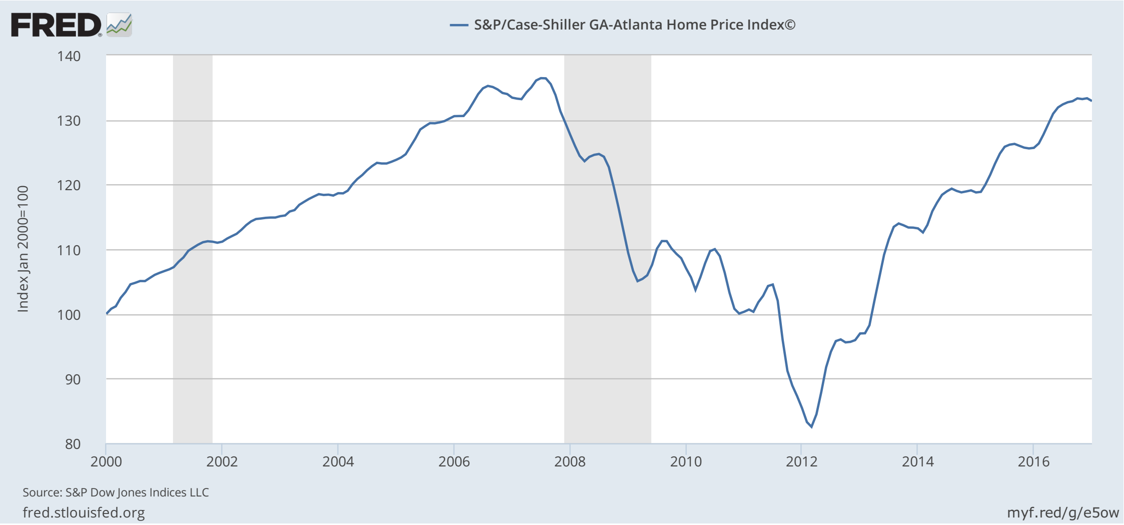
	Mean (1)	Basic (2)	Tokens (3)
Age (Years)	9.934	-0.006	-0.005
Rental	0.027	-0.097	-0.040
Vacant	0.374	-0.060	-0.032
Acres: .5 - 1	0.020	0.243	0.194
Acres: 1 - 2	0.024	0.254	0.196
Acres: 2 - 5	0.005	0.440	0.359
Baths: 1.5	0.143	0.078	0.073
Baths: 2	0.531	0.102	0.090
Baths: 2.5	0.140	0.086	0.098
Baths: 3	0.114	0.113	0.116
Baths: 3.5	0.032	0.184	0.158
Beds: 2	0.131	0.144	0.145
Beds: 3	0.514	0.140	0.166
Beds: 4	0.300	0.115	0.162
Beds: 5	0.052	0.038	0.129
Beds: 6	0.003	-0.036	0.087
Sfla: 500 - 1000	0.023	-0.140	-0.116
Sfla: 1500 - 2000	0.346	0.189	0.150
Sfla: 2000 - 2500	0.198	0.423	0.334
Sfla: 2500 - 3000	0.090	0.657	0.516
Sfla: 3000 - 3500	0.049	0.813	0.647
Sfla: 3500 - 4000	0.020	0.941	0.758
Sfla: 4000 - 4500	0.009	1.036	0.841
Sfla: 4500 - 5000	0.002	1.116	0.922
Sfla: 5000 - 5500	0.001	1.221	1.008
Sfla: 5500 - 6000	0.000	1.280	1.063

Note: Table A2 displays descriptive statistics and implicit prices for the control variables based on transactions in Phoenix, AZ from 2000 to 2013. The age of the house is the only continuous variable. The remaining variables are dummies for the number of bedrooms, bathrooms, living area, lot size, vacant, and rental. All implicit prices are relative to a 1 bed, 1 bath owner-occupied property with less than or equal to half an acre of land and 1,000 to 1,500 square feet of living area. Here we use 500 square feet living area (sfla) bins to save space when approximating the sfla coefficients. The empirical analysis uses 100 sfla bins. Column 2 presents the coefficient estimates without tokens and column 3 presents coefficient estimates when the tokens are included in the regression.

A.3 Subperiod Analysis

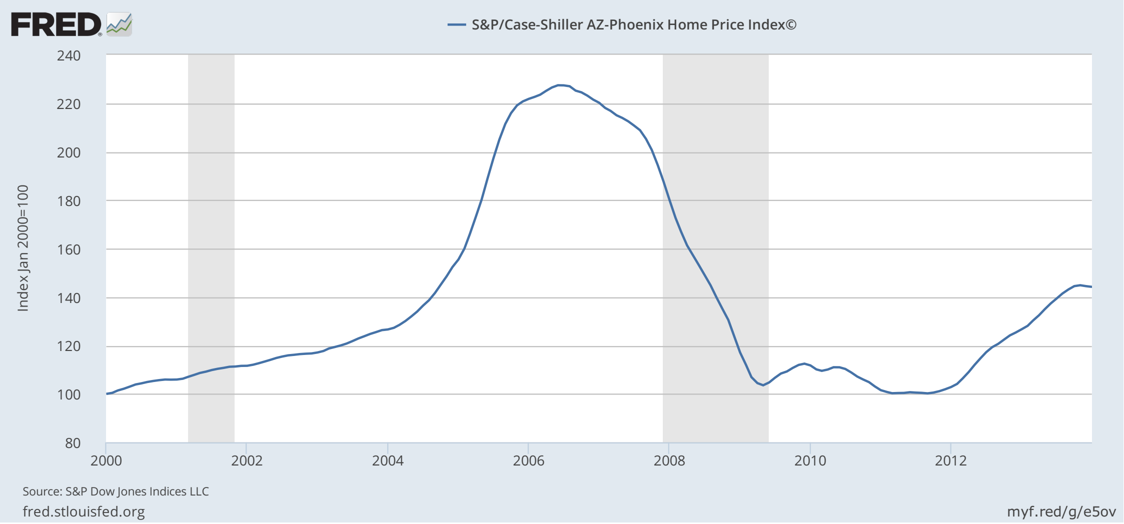
We use the Case-Shiller Repeat Sales Indices for Atlanta and Phoenix to identify the appropriate cutoffs for the subperiod analysis. The subperiods in Atlanta represent bust (2007-2011) and recovery (2012-2016) periods. The subperiods in Phoenix represent pre-boom (2000-2003), boom (2004-2006), bust (2007-2009), and recovery (2010-2013) periods. The repeat-sales indices displayed in Figure A1 and Figure A2 were downloaded from the St. Louis Federal reserve website.

Figure A1: Atlanta Case-Shiller Repeat Sales Index



Source: https://fred.stlouisfed.org/series/ATXRNSA

Figure A2: Phoenix Case-Shiller Repeat Sales Index



Source: https://fred.stlouisfed.org/series/PHXRNSA

B Tokenization Process (Internet Appendix)

B.1 Cleaning

Remarks are cleaned in the following order

1. Convert to lower case.
2. Replace commas (,) periods (.), ampersands (&) and the word *and* with ***" STOP "***. A space is placed at the beginning and end of ***STOP***.
3. Replace all special characters with a space.
4. Replace apostrophes.
5. Remove all remaining single letters.
6. Replace all numbers with a space. Numbers can be in either numeric or character form.
7. Remove repeated ***STOP***s and trim white space at the beginning and end.
8. Depluralize.

B.2 Flex-gram tokens

We use the term *flex-gram* when referring to a phrase of n words. We use the term *token* when referring to a flex-gram of arbitrary length. It is common in textual analysis to use 1-grams (unigrams), 2-grams, (bigrams), or 3-grams (trigrams). Instead of using only unigrams, bigrams, or trigrams we use the following iterative procedure that 1) identifies relevant flex-grams for arbitrary n and 2) removes k-grams ($k < n$) that are constituent parts of larger flex-grams. After n iterations, each remark will include 1-grams, 2-grams,..., $(n-1)$-grams, and n-grams. Alternatively, each remark includes tokens for flex-grams for various n.

Notation: Each remark $R^1(i)$ for $i = 1, ..., I$ is represented as a set of $j = 1, ..., J(i)$ ordered elements, words, 1-grams, or unigrams, R^1_j. Dropping the dependence on i, $R^1 = \{R^1_1, R^1_2, ..., R^1_J\}$.

This procedure is performed using the function `dasher` in `R` and is available from the authors upon request. This procedure can be performed with or without removing stop words from the cleansed remarks. Stop words are unsubstantial words in the text and include many conjunctions, prepositions, and pronouns. Examples include *me, myself, i, as, am are, a, an, the, to, and too.* We remove stop words using the English stop words dictionary in the `tm` library in `R`.

1. Set $n = 2$
2. Paste together consecutive R^{n-1}_j in R_{n-1} creating a new set Q^n where $Q^n = \{R^{n-1}_1 R^{n-1}_2, R^{n-1}_2 R^{n-1}_3, ..., R^{n-1}_{J-1} R^{n-1}_J\} = \{Q^n_1, Q^n_2, ..., Q^n_{J-1}\}$.
3. Remove any $Q^n_j \in Q^n$ that include ***STOP*** in either the first or second position.
4. Count the frequency of the Q^n_j across the $Q^n(i), i = 1, ..., I$.
5. Collect the Q^n_j that occur more than $100 < C$ times. Define this set as X^n with elements $x^n_s, s = 1, ..., S$.
6. Sort the x^n_s from most to least frequent. Beginning with the most frequent, x^n_1, replace the R^{n-1}_j in R^{n-1} with x^n_s wherever $R^{n-1}_{j-1} R^{n-1}_j = x^n_s$. Drop Q^{n-1}_j.
7. Set $R^n = R^{n-1}$.
8. Repeat steps 2-7 for $n = 3, 4, ...$ iterations but only include n-grams in the calculations in Step 4. Stop iterating when there are no n-grams that occur more than C times.
9. Remove all instances of ***STOP*** in R^n and return R^n.

B.3 Flex-gram example

Original: Exquisite home on one acre lot. Professionally decorated with many upgraded features, split floor plan, four bedrooms, spacious master suite with private exit to Pool and Courtyard. Wood shutters through out the house. Formal Living and Dining room. Gourmet Kitchen with exit to paradise backyard with endless entertaining possibilities, salt water pool and spa, built-in gas BBQ and firepit. Just sheer enjoyment with sparkling pool with raised seating area. Full cover patio. Three car garage with lots of built-in storage and generous RV parking on side of the house. True pride of ownership, this home has been impeccably maintained.

Cleaned: *exquisite home on acre lot STOP professionally decorated with many upgraded feature STOP split floor plan STOP bedroom STOP spacious master suite with private exit to pool STOP courtyard STOP wood shutter through out the house STOP formal living STOP dining room STOP gourmet kitchen with exit to paradise backyard with endless entertaining possibilities STOP salt water pool STOP spa STOP built in gas bbq STOP firepit STOP just sheer enjoyment with sparkling pool with raised seating area STOP full cover patio STOP car garage with lot of built in storage STOP generous rv parking on side of the house STOP true pride of ownership STOP this home has been impeccably maintained*

Tokenized: *exquisite home acre-lot professionally-decorated many upgraded-feature split-floor-plan bedroom spacious-master-suite private-exit pool courtyard wood-shutter house formal-living dining-room gourmet-kitchen exit paradise backyard endless entertaining possibilities salt-water-pool spa built gas-bbq firepit just sheer enjoyment sparkling-pool raised seating-area full cover-patio car-garage lot built storage generous rv-parking side house true-pride ownership home impeccably maintained*

B.4 Redundant tokens

A comparison of information provided by the standard attributes and remarks section of the MLS reveals that some of the textual information in the remarks is redundant. For example, the GAMLS and ARMLS data sets both provide an indicator variable that identifies agent-owned properties. Some of the transactions that are flagged as agent-owned also have a token in the remarks that identifies them as *agent owned, owner agent, broker owned, licensed agent,* or *seller agent.*[28] If the standard agent-owned indicator variable is correctly flagged in the MLS then the textual information in the remarks is redundant. However, if the real estate agent mistakeningly overlooks the standard agent-owned field and only includes the information in the remarks, it is not redundant.

Given that the empirical analysis focuses on agent-owned transactions we identify "agent-owned" tokens in the remarks and use them to populate/validate the standard agent-owned field. After updating the agent-owned field we remove the "agent-owned" tokens because they are redundant. There are a total of 2,003 agent-owned transactions in the Atlanta data. Of which, 64 (3.2%) are flagged in both sections, 1,890 (94.4%) are flagged only by the standard attribute field, and 49 (2.4%) are flagged only in the remarks. There are 16,654 agent-owned transactions in the Phoenix data. Of which, 2,092 (12.6%) are flagged in both sections, 13,491 (81.0%) are flagged only by the standard attribute field, and 1,071 (6.4%) are flagged only in the remarks.

When running the vacant discount analysis in Table 9 we remove tokens that include the word *vacant.* The removal of redundant tokens has to be carefully orchestrated as some tokens have unexpected dual meanings. For example, the authors considered removing tokens that include the word *empty* when estimating the vacant discounts. However, doing so would unintentionally flag houses that were marketed to *empty-nesters* as *vacant.*

[28] This is an abridged list of the agent-owned tokens we identify. A complete list is available by request. We suspect that the list of agent-owned keywords will vary by geography and across data sets. For that reason we recommend a careful examination of the tokens and a custom built market specific agent-owned token list.

B.5 Double-selection token tables

The double-selection procedure that we employ selects tokens that can be used to (i) predict house prices and/or (ii) identify agent-owned transactions in the data. The following tables display the top twenty tokens based on the magnitude of the absolute value of the token's coefficient. There are two tables for both Atlanta and Phoenix. The tokens selected by the hedonic equation in Tables B1 and B3 can be used to predict house prices in Atlanta and Phoenix, respectively. Whereas, the tokens selected by the linear probability equation in Tables B2 and B4 can be used to identify agent-owned transactions in Atlanta and Phoenix, respectively.

Table B1: Selected Tokens from the Hedonic Equation (Atlanta)

Unigrams		Bigrams		n-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
townhome	-0.21	investor-special	-0.21	investment-opportunity	-0.19
dock	0.18	sold-asis	-0.20	dock	0.19
investor	-0.14	fixer-upper	-0.19	need-tlc	-0.17
investment	-0.14	tenant-occupied	-0.19	investor	-0.16
fixer	-0.12	need-tlc	-0.18	investment-property	-0.16
gated	0.11	sld-asis	-0.18	sold-asis	-0.15
asis	-0.09	need-work	-0.17	sold-asis-condition	-0.15
craftsman	0.09	investment-opportunity	-0.15	investment	-0.13
potential	-0.09	lot-potential	-0.15	cash	-0.13
tlc	-0.09	investment-property	-0.14	gated	0.12
coffered	0.09	split-foyer	-0.12	tlc	-0.12
cash	-0.09	great-investment	-0.12	fixer-upper	-0.12
restored	0.09	avondale-estate	0.12	craftsman-bungalow	0.11
charleston	0.09	split-level	-0.11	split-level	-0.10
eav	0.09	great-opportunity	-0.11	asis	-0.10
earthcraft	0.09	pre-qual	-0.11	potential	-0.10
tenant	-0.08	coffered-ceiling	0.11	split-level-home	-0.10
finishes	0.08	lock-box	-0.11	repair	-0.10
repair	-0.08	market-village	0.11	coffered-ceiling	0.10
ashford	0.08	submit-offer	-0.11	best-street	0.10

Table B2: Selected Tokens from the Linear Probability Equation (Atlanta)

Unigrams		Bigrams		n-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
realize	0.36	just-renovated	0.24	new-fixture	0.16
highlight	0.12	renovated-brand	0.20	complete-renovation	0.13
taking	0.11	house-just	0.19	entire	0.12
business	0.09	elegant-home	0.12	new-granite-countertops	0.11
friend	0.07	room-area	0.10	highlight	0.09
gathering	0.06	new-fixture	0.08	home-spacious	0.08
entire	0.04	complete-renovation	0.08	new-stainless-steel-appliance	0.07
advantage	0.04	bathroom-home	0.07	friend	0.07
fixture	0.03	entire-house	0.07	large-closet	0.07
brand	0.03	renovated-new	0.06	home-great-neighborhood	0.07
travertine	0.03	home-just	0.06	gathering	0.07
wonderfully	0.03	new-appliance	0.03	new-granite-counter	0.06
renovated	0.02	home-spacious	0.03	new-appliance	0.04
just	0.02	just-waiting	-0.03	bedroom-bathroom	0.04
renovation	0.02	brand-new	0.02	personal	0.04
dream	0.02	new-granite	0.02	new-home	0.03
around	0.02	recently-renovated	0.02	around	0.03
personal	0.02	new-flooring	0.02	wonderfully	0.03
create	-0.02	new-home	0.02	renovated	0.02
pottery	-0.02	spacious-living	0.02	new-paint	0.02

Table B3: Selected Tokens from the Hedonic Equation (Phoenix)

Unigrams		Bigrams		n-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
dock	0.13	th-fairway	0.13	th-fairway	0.12
waterfront	0.11	lake-view	0.13	fairway	0.08
fairway	0.10	sold-asis	-0.08	barn	0.08
cash	-0.09	sub-zero	0.08	fix	-0.07
historic	0.07	golf-course	0.07	golf-course-lot	0.07
barn	0.07	city-light	0.07	golf-course-view	0.07
membership	0.06	huge-loft	-0.07	golf-course	0.06
hillside	0.06	spectacular-view	0.07	finishes	0.06
viking	0.06	panoramic-view	0.07	historic	0.06
disturb	-0.06	travertine-floor	0.06	sold-asis	-0.06
granite	0.05	large-loft	-0.06	large-loft	-0.06
course	0.05	outdoor-living	0.06	biltmore	0.06
loft	-0.05	guard-gated	0.06	granite-counter	0.05
preserve	0.05	custom-cabinetry	0.06	loft	-0.05
irrigated	0.05	large-home	-0.06	guest-house	0.05
fix	-0.05	granite-counter	0.05	city-light	0.05
finishes	0.05	guest-house	0.05	travertine-floor	0.05
biltmore	0.05	pebbletec-pool	0.05	course	0.05
tee	0.05	mountain-preserve	0.05	irrigated	0.05
vintage	0.05	course-community	-0.05	salt-water-pool	0.05

Table B4: Phoenix - Tokens Selected by Linear Probability Model

Unigrams		Bigrams		Flex-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
rubbed	0.07	new-tone	0.09	new-stainless-steel-appliance	0.09
tone	0.05	tone-paint	0.07	arizona	0.07
remodel	0.04	new-ceramic	0.07	new-ceiling-fan	0.06
epoxy	0.03	new-upgraded	0.07	remodel	0.05
hardware	0.03	new-stainless	0.06	tone-paint	0.05
lease	0.03	new-lighting	0.06	new-carpet	0.04
state	0.03	new-ceiling	0.05	new-paint	0.04
carry	0.03	new-berber	0.05	hardware	0.04
new	0.02	complete-remodel	0.05	new-paint-inside	0.04
paint	0.02	new-carpet	0.04	epoxy	0.04
granite	0.02	new-paint	0.04	new-interior	0.04
remodeled	0.02	new-kitchen	0.04	brand-new	0.03
stainless	0.02	new-custom	0.04	faucet	0.03
fixture	0.02	new-countertops	0.04	new-kitchen	0.03
travertine	0.02	can-close	0.04	inch	0.03
pride	-0.02	stainless-appliance	0.03	fresh-paint-inside	0.03
completely	0.02	granite-slab	0.03	tone	0.03
faucet	0.02	new-granite	0.03	new-door	0.03
tenant	0.02	garage-floor	0.03	plumbing-fixture	0.03
baseboard	0.02	never-lived	0.03	new	0.02

C Robustness Checks (Internet Appendix)

C.1 Positive and negative tokens

Although agents are hired to represent the seller, they have to be truthful in their marketing of the property. Listing agents that market a landlocked property in disrepair as "*immaculate with a dock*" will develop a poor reputation. Listing agents also know that prospective buyers may be looking for a *fixer upper* or are willing to put in some *sweat equity.* Thus, if they mention these features in the textual description of the property they will attract the right kind of buyer and increase the likelihood of a sale. Given that listing agents only get paid if the property sells, they have an incentive to properly market the property. In addition, the textual description of the property is often accompanied by photos which allows prospective buyers to validate certain aspects of the property description from afar.

Table C1 examines the positive and negative informational content provided in the public remarks section of the MLS using the occupied housing subsample. Columns 1 to 4 include the standard controls and time by zip code fixed effects. Column 1 does not include the textual information. Column 2 includes both the positive and negative tokens. Whereas column 3 (4) only includes the positive (negative) tokens. Columns 5 to 8 are set up similarly except for the use of multiplicatively separable time and census tract fixed effects. The results show that both the positive and negative tokens address a portion of the omitted variable bias present in the agent-owned estimates. The fact that the estimates are lower in columns 3 and 7 suggests that the agents are more likely to include positive information about the house. Agents do, however, include negative information about the house which, when included alongside the positive information, outperforms the positive information on its own.

Table C1: Agent-owned estimates by token type

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Atlanta	0.029**	0.014	0.017	0.030*	0.031**	0.018	0.020*	0.030**
	(0.016)	(0.014)	(0.014)	(0.016)	(0.014)	(0.013)	(0.012)	(0.015)
Phoenix	0.035***	0.016***	0.022***	0.032***	0.031***	0.014***	0.019***	0.028***
	(0.005)	(0.003)	(0.004)	(0.005)	(0.003)	(0.003)	(0.003)	(0.003)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓	✓				
Time x Tract FE					✓	✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	$\hat{S}_2^{zip^+}$	$\hat{S}_2^{zip^-}$		$\hat{S}_2^{zip}$	$\hat{S}_2^{zip^+}$	$\hat{S}_2^{zip^-}$

*p<0.1; **p<0.05; ***p<0.01

C.2 Heterogeneous subsamples

An agent's informational advantage may vary across neighborhoods. As such, we examine whether the agent-owned estimates vary by neighborhood composition. Income is often used as a proxy for education, so its possible that agent's enjoy a greater informational advantage in lower income neighborhoods relative to higher income neighborhoods. Columns 1 to 4 in Table C2 examine whether agent-owned estimates vary across income levels. The sample is divided into low and high income neighborhoods using the 2010 census median income measure. Columns 1 and 2 provide estimates for the low income neighborhoods and columns 3 and 4 provide estimates for the high income neighborhoods. Even (odd) columns (do not) include the textual information from the public remarks. The agent-owned estimates in Atlanta are insignificant when the textual information is included. Whereas, the agent-owned estimates in Phoenix are significant for both income levels. The results suggest that agent-owned houses sell for a higher premium (as a percent of sales price) in lower income neighbhorhoods.

Levitt and Syverson (2008) note that the degree of heterogeneity in the housing stock may play a role in the informational advantage that agents have. In neighborhoods where the housing stock is homogeneous, sellers can accurately estimate their houses' value by looking at recent sales nearby. However, if the housing stock is heterogeneous, it will be more difficult for sellers to estimate the value of their own house. In columns 5 to 10 we proxy for neighbhorhood heterogeneity using the average difference in square feet of living area within the census block group. Similar to Levitt and Syverson (2008), when the textual information is not included in the model, we find that the sales price difference between agent-owned and non-agent-owned houses is highest in neighborhoods that are more heterogeneous. However, after we incorporate the textual information the difference is minimal. The results suggest that including the textual information in the asset pricing model helps control for heterogeneous nature of the housing stock.

Table C2: Agent-owned estimates across heterogeneous subsamples

	Income				Heterogeneity					
	Low		High		Low		Medium		High	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Atlanta	0.018	−0.013	0.037***	0.016	0.022**	0.015	0.032*	0.019	0.038**	0.013
	(0.020)	(0.024)	(0.009)	(0.012)	(0.010)	(0.014)	(0.018)	(0.019)	(0.019)	(0.020)
Phoenix	0.054***	0.031***	0.033***	0.012***	0.029***	0.014***	0.043***	0.019***	0.039***	0.017***
	(0.007)	(0.007)	(0.005)	(0.004)	(0.005)	(0.005)	(0.006)	(0.005)	(0.004)	(0.005)
Controls	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Tokens		$\hat{S}_2^{zip}$		$\hat{S}_2^{zip}$		$\hat{S}_2^{zip}$		$\hat{S}_2^{zip}$		$\hat{S}_2^{zip}$

*p<0.1; **p<0.05; ***p<0.01

C.3 Sample Selection

The agent-owned premium estimates we report use a sample of houses that were listed and successfully sold on the MLS. They do not, however, incorporate information from houses that were listed on the MLS that did not sell. To address this concern we use the Heckman (1979) selection correction method. The selection correction model estimates a probit model using both sold and unsold records using a sold indicator variable as the dependent variable. The probit estimation includes property controls and time by zip code fixed effects. The probit results are used to construct an inverse Mills ration (IMR) that is included as an additional control. The agent-owned estimates in Table C3 are similar to those reported using a specification without the IMR.

Table C3: Selection correction agent-owned estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Atlanta	0.034**	0.018	0.026**	0.033**	0.021	0.025*
	(0.014)	(0.012)	(0.013)	(0.014)	(0.012)	(0.013)
Phoenix	0.036***	0.017***	0.028***	0.032***	0.015***	0.025***
	(0.005)	(0.003)	(0.004)	(0.003)	(0.003)	(0.003)
Controls	✓	✓	✓	✓	✓	✓
IMR	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓			
Time x Tract FE				✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	L&S		$\hat{S}_2^{zip}$	L&S

*p<0.1; **p<0.05; ***p<0.01

C.4 Two-stage least squares

There is a rich literature on the co-determination of sales price and time-on-market. In this section we re-estimate the agent-owned premiums using a two-stage least squares (2SLS) regression. The first stage of the 2SLS is a regression on time-on-market that includes a degree of overpricing covariate to address the exclusive restriction. The 2SLS agent-owned estimates are reported in Table C4 for the occupied housing subsample. Every column incorporates a set of controls that include the age of the house and indicator variables for bedrooms, bathrooms, living area, and lot size. Columns 1 to 3 include time by zip code fixed effects and columns 4 to 6 include time by census tract fixed effects. Columns 1 and 4 do not use any tokens. Columns 2 and 5 use the unigram token set $\hat{S}_2$. Column 3 and 6 use the token set described in Levitt and Syverson (2008). The estimates in Table C4 are similar to those reported using an OLS specification without controlling for time-on-market.

Table C4: 2SLS agent-owned estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Atlanta	0.031**	0.015	0.023	0.033**	0.019	0.025*
	(0.016)	(0.014)	(0.016)	(0.014)	(0.013)	(0.013)
Phoenix	0.037***	0.015***	0.029***	0.036***	0.015***	0.029***
	(0.005)	(0.003)	(0.005)	(0.004)	(0.003)	(0.004)
Controls	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓			
Time x Tract FE				✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	L&S		$\hat{S}_2^{tract}$	L&S

*p<0.1; **p<0.05; ***p<0.01