Asymmetric or Incomplete Information about Asset Values?

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

September 25, 2018

Abstract

We provide a new framework for using text as data in empirical models. The framework identifies salient information in unstructured text that can control for multidimensional heterogeneity among assets. We demonstrate the efficacy of the framework by reexamining principal-agent problems in residential real estate markets. We show that the agent-owned premiums reported in the extant literature dissipate when the salient textual information is included. The results suggest the previously reported agent-owned premiums suffer from an omitted variable bias, which prior studies incorrectly ascribe to market distortions associated with asymmetric information.

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

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)

[*] Cornell University, SC Johnson College of Business

[†] West Virginia University, College of Business & Economics

[‡] San Diego State University, Fowler College of Business; Contact author: 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, McKay Price, Jenny Schuetz, and Andrey Ukhov for helpful comments. Errors are the responsibility of the authors.

1 Introduction

Many variables in finance and economics are conceptually understood but not easily quantifiable. For example, product quality and managerial ability are often clearly understood in theoretical models, but explicit measurements of these variables are rarely, if ever, available. Although difficult to quantify, these variables cannot be ignored when they are correlated with a variable of interest, since their omission introduces an omitted variable bias. Accompanying bodies of text, and the words and phrases (tokens) contained therein, that either indirectly or directly discuss these variables may provide the information necessary to address this bias. This paper provides a widely applicable, data-driven framework for incorporating salient textual information into empirical models. In a specific application, we demonstrate how the framework can resolve a significant but avoidable bias resulting from textual information that was available, but omitted in previous studies that examine agents' informational advantage in residential real estate markets. After including the salient textual information in the model, we show that the previously reported market distortions associated with asymmetric information dissipate.

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 (one of which is agent-owned transactions). 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.

The extant literature frequently uses agent-owned transactions as an indirect information variable since it offers a relatively clean identification strategy and real estate agents have access to private information.[1,2] Given the idiosyncratic nature of the residential housing stock, a natural concern is that the coefficient estimate for agent-owned transactions is biased absent sufficient controls. Fixed-effects, when employed at the appropriate level, can control for time-invariant attributes that are difficult to quantify. However, property-specific fixed effects will not resolve a bias if real estate agents are more likely to make improvements prior to listing their property for sale. That is, uncontaminated estimates of the value of an informational advantage are only possible after controlling for all relevant public information, both time-varying and time-invariant, regardless of whether or not it is difficult to quantify. The seminal studies by Rutherford et al. (2005) and Levitt and Syverson (2008) were aware of this and report their findings with a caveat that their agent-owned estimates may be biased. 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 when it is only available in a high-dimensional format such as text.[3]

Textual analysis is not a new concept in finance and economics, but several factors have prevented its widespread use.[4] Most notably, a numeric representation of the requisite infor-

[1] 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. Although we cannot control for which individuals choose to become real estate agents in order to gain a perceived informational advantage, we do filter out real estate agents with more than three agent-owned transactions to remove investors that become real estate agents for the express purpose of flipping and/or renting their personal properties.

[2] Real estate agents' private information includes, but is not limited to information acquired through the following channels: private negotations, the number and distribution of unsuccessful bids, delisted properties, and pocket listings. See Levitt and Syverson (2008) for further discussion.

[3] A textual description of the property was not available in all of the aforementioned real estate studies. 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.

[4] 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 (Jegadeesh and Wu, 2013). However, recent studies have also used textual information to examine topics such as financial constraints (Hoberg and Maksimovic, 2014; Bodnaruk et al., 2015; Buehlmaier and Whited,

mation in the text is not readily available. To overcome this obstacle researchers typically specify, ex-ante, a dictionary of words that are associated with a given topic or sentiment and then use the dictionary to map the text into a numeric index that is included in the empirical model. Our framework differs from the extant literature in two distinct ways. First, we use a data-driven approach that does not require a pre-specified dictionary. Second, we include textual information about the asset (not the sentiment of the text) in the empirical model.

Absent an existing well-defined dictionary, implementing textual analysis presents a challenge in terms of both accuracy and costs.[5] This is a concern since Loughran and Mcdonald (2011) note that dictionaries appropriate for one setting may perform poorly in another. Our framework addresses both of these challenges *and* mitigates biases stemming from omitted variables - which is particularly important given the caveats noted in Rutherford et al. (2005) and Levitt and Syverson (2008).[6] Our data-driven approach is similar to Nowak and Smith (2017) in that we do not use a pre-specified dictionary. Nowak and Smith (2017) demonstrate the ability of LASSO to select variables for house price predictions, but do not investigate whether the single-selection LASSO procedure they employ can resolve an asymptotic bias in parameter estimates.[7] Although single-selection LASSO has desirable predictive properties, Belloni et al. (2014) indicate single-selection procedures may not resolve a bias. For this reason, we use a double-selection LASSO procedure that explicitly targets the resolution of an omitted variable bias, thereby allowing for valid asymptotic inference on a parameter of interest. To the best of our knowledge, this is the first study to provide a data-driven

2018). See Loughran and McDonald (2016) for a recent survey of the textual analysis literature.

[5] The high-dimensional nature of text precludes statistical inference using standard econometric techniques for variable selection (Gentzkow et al., 2017). For this reason, researchers created previous dictionaries using a manual process that was both time consuming and susceptible to human error. King et al. (2017) confirm this conjecture in their finding that humans perform poorly when creating dictionaries from scratch, yet perform well when associating words to topics.

[6] Omitted variable bias concerns are ubiquitous in the finance and economics literature. For example, Agarwal et al. (2011) find that securitization hampers mortgage renegotiation, but note that uncertainty induced by Pooling and Servicing Agreements (PSAs) may bias their results. Our framework can incorporate the available, but difficult to quantify textual information in PSAs to examine its effect on servicers' decisions to renegotiate distressed loans. Then, the researcher can determine whether (i) the securitization process, itself, hampers renegotiation or (ii) the PSAs hamper the ability of the servicers to renegotiate.

[7] LASSO is short for least absolute shrinkage and selection operator.

textual analysis framework that mitigates the omitted variable bias stemming from textual information that is frequently available, but not included in empirical models since it is not easily quantifiable.

We create several market-specific real estate dictionaries for Atlanta, GA and Phoenix, AZ using the remarks section of their local 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.[8] 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 incentives, and/or neighborhood amenities that are not easily quantifiable or indicated in the standard MLS data fields. We create separate, market-specific dictionaries since real estate markets are highly localized, as evidenced by the fact that less than 50 percent of the tokens in the Atlanta and Phoenix dictionaries overlap.

The dictionaries we create include both positive and negative information about the condition and quality of the house and neighborhood. More importantly, the tokens include attributes that are binary in nature, objective, observable, and verifiable. The tokens that strongly predict agent-owned transactions identify low intensity property improvements (new paint, new appliances, etc.) that are not strong predictors of price. This finding further motivates the double-selection procedure we employ, since omitting these tokens from the model generates a positive bias in the agent-owned coefficient estimate. The double-selection method selects approximately 600 tokens in each of the market-specific dictionaries we create. In contrast, Levitt and Syverson (2008) use a dictionary of 61 words and phrases. This

[8] 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."

difference in the number of tokens reflects more the value of machine learning and other high-dimensional methods rather than the limitation of pre-defined dictionaries.

Sans textual information, we estimate a 3 to 4 percent premium for agent-owned transactions in Atlanta and Phoenix. These naive estimates are similar to those reported in Rutherford et al. (2005) and Levitt and Syverson (2008). However, after we incorporate the publicly available 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. These null results are bolstered by simulations that indicate that the procedure is quite powerful and can reliably detect economically significant price effects in both Atlanta and Phoenix. We also show that these findings are not a result of overfitting and that the framework performs well in out-of-sample tests.

In addition to providing a new methodology, our empirical application of the methodology makes several important contributions to the literature. Contrary to previous studies, 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 publicly available, but difficult to quantify, textual information from the remarks in the pricing model explains a large portion of the naive agent-owned premium. Previous studies attribute the premium to asymmetric information. However, we show that a large portion of the naive premium reflects a bias related to omitted variables. These findings highlight (i) the difficulty in testing for information asymmetry using incomplete information about heterogenous assets and (ii) the need to include textual information in empirical models in order to more closely align the information set in the model with the information set of the market participants involved in the transaction.

The plan of our analysis is as follows. Section 2 describes the framework and methodology that we employ. 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

Implicitly, a home buyer and seller negotiate a sales price based on a set of property attributes that are observable to both 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 viewing the property online, visiting the property, reviewing property improvement and tax records, hiring a professional home inspector, and consulting with their real estate agent.[9] In this sense, X is fully revealed to both the buyer and seller, so there is no information asymmetry between the market participants.[10]

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

[9] 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").

[10] 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 has an informational 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' informational advantage or test whether the advantage can be exploited for financial gain. For these reasons we focus on incorporating the textual information that is available to all market participants in the pricing model.

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 interest 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 that the listing agent provides 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. However, 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 *custom kitchen* and *beautiful in ground pool & spa*. In contrast, the less expensive house *needs cosmetic fix-up*, has a pool that *needs to be replastered*, and is *vacant*. This simple example highlights the property-observable textual information contained in the remarks and the need to include it in the 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*).[11] 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 stemmed words. Recognizing that some of the textual information in the remarks is repetitive, we also remove tokens that are redundant for the variable of interest.[12]

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 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

[11] 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.

[12] 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 also experimented with other spell checking and stemming packages. The use of the spell checking and stemming packages reduces the number of tokens included in the dictionary. However, they have no effect on the agent-owned coefficient estimates we report. In addition, the agent-owned coefficient estimates do not change when we use the raw text. A step-by-step overview of the tokenization process and removal of redundant tokens is provided in an internet appendix.

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 tokens define a set of *candidate 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 the 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 client-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.[13] 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.

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

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 this 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 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 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 $\mu_n + \psi_{nt}$. Thus, the tokens serve 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.[14] More importantly,

[14] 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.[15] 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

[15] The ℓ_1 length of a $K \times 1$ vector x is $\|x\|_1 = \sum_k |x_k|$. λ_p, λ_d are determined using the penalty parameters ($c = 1.10$, $\gamma = 0.10$) recommended in Belloni et al. (2014). In an internet appendix, we show that the choice of penalty parameters affects $\hat{S}$ and $\hat{Q}$, but not $\hat{\tau}$.

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

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

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.[16] 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 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 Section A.1 of the appendix. 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.

[16] 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.

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.[17]

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.

[17] 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.[18] 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 421 (548) of the 2,000 candidate tokens that explain price in $\widehat{S}_p^{tract}$ and 223 (126) 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 (Phoenix) data set. In total, there are 624 (615) unique tokens from both of these sets in the Atlanta (Phoenix) $\widehat{S}_2^{tract}$ dictionary. When using additively separable time and zip code fixed effects the process selects 586 (613) unique tokens in the Atlanta (Phoenix) $\widehat{S}_2^{zip}$ dictionary. The results show that the within-city $\widehat{S}_2^{tract}$ and $\widehat{S}_2^{zip}$ dictionaries are highly correlated (88.5%) especially in their selection of the token set in $\widehat{S}_d$ (99.7%). In unreported results we create the dictionaries for every subperiod in Atlanta and Phoenix. The agent-owned subperiod estimates remain the same whether we use the dictionaries for the entire study period or the market-specific subperiods. For this reason, we only report estimates using the dictionaries for the entire study period.[19]

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

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

[19] In an internet appendix we recreate the dictionaries using a different subsample of data and various penalty parameters. We show that these changes do not affect the agent-own coefficient estimates we report.

determined by the true condition and quality of the property. For example, removing the phrase *fixer upper* from a description while not making any repairs to the property is unlikely to increase its sales price.

For informational purposes, Figure 1 plots the ten largest positive and negative unigram 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, six of the ten positive unigram tokens selected using zip code fixed effects are neighborhoods in Atlanta (Walton, Ormewood, Collier, Ashford, Grant, and Vinings). Similarly, five of the ten positive tokens selected using census tract fixed effects are neighborhoods - although the neighborhoods selected are not identical (Chastain, Brookhaven, Ashford, Walton, and Buckhead). This suggests that both zip code and census tract 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. We define and provide an example of every unigram token listed in Figure 1 in an internet appendix.

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 client-owned houses. In Atlanta (Phoenix), 70.7 (75.2) percent of the tokens selected in $\widehat{S}_2^{tract}$ and 71.3 (72.7) percent of the tokens selected in $\widehat{S}_2^{zip}$ are positive. The ratio of positive-to-negative tokens is relatively constant across the market-specific subperiods. We examine the extent to which the positive and negative tokens in the MLS remarks address the bias associated with agent-owned sales transactions in an internet appendix. A high level comparison of agent-owned versus client-owned remarks suggests that agents expend a similar amount of effort writing remarks for client-owned houses. For example, in Phoenix the average length of an agent-owned remark (437 characters) is similar to a client-owned remark (426 characters).

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. For clarity, only twenty of the strongest predictors in either $\hat{S}_p$ or $\hat{S}_d$ are presented. Properties with *granite-slab* or *stainless-appliances* sell for a higher price and are more likely to be agent-owned. In contrast, properties sold via an *estate-sale* sell for a lower price and are less likely to be agent-owned. This finding aligns closely with Campbell et al. (2011) who find that death-related discounts (estate-sale) reflect poor maintenance. 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.[20]

By construction, tokens in $\hat{S}_p$ explain large variations in transaction price, and many 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 improve their property. However, because these tokens are not in $\hat{S}_p$, these improvements 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 improvements are time-varying and do not require a permit, they most likely are not included in county tax assessor data sets.

It is also 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 real estate studies that used a pre-specified dictionary included words related to subjective measures of property quality in the dictionary. For example, the dictionary in Levitt and Syverson (2008) includes *amazing, fantastic*, and *tasteful* while the dictionary in Goodwin et al. (2014) is comprised of 17 positive words such as *beautiful, cozy, gorgeous*, and *lovely*.[21] Furthermore, because these features are verifiable and posted

[20] *th-fairway* is an artifact of the cleaning procedure where all numbers are removed from the remarks and can indicate the 4th, 5th,. . . , or 18th fairway. See the internet appendix for further details of the cleaning procedure.

[21] 55 of the 59 tokens (93.2%) in the Levitt and Syverson (2008) dictionary are included in the K = 2,000

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 Figures 1 and 2 do not appear in dictionaries that are not specific to real estate, and even when they do the words often take on a different meaning. 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.[22] This finding further emphasizes the need for researchers to use a dictionary specific to the research question at hand. This is especially important when an asset, such as real estate, trades in highly localized markets as we find that less than 50 percent of the unigram tokens selected for the Atlanta dictionary are included in the Phoenix dictionary.

4.2 Agent-Owned Estimates

When comparing agent-owned and client-owned houses it is 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.[23] 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 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

most frequent tokens for Phoenix. However, only 34 of the 59 tokens (57.6%) are included in the $\widehat{S}_2^{zip}$ dictionary. Although the Levitt and Syverson (2008) dictionary has 61 tokens, we remove *bank owned* and *foreclosure* transactions prior to running the empirical analysis. For this reason, we did not include these tokens in the comparison.

[22] This is similar to the finding in Loughran and Mcdonald (2011) that almost three-fourths of the negative words in the Harvard IV-4 Dictionary are attributable to words that are typically not negative in a financial context.

[23] Following Levitt and Syverson (2008), we filter out houses that sold more than two times within a three year period. In an internet appendix, we rerun the empirical analysis using only the subsample of houses that sold more than two times within a three year period. The agent-owned coefficient estimates are similar to those reported in Table 4.

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 client-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 controls for 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 4.1% 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, the agent-owned estimates reported in column 1 likely suffer from a model-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 (i.e. tokens in $\widehat{S}_2^{zip}$) in column 2, the agent-owned estimate is no longer statistically significant and is statistically different than the naive baseline coefficient estimate in column 1 at the 10 percent level.[24] Column 3 includes the dictionary described in the appendix of Levitt and Syverson (2008). The L&S dictionary reduces the magnitude of the agent-owned estimate, although it remains statistically significant at the 10% level. Also note that the $\widehat{S}_2^{zip}$ dictionary reduces the vacant (rental) coefficient estimate from -12.0% (-3.8%) in column 1 to -7.4% (-1.7%) in column 2. In contrast, the L&S dictionary has a smaller impact on the vacant coefficient estimate and actually increases the rental discount from -3.8% to -4.9%. This is probably due to the fact that the L&S dictionary includes mostly "positive"

[24] Formally, we reject $H_0 : \tau = 0.033$ at the 10% level. Where applicable, similar tests relative to baseline estimates are indicated with a † going forward.

tokens.

Similar results are reported in Panel B for Phoenix. The agent-owned estimate drops from 4.1% to 1.7% when the textual information from the remarks is included. Although the agent-owned estimate remains statistically significant, we reject the null hypothesis that the coefficient estimate in column 2, which includes the tokens, is equal to the coefficient estimate in column 1. In other words, the naive baseline estimate in column 1 suffers from a model-omitted variable bias that we address using our textual analysis framework. After we include the textual information in column 2, the naive agent-owned coefficient estimate decreases by over 58 percent.

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 the textual information from the remarks, we naively estimate an agent-owned premium in column 4 of 3.3% for Atlanta and 3.6% for Phoenix. When the tokens in $\widehat{S}_2^{tract}$ are included in column 5, the agent-owned premium drops to 1.7% in Atlanta and 1.5% in Phoenix. Atlanta's agent-owned coefficient estimate in column 5 is not statistically significant and not statistically different than the naive estimate in column 4. This finding suggests that the statistical significance in column 4 is an artifact of the model-omitted variable bias that our framework addresses. After adjusting for this bias, the agent-owned estimate is no longer statistically significant. This finding demonstrates that ignoring available, but difficult to quantify textual information can lead to spurious results.

In contrast to Atlanta, the coefficient estimates in Phoenix are statistically significant after including the textual information in columns 2 and 5. However, every token-adjusted estimate of τ in Table 4, and the subsequent analysis, is significantly closer to zero than its naive counterpart. In other words, the naive coefficient estimates in columns 1 and 4 are heavily biased. Also note that the length of the agents' remarks are limited, so agents cannot include textual information about *every* housing attribute that affects price. We show that agents can (and do) provide a substantial amount of textual information that, when included in the pricing model using our framework, either affects the statistical significance of the model, significantly reduces the coefficient estimate, or both.

4.2.1 Alternative Token Sets and Overfitting

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 client-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 3.6% in Phoenix. Column 2 includes the 586 (613) 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. To answer this, we include 1,414 (1,387) of the 2,000 candidate tokens that are 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 2.2% for Atlanta and 3.6% to 2.8% 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 92.1%. Whereas, the $\widehat{S}_2^{zip}$ token set increases the R^2 to 93.6%.

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 by performing three experiments. First, we permute the remarks by randomly drawing the remarks without replacement and treating these remarks as the true remarks. Then, we create token indicators based on the $\widehat{S}_2^{zip}$ token set that was created using the non-permuted remarks. Results for a random permutation are reported in column 4. The agent-owned premium in column 4 is nearly identical to the coefficient estimate in column 1.[25] Thus, it does not appear as though the estimates in column 2 are the result of overfitting noise in the data. Rather, the approach we describe can accurately identify a subset of relevant tokens from a set of many candidate tokens using an ℓ_1 penalty and suitable penalty parameters.

Second, we investigate the sensitivity of the results to the size of the penalty parameters. In Equations 5 and 6, $\hat{Q}_d$ and $\hat{Q}_d$ are determined by the penalty parameters $\phi_{p,k}, \phi_{d,k}$ and λ_p, λ_d. Belloni et al. (2012) describe a feasible, data-driven procedure for choosing $\phi_{p,k}, \phi_{d,k}$ that implicitly relies on the choice of λ_p, λ_d. The tuning parameters $1 < c$ and $\gamma \approx 0$ determine λ_p, λ_d via $\lambda_p = \lambda_d = 2c\sqrt{N}\Phi^{-1}(1 - \gamma/2P)$ where Φ^{-1} is the inverse cdf of the standard normal and P is the total number of parameters in the model.[26] As λ_p, λ_d increases (decreases), the penalty on the token coefficients increases (decreases) and fewer (more) tokens are selected. In this study, we use the values $c = 1.1$ and $\gamma = 0.1$ suggested in Belloni et al. (2012). In order to determine if the results are sensitive to these tuning parameters, we consider $c \in \{1.05, 1.10, 1.25\}$ and $\gamma \in \{0.01, 0.10, 0.25\}$. Results provided in an internet appendix indicate that $\hat{Q}_d$ and $\hat{Q}_d$ vary in a predictable way as the penalty changes. More importantly, $\hat{\tau}$ is not sensitive to these alternative tuning parameters. This experiment indicates that the results using $c = 1.1$ and $\gamma = 0.1$ are not sensitive to alternative tuning parameters that increase or decrease the penalty on the token coefficients.

Third, we investigate the extent to which $\hat{S}_2$ can be used for out-of-sample estimation. The previous two experiments indicate that in-sample estimates of τ are neither spurious nor the result of an under-penalized and thus overfit model. To examine the out-of-sample performance of our approach we (i) randomly split the data in both cities into equally sized training (A) and testing (B) subsamples, (ii) identify the relevant set of tokens in the training

[25] Additional permutations of the remarks produce nearly identical coefficient estimates and are available upon request.

[26] The results are not sensitive to the alternative $\lambda_p = \lambda_d = 2c\sqrt{N}\Phi^{-1}(1 - \gamma/2P \max \log N, \log P)$.

subsample ($\hat{S}_2^A$), and (iii) use the tokens in $\hat{S}_2^A$ as a set of controls when estimating τ using the transactions in B. The results, which we provide in an internet appendix, are nearly identical to those reported in Tables 4 and 5.

As noted earlier, the bulk of the analysis reported in this study uses market-specific unigram dictionaries. 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 an 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.

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 an internet 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 information asymmetry and 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 informational 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 to 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 statistically significant. The results validate our earlier findings that agent-owned houses did not sell for a premium relative to client-owned houses in Atlanta from 2007-2016. The results highlight the fact that our approach effectively incorporated the textual information from the agents' remarks to control for the variability in location, physical characteristics, and potential agency issues. The results also highlight the need to include 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 column 8, the agent-owned estimates are no longer statistically 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 bias that is present in previous studies.

4.2.3 Power Simulations

We run a series of simulation experiments to assess the finite sample performance of the asymptotic theory behind the post double-selection estimator for agent-owned premiums. We focus on the 0 to 3 percent range that corresponds with the agent-owned coefficient estimates reported in this study and the related literature. For computational considerations and, more importantly, because the agent-owned coefficient estimate is statistically insignificant in Table 6, we run the simulations using the entire period in Atlanta (2007-2016) and the pre-boom subperiod in Phoenix (2000-2003). Based on the parameter estimates in the data, we simulate 500 agent-owned indicators and prices for each $\tau \in \{0.000, 0.005, \ldots, 0.030\}$ in both cities. We then perform the post double-selection procedure on the simulated data sets. A full description of the simulation procedure is provided in the appendix.

Figures 3a and 3b report the fraction of simulations where $H_0 : \tau = 0$ is rejected at the 5% and 10% significance level for each value of τ in Atlanta and Phoenix, respectively. The results indicate the procedure is quite powerful for both cities. In Atlanta (Phoenix), the procedure can reliably detect price effects as small as 1.5 (1.0) percent. As a whole, the simulation results indicate the procedure can detect economically significant agent-owned price effects. More importantly, the results demonstrate that the null results we report are not driven by a finite sample approximation error.

4.2.4 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). We show that the bias is not resolved when a repeat-sales methodology is employed. Although repeat-sales estimators control for unobserved time-invariant attributes, unobserved time-varying attributes can still bias the estimates. Differencing Equation 1 gives us:

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

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, $E[\Delta d_{nt}\Delta\psi_{nt}] = 0$.[27] 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, 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 agent-owned estimates 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

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

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

4.2.5 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. The estimates for Atlanta are statistically insignificant across the entire study period regardless of the functional form and tokens employed. The TOM estimates are statistically 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 client-owned houses.[28]

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" their house. 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.[29]

[28] Similar results are found when log(TOM) is the dependent variable. We examine the co-determination of sales price and time-on-market in an internet appendix.

[29] 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.[30] 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 38% from 8.6% to 5.3% using zip code controls. The results in Phoenix mimic Atlanta. The vacant discount estimate drops 46% from 4.8% to

[30] When estimating agent-owned premiums, Atlanta's entire study period is 2007 to 2016 since 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 since the vacant variable is populated and available for the entire study period.

2.6%. The results in Table 9 highlight the effect the real estate market cycle has on vacancy discounts. The magnitude of the vacant coefficient estimate 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 coefficient estimate, thereby demonstrating our approach's ability to address a portion of the model-omitted variable bias inherent in the naive estimates.

5 Conclusion

Although researchers often have access to unstructured text, its high-dimensional nature precludes the use of conventional econometric techniques. For this reason, text is frequently ignored, resulting in a model-omitted variable bias. A model-omitted variable bias occurs when a correlated variable is available in the data, but not included in the model. In other words, the bias is self-imposed and can be easily addressed by including the variable.[31] This, of course, assumes the variable is quantifiable and available in a well-structured format. We provide a framework for using unstructured text as data. The framework we employ differs from previous textual analysis applications in that it (i) does not require a pre-specified dictionary to map text into usable predictors, (ii) includes the usable predictors (and not their tone or sentiment) in the pricing model, and (iii) uses a double-selection LASSO procedure to address model-omitted variable biases. Our framework is incredibly flexible and can be used to incorporate textual information in models beyond the real estate application highlighted in this study.[32]

This is the first study, to the best of our knowledge, to use the double-selection LASSO procedure for variable selection in high-dimensional unstructured data. To demonstrate the efficacy of our framework, we revisit the seminal studies by Rutherford et al. (2005), Levitt

[31] A data-omitted variable bias occurs when a variable that is correlated with the parameter of interest is not available in the data. In which case, the researcher has incomplete information and cannot easily address the bias. See Section 2.1 for further discussion.

[32] For example, the framework can be applied in financial research that uses 10-K filings, conference-call records, or FOMC-meeting minutes.

and Syverson (2008), and Kurlat and Stroebel (2015) that use agent-owned transactions as an indirect information variable (i.e. a proxy for asymmetric information). Using MLS data from Atlanta, Georgia and Phoenix, Arizona we replicate the naive agent-owned premiums in the extant literature of 3 to 4 percent. However, after we incorporate textual information from the remarks section of the MLS, the agent-owned premium shrinks by more than half in Phoenix and is no longer statistically significant in Atlanta. Using a restricted subsample we also find that the agent-owned coefficient estimate is statistically insignificant in Phoenix during the pre-boom (2000-2003) and bust (2007-2009) subperiods. The results are robust to several model specifications and are not the result of overfitting. Additionally, we show that our data-driven textual analysis framework performs well in out-of-sample tests.

These findings demonstrate that (i) the naive estimates in the extant literature suffer from a model-omitted variable bias and (ii) the bias can be mitigated by including textual information in the empirical model in order to more closely align the information set in the model with that of the market participants involved in the transaction. The results also suggest 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 informational advantage to their clients' detriment (i.e. an agency problem does not exist).

We also show the repeat-sales estimator approach does not resolve the model-omitted variable bias. The repeat-sales approach assumes 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 controls for the time-varying condition of the house and neighborhood, allowing us to estimate the pricing differential for agent-owned properties conditional on changes in the relevant time-varying attributes of the property. These results highlight the value of including textual information when heterogenous assets vary along multiple dimensions that are difficult to quantify.

References

Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., and Evanoff, D. D. (2011). The role of securitization in mortgage renegotiation. *Journal of Financial Economics*, 102(3):559–578.

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.

Buhlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science &amp; Business Media.

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.

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3):712–729.

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 MLS listings

Sample Listing 1

Tract	Beds	Baths	Sqft	Date	Price	$\hat{u}$	$\hat{u}_2$
04013107000	3	2	1,484	10/15/2002	$124,000	0.141	0.012

MLS Remarks: hallcraft red brick beauty!! in all the years i have been working this neighborhood this is the nicest home i have ever seen! home is beautifully decorated. custom kitchen is magnificent! panelled family room w/fireplace. wonderful back yard is tiled with beautiful in ground pool & spa! large windows in kitchen and front room for light, cheery atmosphere! lots of extras!

Sample Listing 2

Tract	Beds	Baths	Sqft	Date	Price	$\hat{u}$	$\hat{u}_2$
04013107000	3	2	1,480	10/28/2002	$88,000	-0.175	-0.037

MLS Remarks: super bargain price on this great family home. needs cosmetic fix-up and cleaning and the pool needs to be replastered. great opportunity for someone who can do a little work. easy to show - vacant and on lockbox.

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.

Table 2: Descriptive statistics

Panel A: Atlanta (2007-2016)						
	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Price (000s)	50.00	137.00	225.89	193.45	284.00	2,600.00
Sfla (000s)	0.57	1.64	2.21	2.12	2.67	6.00
Age	2	14	29.86	24	42	196
Bedrooms	1	3	3.63	4	4	6
Bathrooms	1.00	2.00	2.49	2.50	3.00	3.50
Agent-owned	0	0	0.02	0	0	1
Rental	0	0	0.03	0	0	1
Vacant	0	0	0.35	0	1	1

Panel B: Phoenix (2000-2013)						
	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Price (000s)	50.00	142.50	247.56	204.00	293.00	3,000.00
Sfla (000s)	0.50	1.44	1.92	1.77	2.24	6.00
Age	2	7	19.85	16	29	122
Bedrooms	1	3	3.28	3	4	6
Bathrooms	1.00	2.00	2.16	2.00	2.50	3.50
Agent-owned	0	0	0.06	0	0	1
Rental	0	0	0.03	0	0	1
Vacant	0	0	0.37	0	1	1

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, Agent-owned	-0.105	1				
zip, Price	0.840	-0.088	1			
zip, Agent-owned	-0.106	0.997	-0.088	1		
tract, Both	0.767	0.526	0.643	0.524	1	
zip, Both	0.619	0.550	0.756	0.552	0.885	1
Non-Zero	421	223	383	224	624	586

Panel B: Phoenix (2000-2013)

	tract, Price	tract, Agent-owned	zip, Price	zip, Agent-owned	tract, Both	zip, Both
tract, Price	1					
tract, Agent-owned	0.113	1				
zip, Price	0.83	0.119	1			
zip, Agent-owned	0.102	0.975	0.113	1		
tract, Both	0.922	0.389	0.77	0.374	1	
zip, Both	0.764	0.39	0.918	0.397	0.847	1
Non-Zero	548	126	543	130	615	613

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) 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^{**}	$0.015^{\dagger}$	0.020^{*}	0.033^{***}	0.017	0.021^{*}
	(0.013)	(0.010)	(0.011)	(0.012)	(0.010)	(0.011)
Vacant	-0.120^{***}	$-0.074^{***\dagger}$	$-0.099^{***\dagger}$	-0.102^{***}	$-0.065^{***\dagger}$	$-0.085^{***\dagger}$
	(0.016)	(0.008)	(0.012)	(0.014)	(0.007)	(0.010)
Vacant x Agent-owned	0.054^{*}	$0.010^{\dagger}$	0.014	0.062^{***}	$0.016^{\dagger}$	$0.024^{\dagger}$
	(0.032)	(0.019)	(0.028)	(0.023)	(0.013)	(0.020)
Rental	-0.038^{**}	$-0.017^{*\dagger}$	-0.049^{***}	-0.013	-0.004	-0.026^{**}
	(0.018)	(0.009)	(0.014)	(0.014)	(0.008)	(0.011)
Rental x Agent-owned	0.019	-0.004	0.013	0.027	0.003	0.022
	(0.027)	(0.021)	(0.026)	(0.023)	(0.017)	(0.021)
N	105,223	105,223	105,223	105,223	105,223	105,223
P	1,110	1,696	1,161	5,639	6,263	5,690
R^2	0.809	0.857	0.826	0.862	0.895	0.876
Panel B: Phoenix (2000-2013)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.041^{***}	$0.017^{***\dagger}$	$0.032^{***\dagger}$	0.036^{***}	$0.015^{***\dagger}$	$0.027^{***\dagger}$
	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
Vacant	-0.048^{***}	$-0.026^{***\dagger}$	-0.040^{***}	-0.041^{***}	$-0.024^{***\dagger}$	-0.035^{***}
	(0.006)	(0.004)	(0.006)	(0.004)	(0.003)	(0.004)
Vacant x Agent-owned	-0.026^{***}	$-0.016^{***\dagger}$	-0.030^{***}	-0.013^{***}	-0.009^{***}	-0.018^{***}
	(0.007)	(0.005)	(0.008)	(0.004)	(0.003)	(0.004)
Rental	-0.067^{***}	$-0.029^{***\dagger}$	-0.055^{***}	-0.061^{***}	$-0.029^{***\dagger}$	$-0.050^{***\dagger}$
	(0.009)	(0.004)	(0.008)	(0.007)	(0.003)	(0.006)
Rental x Agent-owned	-0.029^{**}	$-0.004^{\dagger}$	-0.023^{*}	-0.024^{***}	$-0.006^{\dagger}$	-0.021^{**}
	(0.013)	(0.009)	(0.012)	(0.009)	(0.007)	(0.009)
N	274,824	274,824	274,824	274,824	274,824	274,824
P	1,838	2,451	1,890	11,570	12,185	11,622
R^2	0.901	0.928	0.908	0.929	0.946	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 $\hat{S}_2$ dictionaries that differ only in the use of zip or tract fixed effects during the variable selection process. Columns 3 and 6 use the dictionary described in Levitt and Syverson (2008). N represents the number of observations and P represents the number of variables (controls + fixed effects + tokens) included in the specification. † denotes coefficient estimates in columns 2 and 3 (5 and 6) that are statistically different than the baseline estimate in column 1 (4) at the 10 percent level. Standard errors clustered at the time, location, and agent level are reported in brackets.

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

Panel A: Atlanta						
	(1)	(2)	(3)	(4)	(5)	(6)
Entire	0.029^{**}	0.017	0.022^{*}	0.030^{**}	0.019	0.018
Bust	0.021^{**}	0.009	0.014	0.022^{**}	0.010	0.011
Recovery	0.033^{***}	0.020	0.026^{*}	0.034^{***}	0.023	0.021
Panel B: Phoenix						
	(1)	(2)	(3)	(4)	(5)	(6)
Entire	0.036^{***}	$0.015^{***\dagger}$	$0.028^{***\dagger}$	0.036^{***}	$0.020^{***\dagger}$	$0.019^{***\dagger}$
Pre-boom	0.025^{***}	$0.009^{**\dagger}$	0.020^{***}	0.024^{***}	$0.013^{***\dagger}$	$0.012^{**\dagger}$
Boom	0.037^{***}	$0.015^{***\dagger}$	$0.028^{***\dagger}$	0.037^{***}	$0.020^{***\dagger}$	$0.019^{***\dagger}$
Bust	0.045^{***}	$0.019^{***\dagger}$	$0.033^{***\dagger}$	0.043^{***}	$0.023^{***\dagger}$	$0.021^{***\dagger}$
Recovery	0.049^{***}	$0.026^{***\dagger}$	0.039^{***}	0.049^{***}	$0.030^{***\dagger}$	$0.031^{***\dagger}$
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 tokens in the $\hat{S}_2^{zip}$ dictionary. 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 dictionary and column 6 uses the flex-gram dictionary. Agent-owned estimates are provided for the entire study period and several market-specific subperiods. † denotes coefficient estimates in columns 2 to 6 that are statistically different than the baseline estimate in column 1 at the 10 percent level.

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	0.017	0.012	0.018	0.014
Bust	0.021^{**}	0.015	0.012	0.009	0.009	0.005	0.009	-0.005
Recovery	0.033^{***}	0.024	0.021	0.010	0.020	0.014	0.023	0.020
Panel B: Phoenix								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Entire	0.036^{***}	0.029^{***}	0.033^{***}	0.024^{***}	$0.015^{***\dagger}$	$0.010^{***\dagger}$	$0.017^{***\dagger}$	$0.010^{***\dagger}$
Pre-boom	0.025^{***}	0.019^{***}	0.024^{***}	0.016^{***}	$0.009^{**\dagger}$	$0.004^{*\dagger}$	$0.008^{**\dagger}$	$0.004^{\dagger}$
Boom	0.037^{***}	0.030^{***}	0.032^{***}	0.023^{***}	$0.015^{***\dagger}$	$0.009^{***\dagger}$	$0.014^{***\dagger}$	$0.008^{***\dagger}$
Bust	0.045^{***}	0.039^{***}	0.044^{***}	0.033^{***}	$0.019^{***\dagger}$	$0.014^{*\dagger}$	$0.023^{***\dagger}$	$0.016^{\dagger}$
Recovery	0.049^{***}	0.036^{***}	0.046^{***}	0.036^{**}	$0.026^{***\dagger}$	$0.016^{***\dagger}$	$0.030^{***\dagger}$	$0.024^{**\dagger}$
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 tokens in the $\hat{S}_2^{zip}$ dictionary. Columns 1 and 5 include the entire occupied housing transaction sample. 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. † denotes coefficient estimates in columns 5 to 8 that are statistically different than their corresponding estimate in columns 1 to 4 at the 10 percent level.

Table 7: Repeat-sales agent-owned estimates

	Atlanta (2007-2016)			Phoenix (2000-2013)		
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.062^{***}	$0.020^{\dagger}$	0.033	0.029^{***}	$0.013^{***\dagger}$	$0.018^{***\dagger}$
	(0.023)	(0.024)	(0.021)	(0.005)	(0.004)	(0.004)
N	13,198	13,198	13,198	58,759	58,759	58,759
P	7,394	7,978	7,445	30,318	30,931	30,370
R^2	0.970	0.978	0.973	0.982	0.985	0.983
House FE	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓	✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	L&S		$\hat{S}_2^{zip}$	L&S

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

Note: The repeat-sales estimates include a subsample of houses that sold at least twice during the sample 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. Columns 1 to 3 include transactions in Atlanta, GA from 2007 to 2016 and columns 4 to 6 includes transactions in Phoenix, AZ from 2000 to 2013. Every column includes house fixed effects and time by zip code fixed effects. Columns 1 and 4 do not include any tokens. Columns 2 and 5 include unigram tokens in the set $\hat{S}_2^{zip}$ and columns 3 and 6 use the dictionary described in Levitt and Syverson (2008). N represents the number of observations and P represents the number of variables (fixed effects + tokens) included in the specification. † denotes coefficient estimates in columns 2 and 3 (5 and 6) that are statistically different than the baseline estimate in column 1 (4) at the 10 percent level. Standard errors clustered at the time, location, and agent level are reported in brackets.

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

Panel A: Atlanta

	(1)	(2)	(3)	(4)	(5)	(6)
Entire	1.87	0.66	1.94	2.74	1.54	2.81
Bust	6.86	6.12	6.73	8.72	8.14	8.55
Recovery	−1.17	−2.38	−0.99	−0.78	−2.14	−0.61

Panel B: Phoenix

	(1)	(2)	(3)	(4)	(5)	(6)
Entire	2.11	1.72	1.93	2.02	1.75	1.81
Pre-boom	0.69	0.09	0.62	0.33	−0.07	0.24
Boom	−1.36	−2.02	−1.55	−1.48	−2.12	−1.68
Bust	9.59*	9.82**	9.39*	10.24	10.89*	10.03*
Recovery	5.00	4.86	4.41	5.49	5.57	4.92
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: 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 tokens in the $\hat{S}_2$ dictionary. Column 3 and 6 use the dictionary described in Levitt and Syverson (2008). The agent-owned TOM estimates are provided for the entire study period and several subperiods. † denotes coefficient estimates in columns 2 and 3 (5 and 6) that are statistically different than the baseline estimate in column 1 (4) at the 10 percent level.

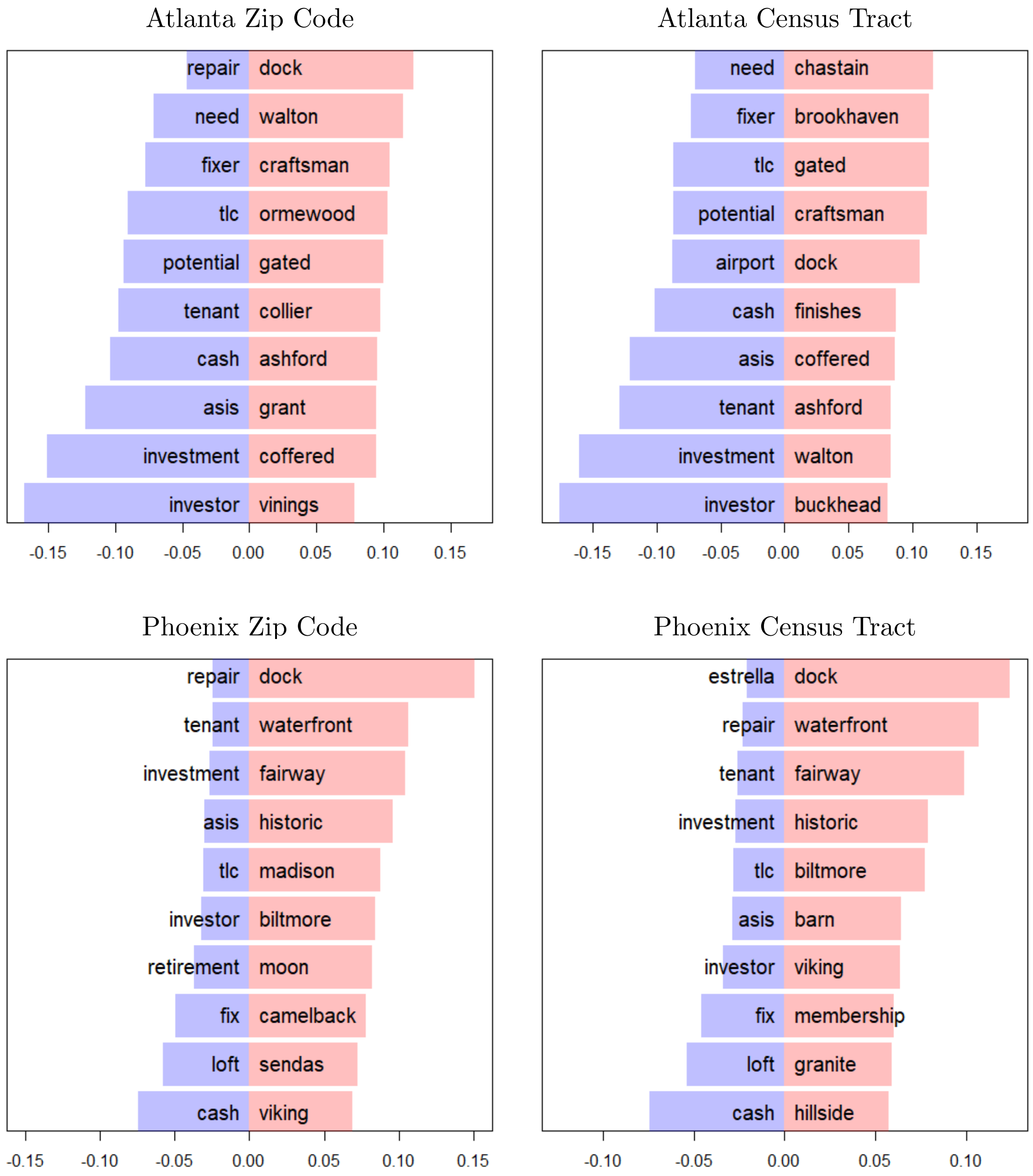
Table 9: Vacant discount estimates

	Atlanta				Phoenix			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Entire	-0.086^{***}	$-0.053^{***\dagger}$	-0.074^{***}	$-0.047^{***\dagger}$	-0.048^{***}	$-0.026^{***\dagger}$	-0.042^{***}	$-0.023^{***\dagger}$
Pre-boom	-0.033^{***}	$-0.017^{***\dagger}$	-0.032^{***}	$-0.017^{***\dagger}$	-0.032^{***}	$-0.015^{***\dagger}$	-0.032^{***}	$-0.016^{***\dagger}$
Boom	-0.055^{***}	$-0.029^{***\dagger}$	-0.049^{***}	$-0.028^{***\dagger}$	-0.032^{***}	$-0.016^{***\dagger}$	-0.030^{***}	$-0.015^{***\dagger}$
Bust	-0.153^{***}	$-0.084^{***\dagger}$	-0.131^{***}	$-0.076^{***\dagger}$	-0.072^{***}	$-0.039^{***\dagger}$	-0.063^{***}	$-0.035^{***\dagger}$
Recovery	-0.106^{***}	$-0.071^{***\dagger}$	-0.090^{***}	$-0.062^{***\dagger}$	-0.075^{***}	$-0.039^{***\dagger}$	-0.059^{***}	$-0.034^{***\dagger}$
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 $\hat{S}_2$ dictionary 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 an internet appendix. † denotes coefficient estimates in even columns that are statistically different than their corresponding estimate in odd columns at the 10 percent level.

Figure 1: Implicit prices for unigram tokens



Atlanta Zip Code

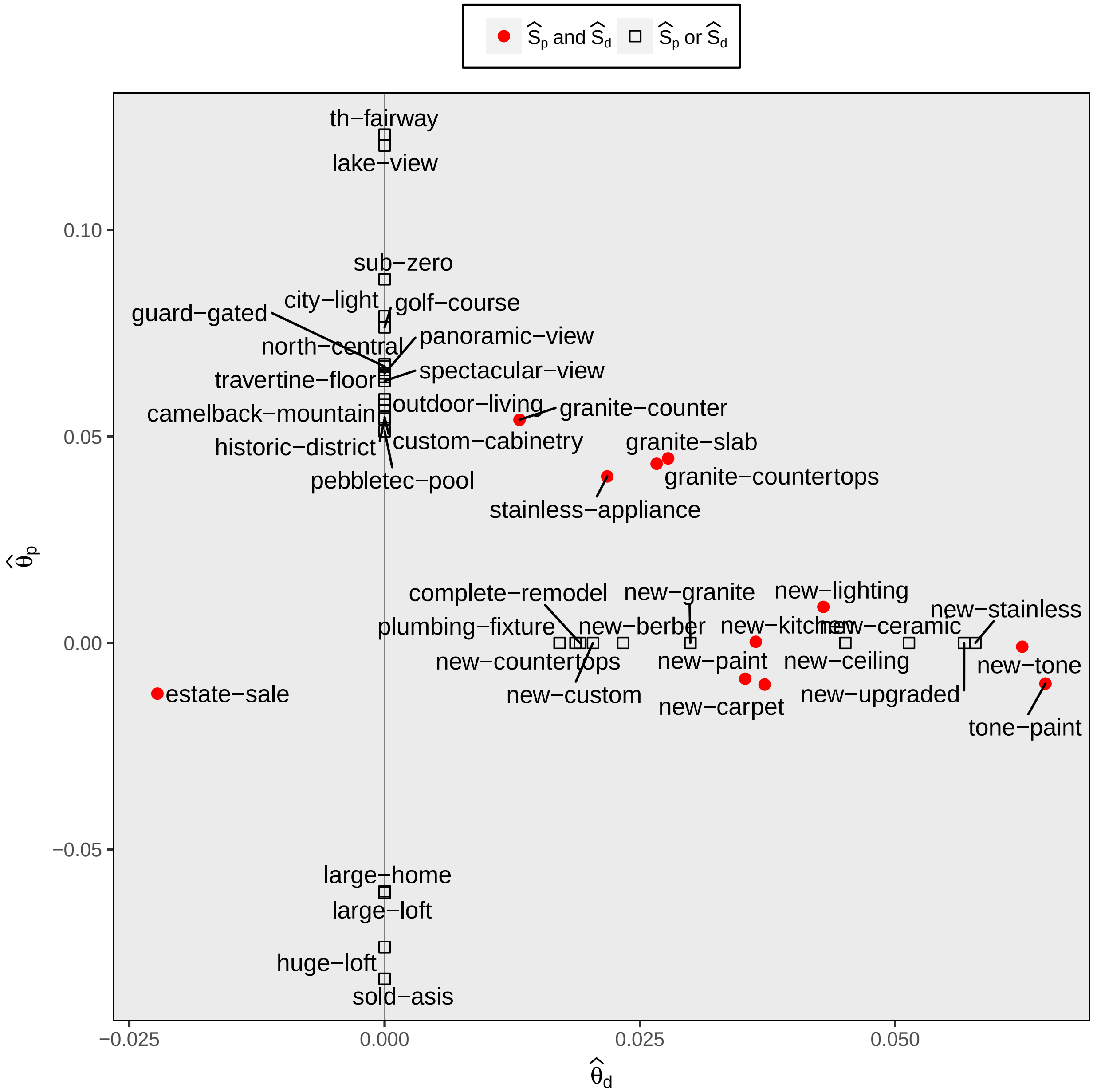
Atlanta Census Tract

Phoenix Zip Code

Phoenix Census Tract

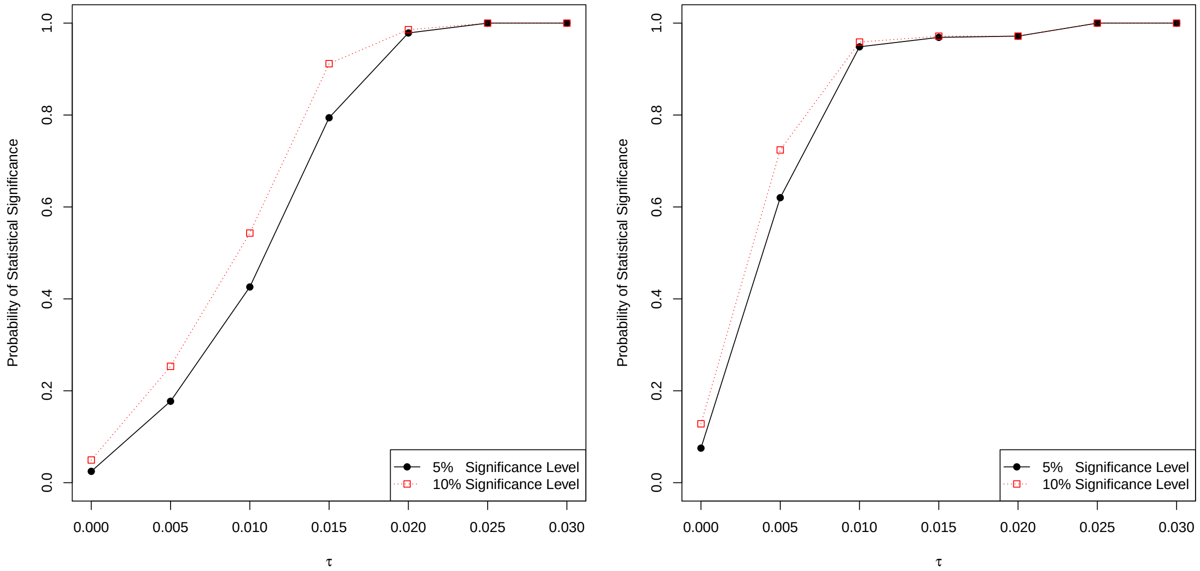
Note: The top ten positive and negative unigram tokens are displayed for Atlanta and Phoenix. The tokens with the largest magnitudes were selected from the market-specific zip code and census tract dictionaries.

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.

Figure 3: Post double-selection in finite samples



(a) Atlanta: 2007-2016

(b) Phoenix: 2000-2003

Note: Figures 3a and 3b display simulation results for the finite sample performance of the post double-selection estimator. Each figure displays the fraction of simulations where $H_0 : \tau = 0$ is rejected at either a 5% or 10% significance level. Details of the simulation are provided in the appendix.

Appendices

Contents

A Data Overview

A.1 Data filters

We drop single-family detached 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 ≤ 6
4. $1 \leq$ bathrooms ≤ 3.5
5. age ≥ 2
6. acres ≤ 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)	29.845	-0.001	-0.001
Rental	0.031	-0.048	-0.020
Vacant	0.350	-0.123	-0.077
Acres: .5 - 1	0.108	0.024	0.028
Acres: 1 - 2	0.031	0.103	0.095
Acres: 2 - 5	0.010	0.227	0.228
Baths: 1.5	0.021	0.067	0.051
Baths: 2	0.286	0.245	0.192
Baths: 2.5	0.361	0.302	0.235
Baths: 3	0.167	0.326	0.264
Baths: 3.5	0.127	0.450	0.358
Beds: 2	0.034	-0.049	-0.050
Beds: 3	0.424	-0.057	-0.027
Beds: 4	0.429	-0.017	0.016
Beds: 5	0.105	0.000	0.036
Beds: 6	0.008	-0.056	0.005
Sfla: 500 - 1000	0.015	-0.081	-0.071
Sfla: 1500 - 2000	0.263	0.135	0.115
Sfla: 2000 - 2500	0.242	0.272	0.229
Sfla: 2500 - 3000	0.177	0.410	0.340
Sfla: 3000 - 3500	0.088	0.510	0.424
Sfla: 3500 - 4000	0.035	0.597	0.501
Sfla: 4000 - 4500	0.013	0.655	0.557
Sfla: 4500 - 5000	0.004	0.732	0.614
Sfla: 5000 - 5500	0.002	0.814	0.703
Sfla: 5500 - 6000	0.000	0.902	0.807

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	Basic	Tokens
	(1)	(2)	(3)
Age (Years)	9.928	-0.006	-0.007
Rental	0.027	-0.092	-0.038
Vacant	0.374	-0.060	-0.032
Acres: .5 - 1	0.020	0.242	0.196
Acres: 1 - 2	0.024	0.255	0.199
Acres: 2 - 5	0.005	0.439	0.361
Baths: 1.5	0.143	0.078	0.071
Baths: 2	0.531	0.102	0.088
Baths: 2.5	0.140	0.086	0.095
Baths: 3	0.114	0.112	0.112
Baths: 3.5	0.032	0.184	0.154
Beds: 2	0.131	0.144	0.143
Beds: 3	0.514	0.140	0.165
Beds: 4	0.300	0.115	0.161
Beds: 5	0.052	0.039	0.127
Beds: 6	0.003	-0.036	0.085
Sfla: 500 - 1000	0.023	-0.141	-0.117
Sfla: 1500 - 2000	0.346	0.189	0.152
Sfla: 2000 - 2500	0.198	0.423	0.337
Sfla: 2500 - 3000	0.090	0.657	0.519
Sfla: 3000 - 3500	0.049	0.813	0.650
Sfla: 3500 - 4000	0.020	0.941	0.762
Sfla: 4000 - 4500	0.009	1.037	0.843
Sfla: 4500 - 5000	0.002	1.117	0.924
Sfla: 5000 - 5500	0.001	1.221	1.012
Sfla: 5500 - 6000	0.000	1.280	1.064

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 Power simulations

Price parameters and residuals for simulation

This section describes the simulation experiments we use to calculate the power of the double-selection procedure in Belloni et al. (2014) for $\tau \in \{0.00, 0.05, 0.010, 0.015, 0.020, 0.025, 0.030\}$. We use $J = 500$ simulations for each choice of τ. Based on computational considerations, we run the simulations for periods in which we find null results (2007-2016 in Atlanta and 2000-2003 in Phoenix) and use $K = 2000$ candidate tokens.[33] For exposition purposes, we drop the t subscript.

In order to simulate price data, we estimate parameters of the price equation using the heteroskedastic LASSO

$$(\hat{\beta}_p^{*'}, \hat{\tau}_p^{*'}, \hat{\theta}_p^{*'})' = \arg\min_{\beta,\tau,\theta} \sum_n (p_n - x_n\beta - d_n\tau - w_n\theta)^2 + \lambda_p \sum_k |\theta_k \phi_{p,k}| \tag{10}$$

Because LASSO coefficient estimates are shrunk towards 0, we then calculate the post-LASSO (Belloni et al., 2013) estimates as the least-squares coefficients when regressing p_n on d_n and the $\hat{Q}^*$ variables in x_n and w_n corresponding to the non-zero elements in $\hat{\beta}_p^*$ and $\hat{\theta}_p^*$. Define the $\hat{\beta}_p^{PL}$ and $\hat{\theta}_p^{PL}$ as the post-LASSO estimates and x_n^{PL} and w_n^{PL} as the corresponding regressors. Define $\hat{\tau}^{PL}$ as the post-LASSO estimate of τ. Define $\hat{e}_n$ as the residual from the post-LASSO estimator. We calculate the predicted value of price excluding the agent-owned effect as $\hat{p}_n = x_n\hat{\beta}_p^{PL} + w_n\hat{\theta}_p^{PL}$.

Agent-owned parameters for simulation

In order to simulate the agent-owned variable, we estimate an ℓ_1 penalized logit model for d_n

$$(\hat{\beta}_d^{*'}, \hat{\theta}_d^{*'})' = \arg\max_{\beta,\theta} \prod_n \Lambda(x_n\beta, w_n\theta)^{d_n}(1 - \Lambda(x_n\beta, w_n\theta))^{1-d_n} + \lambda_d \sum_k |\theta_k| \tag{11}$$

In Equation 11, Λ is the logistic cdf, and the objective function is a penalized likelihood model. We use the λ_d that minimizes the 5-fold cross-validated likelihood. Simulation results are not sensitive to other choices of λ_d near λ_d and are available upon request. We then calculate the (unpenalized) maximum likelihood estimates, $\hat{\beta}_d^{MLE}$ and $\hat{\theta}_d^{MLE}$, for the logit model using the variables corresponding to the non-zero elements in $\hat{\beta}_d^{*'}$ and $\hat{\theta}_d^{*'}$. Based on the maximum likelihood estimates, we then calculate $\hat{\pi}_n = \Pr(d_n = 1 | x_n, w_n, \hat{\beta}_d^{MLE}, \hat{\theta}_d^{MLE})$.

[33] Each simulation requires approximately 15 minutes of run time.

Simulation

For each simulation $j = 1, ..., J$, we generate y_n^j and d_n^j as

1. Draw d_n^j as a Bernoulli random variable with $\Pr(d_n^j = 1) = \hat{\pi}_n$

2. Draw $\epsilon_n^j \sim \mathcal{N}(0, 1)$ and create $p_n^j = \hat{p}_n + d_n^j \tau + \epsilon_n^j \hat{e}_n$

The first step ensures d_n^j is a binary variable. The second step is similar to the wild bootstrap and allows for heteroskedasticity in the simulated errors. Note, $\hat{p}_n$ is the predicted price minus the least-squares estimate of τ in the first stage of the simulation. Based on the simulated data, we then estimate τ using the post double-selection procedure outlined in the paper.

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.*[34] 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*.

[34] 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 Duplicate remarks

An earlier version of this paper included MLS listings that had duplicate remarks. The current version of the paper removes remarks that show up in more than three listings since they (i) do not provide unique information about the house listed for sale and (ii) typically identify an atypical transaction. For example, the following remarks are used 108, 105, and 78 times each in the Phoenix data. Although the first and second remark state that the listing is "not a short-sale", the remark provides little to no information about the house listed for sale. Instead, the remarks state that the house is either "aggressively priced" or "priced to sell", which suggests that the seller is highly motivated and may be under some type of distress. Similarly, the third duplicate remark identifies corporate owned properties.

Duplicate Remark 1 (N = 108)
An acquired home, NOT a short sale. This home shows well. Aggressively priced and sold as-is. LSR with offer if a loan is needed or proof of funds if cash. Seller may pay up to 3% towards closing costs for owner occupant purchaser!
Duplicate Remark 2 (N = 105)
An acquired home, NOT a short sale. This home shows well. Priced to sell and sold as-is. Please have LSR with offer if a loan is needed or proof of funds if cash. Seller may contribute towards buyers closing costs for owner occupant purchases.
Duplicate Remark 3 (N = 78)
This property is corporate owned and sold AS-IS in current condition. No repairs, warranties, disclosures or inspection provided by the Seller. Your proof of funds or a prequalification letter must accompany all offers. A special addendum will be made part of the final contract. Seller reserves the right to negotiate offers in any order regardless of date/time submitted. Buyer to verify all facts and figures.

Removing the duplicate remarks has a large impact on the variable selection process. The double-selection process selects 1,162 tokens in $\widehat{S}_2^{tract}$ and 1,167 tokens in $\widehat{S}_2^{zip}$ when the duplicate remarks are included in the Phoenix dataset. In contrast, the double-selection process selects 615 tokens in $\widehat{S}_2^{tract}$ and 613 tokens in $\widehat{S}_2^{zip}$ when the duplicate remarks are filtered out (see Table 3). Removing the duplicate remarks does not, however, have a material impact on the agent-owned coefficient estimates since the estimates in Table B1, which include the duplicate remarks, are similar to Table 4, which does not include the duplicate remarks.

Table B1: Baseline agent-owned estimates by tenure with duplicate remarks

Panel A: Atlanta (2007-2016)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.033^{**}	$0.012^{\dagger}$	0.021	0.033^{**}	$0.015^{\dagger}$	0.021^{*}
	(0.015)	(0.012)	(0.013)	(0.013)	(0.011)	(0.012)
Vacant	-0.121^{***}	$-0.072^{***\dagger}$	$-0.099^{***\dagger}$	-0.103^{***}	$-0.063^{***\dagger}$	-0.085^{***}
	(0.017)	(0.008)	(0.013)	(0.014)	(0.007)	(0.011)
Vacant x Agent-owned	0.055^{*}	$0.015^{\dagger}$	0.013	0.063^{**}	$0.021^{\dagger}$	$0.023^{\dagger}$
	(0.033)	(0.020)	(0.030)	(0.025)	(0.014)	(0.021)
Rental	-0.038^{*}	$-0.017^{\dagger}$	-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
P	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^{***}	$0.017^{***\dagger}$	$0.031^{***\dagger}$	0.035^{***}	$0.015^{***\dagger}$	$0.027^{***\dagger}$
	(0.005)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)
Vacant	-0.048^{***}	$-0.025^{***\dagger}$	-0.040^{***}	-0.041^{***}	$-0.023^{***\dagger}$	-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^{***}	$-0.030^{***\dagger}$	-0.057^{***}	-0.063^{***}	$-0.029^{***\dagger}$	-0.052^{***}
	(0.011)	(0.005)	(0.010)	(0.008)	(0.004)	(0.007)
Rental x Agent-owned	-0.026^{**}	$-0.004^{\dagger}$	-0.020^{*}	-0.021^{**}	$-0.006^{\dagger}$	-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
P	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 dictionary described in Levitt and Syverson (2008). N represents the number of observations and P represents the number of variables (controls + fixed effects + tokens) included in the specification. † denotes coefficient estimates in columns 2 and 3 (5 and 6) that are statistically different than the baseline estimate in column 1 (4) at the 10 percent level. Standard errors clustered at the time and location level are reported in brackets.

B.6 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 B2 and B4 can be used to predict house prices in Atlanta and Phoenix, respectively. Whereas, the tokens selected by the linear probability equation in Tables B3 and B5 can be used to identify agent-owned transactions in Atlanta and Phoenix, respectively.

Table B2: Selected Tokens from the Hedonic Equation (Atlanta)

Unigrams		Bigrams		n-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
investor	-0.18	sold-asis	-0.26	investor	-0.20
investment	-0.16	tenant-occupied	-0.22	sold-asis	-0.19
tenant	-0.13	need-tlc	-0.18	investment-opportunity	-0.18
asis	-0.12	investor-special	-0.17	investment-property	-0.16
chastain	0.12	split-foyer	-0.16	need-tlc	-0.15
craftsman	0.11	split-level	-0.15	cash	-0.14
brookhaven	0.11	fixer-upper	-0.15	asis	-0.13
gated	0.11	great-investment	-0.14	split-level	-0.12
dock	0.11	investment-property	-0.14	split-level-home	-0.12
cash	-0.10	chastain-park	0.14	tenant	-0.12
potential	-0.09	pre-qual	-0.13	investment	-0.12
tlc	-0.09	investment-opportunity	-0.13	dock	0.12
finishes	0.09	lot-potential	-0.13	sold-asis-condition	-0.12
airport	-0.09	chef-kitchen	0.13	brookhaven	0.11
coffered	0.09	first-time	-0.12	craftsman	0.11
walton	0.08	coffered-ceiling	0.12	split-foyer	-0.11
buckhead	0.08	need-work	-0.12	tlc	-0.11
kirkwood	0.08	starter-home	-0.10	repair	-0.10
sandy	0.08	sandy-spring	0.10	airport	-0.10
ashford	0.08	best-street	0.10	coffered-ceiling	0.10

Table B3: Selected Tokens from the Linear Probability Equation (Atlanta)

Unigrams		Bigrams		n-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
realize	0.27	house-just	0.25	new-fixture	0.13
taking	0.07	just-renovated	0.23	entire	0.09
brand	0.03	entire-house	0.08	new-granite-countertops	0.06
entire	0.03	new-fixture	0.07	complete-renovation	0.06
highlight	0.03	complete-renovation	0.04	new-appliance	0.04
renovated	0.02	bathroom-home	0.03	new-stainless-steel-appliance	0.04
renovation	0.02	brand-new	0.02	highlight	0.03
fixture	0.02	new-appliance	0.02	friend	0.02
friend	0.02	new-flooring	0.02	renovated	0.01
advantage	0.02	renovated-new	0.02	new-paint	0.01
appliance	0.01	home-just	0.02	renovation	0.01
house	0.01	floor-entire	-0.02	beautiful-home	0.01
dream	0.01	supra-lock	-0.02	fireside-family-room	0.01
absolutely	-0.01	elegant-home	0.02	meticulously-maintained	-0.01
waiting	-0.01	new-paint	0.01	absolutely	0.01
growing	-0.01	new-granite	0.01	story-foyer	-0.01
send	-0.01	new-home	0.01	finished-terrace-level	-0.01
attorney	-0.01	lovely-home	-0.01	main-level	-0.01
approval	-0.01	home-well	-0.01	beautifully-maintained	-0.01
loved	-0.01	open-kitchen	-0.01	family-room	-0.01

Table B4: Selected Tokens from the Hedonic Equation (Phoenix)

Unigrams		Bigrams		n-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
dock	0.12	th-fairway	0.12	th-fairway	0.12
waterfront	0.11	lake-view	0.12	biltmore	0.09
fairway	0.10	sub-zero	0.09	fairway	0.08
historic	0.08	golf-course	0.08	fix	-0.07
biltmore	0.08	sold-asis	-0.08	golf-course-lot	0.07
cash	-0.07	city-light	0.08	barn	0.07
granite	0.06	guard-gated	0.07	golf-course	0.06
preserve	0.06	north-central	0.07	city-light	0.06
camelback	0.06	huge-loft	-0.07	travertine-floor	0.06
membership	0.06	panoramic-view	0.07	finishes	0.06
hillside	0.06	travertine-floor	0.06	historic	0.06
barn	0.06	large-loft	-0.06	large-loft	-0.06
viking	0.06	outdoor-living	0.06	golf-course-view	0.06
pool	0.05	spectacular-view	0.06	granite-counter	0.05
loft	-0.05	large-home	-0.06	loft	-0.05
wine	0.05	camelback-mountain	0.06	granite	0.05
fix	-0.05	granite-counter	0.05	travertine	0.05
finishes	0.05	heated-pool	0.05	pebble-tec-pool	0.05
subzero	0.05	custom-home	0.05	salt-water-pool	0.05
view	0.04	hardwood-floor	0.05	arcadia	0.05

Table B5: Phoenix - Tokens Selected by Linear Probability Model

Unigrams		Bigrams		Flex-grams	
Token	$\hat{\theta}$	Token	$\hat{\theta}$	Token	$\hat{\theta}$
tone	0.05	tone-paint	0.06	new-stainless-steel-appliance	0.07
remodel	0.04	new-stainless	0.06	remodel	0.05
granite	0.03	new-upgraded	0.06	arizona	0.05
hardware	0.03	new-tone	0.06	new-ceiling-fan	0.05
rubbed	0.03	new-ceiling	0.05	new-carpet	0.04
new	0.02	new-ceramic	0.05	tone-paint	0.04
paint	0.02	new-carpet	0.04	hardware	0.04
stainless	0.02	new-paint	0.04	brand-new	0.03
fixture	0.02	new-kitchen	0.04	new-paint	0.03
travertine	0.02	new-lighting	0.04	faucet	0.03
pride	-0.02	granite-countertops	0.03	new-kitchen	0.03
faucet	0.02	granite-slab	0.03	new-paint-inside	0.03
state	0.02	new-granite	0.03	new	0.02
carpet	0.01	pride-ownership	-0.02	remodeled	0.02
newer	-0.01	well-cared	-0.02	granite-countertops	0.02
neutral	-0.01	stainless-appliance	0.02	granite	0.02
call	-0.01	leave-message	-0.02	fixture	0.02
breakfast	-0.01	garage-floor	0.02	travertine	0.02
remodeled	0.01	granite-kitchen	0.02	stainless-appliance	0.02
maintained	-0.01	plumbing-fixture	0.02	new-interior	0.02

B.7 Top ten unigram token descriptions

Figure 1 presents the top ten positive and negative unigram tokens for Atlanta and Phoenix. The bulk of the tokens are self-explanatory. However, some tokens require additional explanation. In this section we describe each token and provide an example of the token in the remarks. Tokens that are listed multiple times in Figure 1 are only listed once below.

B.7.1 Atlanta negative tokens

Token: **investor** (N=1,902)

Description: The investor token identifies two types of properties: rentals and/or properties that need work (i.e. flips).

Example: Great opportunity for the investor who just wants minimal work. Cute home has great potential for the owner occupant who isn't afraid to take on a project. Very attractive street/area. Conveniently located near Clayton State University.

Token: **investment** (N=1,092)

Description: The investment token identifies two types of properties: rentals and/or properties that need work (i.e. flips).

Example: Back on the market at a severely reduced price. 3 bedroom 2 bath investment property in East Cobb. Great rehab project or investment property. Flip, rent or seek a variance to increase the value.

Token: **asis** (N=3,458)

Description: The as-is token signals that the seller is not willing to pay to fix/repair the house, so buyers must negotiate accordingly (i.e. factor required repairs into their offer.)

Example: This is in a lovely older established neighborhood. The home has huge rooms and is very spacious, cozy and inviting. With the right vision this home would be a wonderful place to call home. Sold as-is. Equal Housing Opportunity. UI(Uninsured).

Token: **cash** (N=328)

Description: The cash token is included in the remarks for two distinct purposes. In the example below the cash token identifies a seller who will only take cash offers. In other listings, the cash token identifies rental properties that generate cash flows for their owner.

Example: Large four sided brick ranch. Separate apartment in lower level of home with full kitchen, bedroom, bathroom and rec room. Hardwood floors on main level of home. This is a Great Deal at this low price. Cash offers only.

Token: **tenant** (N=615)

Description: The tenant token identifies rental properties.

Example: Do not disturb tenants! Great tenant in place currently paying $820 per month. Lease expires May 2016. This 4 bed / 2 bath home sits in a quiet neighborhood. Tenants will love the spacious fenced in yard outside and the open floor plan inside.

Token: **potential** (N=1,089)

Description: The potential token identifies properties that need work.

Example: Great East Atlanta opportunity to renovate current home to easily reconfigure into a 3 bed 2 bath Endless potential, perfect location on great lot with private backyard, and so much more. Corporate owned. Closing attorney to he [name redacted].

Token: **tlc** (N=876)

Description: The tlc token is a common abbreviation of tender, love and care. In other words, the house needs some work.
Example: All the right schools you want to own a home! This house is price to sell. 4 bedroom with an addtional room that could serve as a bonus room or even a 5th bedroom. Home needs a little TLC but price reflects some of the minor repairs that are needed.

Token: **fixer** (N=272)

Description: The fixer token identifies houses that need work (i.e. fixer upper).
Example: Fixer upper in East Cobb's BEST school district, Sope Creek/Dickerson/Walton! Multi-level home would be perfect to flip or rent. Granite and stainless kitchen. Home backs up to the Chattahoochee Nation Forest that gives you privacy and outdoor opportunities.

Token: **need** (N=2,350)

Description: The need token identifies houses that need work.
Example: Great investor opportunity in Smyrna. Close to new stadium! Ranch in Highlands Subdivision sits on level lot. 3 BR, 1 bath, LR, Bonus Room. Laundry on main. Covered patio. Hardwood floors. Home needs rehab. Cash only. Sold as is with only a lead based pain

Token: **repair** (N=493)

Description: The repair token identifies houses that need work.
Example: Spacious 3 bedroom home with lots of room. Sep dining & living room. Bonus room. Kitchen with breakfast area, bar, island, solid counters & pantry. Powder room on main level. Full unfinished basement. 2 Car garage. Will need your attention to repairs.

Token: **airport** (N=454)

Description: The airport token identifies houses located near the airport. Although the location may be convenient for traveling, the noise from airplanes taking off and landing is often considered a disamenity.
Example: Newly renovated 3 bedroom 2 Bath gem located minutes from the airport. Perfect for a first time home buyer or a Buy & Hold investor looking to invest in a rental unit with no work required. This property has been totally renovated including upgraded kitchen.

B.7.2 Atlanta positive tokens

Token: **dock** (N=173)

Description: The dock token identifies houses that are located on or nearby a lake.
Example: Very nice all brick home with a beautiful lake view. Yard is landscaped, open lake access views and a big dock on 22 acre lake! Master Bedroom is large & the master bath has a lovely seperate tub & shower with a large bay window where the garden tub is.

Token: **walton** (N=672)

Description: The walton token identifies a high performing school zone in Atlanta.
Example: East Cobb & Walton Schools in Bridgegate w/Quick Access to Roswell Rd! 2-Story Foyer w/Hardwood & Crown/Chair Molding. Family Rm w/Hardwood, Gas Fireplace & Open to Sunroom. Eat-in Kitchen w/Granite, Stainless Steel Appliances, Island, Tile & French Doors.

Token: **craftsman** (N=1,541)

Description: The craftsman token identifies a popular architectural style.
Example: 2 story craftsman in thriving East Atlanta, Brand new exterior paint. Hardwood floor throughout, separate dining, tons of natural light. Features include open floor family room w/ fireplace. Kitchen w/ granite countertops, custom cabinetry, and stainless steel appliances.

Token: **ormewood** (N=199)

Description: The ormewood token identifies a popular neighborhood in Atlanta.
Example: You don't want to miss this light-filled Classic Ormewood Park Bunglow loaded with charm, great curb appeal and an awesome corner lot! You will love the sun-filled master suite complete with walk-in closet and nice size master bath. Fenced yard and covered porch.

Token: **gated** (N=545)

Description: The gated token identifies houses located in gated communities.
Example: UPGRADES GALORE! This home has it all... Gated community, best lot in subdivison on cul-de-sac. Wood cabinets, stone countertops with top of the line stainless appliances. Wired for surround sound. Master suite with trey ceilings and spa like bath.

Token: **collier** (N=87)

Description: The collier token identifies a popular neighborhood in Atlanta.
Example: Adorable renovated Collier Hills cottage! Completely gutted and renovated kitchen opened up to the dining room in 2013. Incredible backyard fully fenced. Renovated baths and great master suite with walk in closets and double vanity w/tons of storage.

Token: **ashford** (N=287)

Description: The ashford token identifies a popular neighborhood in Atlanta.
Example: Charming bungalow situated on large private lot in sought after Ashford Park. Rocking chair front porch! Gleaming hardwoods throughout the home. Master Suite with newly renovated high end bath and two closets. Updated kitchen with Dacor gas range, and breakfast area.

Token: **grant** (N=515)

Description: The grant token identifies a popular neighborhood in Atlanta.
Example: Stunning Grant Park, 2 bedroom/1 bath bungalow, new windows, HVAC and water heater only 4 years old, house was added onto and most of the interior was renovated in 2012, open kitchen featuring cherry kitchen cabinets, stainless appliances, and hardwood floors.

Token: **coffered** (N=434)

Description: The coffered token identifies houses that have coffered ceilings.
Example: Better than new This home nestled in nature is one of the few with the private backyard overlooking wooded natural area. Large entry foyer flanked by formal dining room w/ coffered ceilings. Chef's gourmet kitchen w/granite counter tops, and SS appliances.

Token: **vinings** (N=304)

Description: The vinings token identifies a popular neighborhood in Atlanta.
Example: Welcome to Sophisticated Luxury in Vinings Estates! Gorgeous Brick home in quiet cul-de-sac. Boasts generous open spaces. Entertain in your own Backyard Oasis. Two story great room with wall of windows offers privacy among the green tree tops.

Token: **chastain** (N=183)

Description: The chastain token identifies a popular neighborhood in Atlanta.
Example: Beautiful traditional 2 story brick home just steps away from Chastain Park Located on desirable Tall Pines Drive adjacent to Tall Pines Ct. 9ft ceilings and hardwood floors. Cook's kitchen with granite/stainless opens to charming vaulted breakfast room.

Token: **brookhaven** (N=716)

Description: The brookhaven token identifies a popular neighborhood in Atlanta.
Example: Adorable cottage in hot Brookhaven at super price. Home sits on oversized lot (huge side yard) plus backyard is large and fenced. New tankless water heater and recent HVAC system. New up tub/shower combo in second bedroom. Great Montgomery Elementary School.

Token: **finishes** (N=554)

Description: The finishes token identifies houses that have high-end features.
Example: Striking Craftsman design has curb-appeal and pristine interior! Iron rail fenced front lawn & inviting front veranda. High-end finishes throughout. Hardwoods, crown molding, 9ft ceilings & open concept living. Fireside family room opens to formal dining room.

Token: **buckhead** (N=498)

Description: The buckhead token identifies a popular neighborhood in Atlanta.
Example: Charming 3 bedroom 2 bath home in the heart of Buckhead. Spacious formal living room with fireplace, views to dining area open to kitchen. Kitchen features granite counters and stainless steel appliances. Sunroom and bright den with French doors to deck.

B.7.3 Phoenix negative tokens

Token: **loft** (N=12,866)

Description: The loft token identifies houses that have lofts.

Example: Reduced Reduced Motivated seller Beautiful open floor plan is perfect for entertaining Custom permitted addition has added 960 sqft to this home to make it functional and flow flawlessly The second floor has a cozy loft and 3 good size bedrooms.

Token: **retirement** (N=1,265)

Description: The retirement token identifies houses in communities with age restrictions.

Example: Located in the awesome 55 plus Sun Air Estates Retirement Community. Super Recreation Center with Clubhouse, Pool, Tennis Court, Shuffleboard, etc. Gorgeous 11 high coffered-ceiling in 8' circular formal-entry; split-bedroom floor plan for privacy; large Great Room; super Kitchen with Kitchen Island, Refrigerator, Microwave & Pantry.

Token: **repair** (N=2,904)

Description: The repair token identifies houses that need work.

Example: Regular sale in historic Rancho Ventura neighborhood. This is a Mid-Century Modern Haver Home designed by architect Ralph Haver. Large diving pool, Pebble-Tec surface in good shape. Home needs some repairs, ideal for homeowner with 203(k) financing or an investor. Do NOT submit any offer prior to viewing the home.

Token: **money** (N=1,887)

Description: The money token identifies sellers who are motivated and want cash buyers.

Example: INVESTOR SPECIAL! This 4 bedroom, 2 bath home located in Scottsdale has already been cleaned out and is ready for someone to put the finishing touches on it. Looking for Cash/ Hard money and a quick close.

B.7.4 Phoenix positive tokens

Token: **waterfront** (N=989)

Description: Although Phoenix has a desert landscape, there are several lakes with waterfront houses. The waterfront token identifies these houses.
Example: Great waterfront lot in Crystal Gardens, 3 Bedroom with Bonus Room, split floor plan, spacious kitchen, large master bathroom with walk-in closet, wood shutters, covered patio, lake access gate, garden area, concrete side yard walkway, side garage door access, close to freeways, shopping, restaurants, entertainment and so much more

Token: **fairway** (N=2,359)

Description: The fairway token identifies houses located on golf courses.
Example: A beautiful Serenitas on the 14th fairway of the Trilogy golf course.This gorgeous home boasts over $80,000 in upgrades including granite counter tops and island in kitchen, crown moulding ,plantation shutters, custom paint, and custom security doors. House repainted fall of 2012. Full Membership/Access to Kiva Club included in HOA.

Token: **historic** (N=2,128)

Description: The historic token identifies older houses that are (typically) well maintained.
Example: Major remodeling & expansion with modern amenities while retaining it's classic 1939 architecture & style. Main House contains 1,585 SF with 3 bedrooms & 2 baths. Remodeling includes A/C unit, kitchen cabinets, granite counters, stainless steel appliances, baths, copper plumbing, recessed lighting, skylights, etc. Your clients will love the historic charm complete with today's modern amenities.

Token: **madison** (N=864)

Description: The madison token identifies a high performing school zone in Phoenix.
Example: Four bedroom home in highly sought after Madison School District. New A/C, evap cooler and flooring all with in the last 2 years. Irrigated lot with mature landscaping and lovely grass. HIGHLY SOUGHT AFTER central Phoenix area close to schools, shopping, freeway and downtown.

Token: **biltmore** (N=1,111)

Description: The biltmore token identifies a popular neighborhood in Phoenix.
Example: Red Brick Arcadia Home located near the Biltmore corridor. 3 nice bedrooms. Newer High efficiency HVAC unit. Kitchen with newer stainless range/oven. dining room and living room. covered patio with pool and large yard. Block fence. single car garage. Nice area close to the Biltmore shops and easy access to downtown Phoenix.

Token: **moon** (N=976)

Description: The moon token identifies a popular neighborhood in Phoenix.
Example: Wonderful Moon Valley culdesac property with huge oversized lot at 17,271 sq. ft! This 4 bedroom, 2.5 bath home sits dead end in a quiet interior street in the Heart of Moon Valley. Vaulted ceilings in the Living room with Beams prow front window, 2X6 Construction, thermal pane windows, roof new shingle in 2010 dramatic fireplace for focal point, and huge green backyard!

Token: **camelback** (N=2,161)

Description: The camelback token identifies a popular neighborhood in Phoenix.
Example: Million dollar views! You may be washing dishes, but you are looking right at Camelback Mountain. This home has been completely remodeled with fabulous finishes including wood floors and marble floors. Relax in the truly spa-like master bath. Enjoy the lush extensive landscaping with common area behind; two patio areas including one to enjoy the Praying Monk. All this, and it is located in a simply wonderful community.

Token: **sendas** (N=543)

Description: The sendas token identifies a popular neighborhood in Phoenix.
Example: Absolutely stunning, Tuscan home in the prestigious community of Las Sendas Mountain! This home has been lovingly upgraded & meticulously maintatined. The spacious open great room boasts a Cantera Fireplace & plush carpet. The Gourmet Chef's Kitchen has a newer induction cooktop & built-in refrigerator. The private Backyard is ideal for entertaining w/the sparkling pool & spa, built-in BBQ & breathtaking Mountain Views!

Token: **viking** (N=739)

Description: The viking token identifies a popular high-end brand of appliances.
Example: Mountain views, greatest schools in the state, and the exclusive Ahwatukee Foothills community all enhance this stunning 5 bedroom, single story, fully remodeled home. New kitchen, new cherry cabinets, all Viking appliances, new travertine flooring throughout the home and of course a crystal blue pool and 3 car garage.

Token: **barn** (N=978)

Description: The barn token identifies properties that have an additional structure.
Example: Beautifully Remodeled 1 Acre Horse Property In Highly Desirable Sunburst Farms, 4 bdrms, 2 Bthrms. Split Floor plan, 2 Car Garage, 2 Car Tandem Work Shop or Garage, RV Parking, Barn & Newly Resurfaced Pool With A Large Covered Patio Great For Entertaining! Plenty Of Room For Your Horses & Toys.

Token: **membership** (N=1,336)

Description: The membership token identifies neighborhoods with community clubs.
Example: Move Right In To This Amazing 2Bdrm 2b Home With A Finished AZ Rm not in the SQFT Plus 2Car Garage w/Built In Storage Cabinets. Just Over From The Golf Course and Clubhouse. Community Club Membership Optional.

Token: **granite** (N=36,408)

Description: The granite token identifies houses with granite countertops.
Example: Feels like new 3 bedroom/2.5 bath house with downstairs den and large upstairs loft. 18' tile throughout with carpeted bedrooms and 2 tone neutral paint; kitchen includes granite counter tops and gas range/oven; wired for security system and surround sound. 2 car garage; front and back landscaping with watering systems. Washer/dryer included.

Token: **hillside** (N=1,039)

Description: The hillside token identifies houses located on the side of a hill (i.e. nice views).
Example: It always comes down to location in real estate & this house has it! Tranquil & private hillside lot located in a popular Ahwatukee Foohills neighborhood. Walking distance to top rated schools, hiking trails, shopping & restaurants. It's obvious from the time you pull in the driveway that the original owners have lovingly cared for this home.

C Robustness Checks (Internet Appendix)

C.1 Tuning parameters

The penalty parameters $1 < c, \gamma \approx 0$ determine λ_p, λ_d via $\lambda_p = \lambda_d = 2c\sqrt{N}\Phi^{-1}(1 - \gamma/2P)$ where Φ^{-1} is the inverse cdf of the standard normal and P is the total number of parameters in the model. Implicitly, c, γ determine $\hat{S}_2, \hat{Q}_2, \hat{\tau}$ by controlling the amount of regularization that prevents over-fitting. In the body of the paper, we use the penalty parameters (c= 1.10 and γ = 0.10) suggested in Belloni et al. (2014) to create the real estate specific dictionaries for Atlanta and Phoenix. Here, we consider $c \in \{1.01, 1.05, 1.1\}$ and $\gamma \in \{0.01, 0.10, 0.25\}$ in order to determine the sensitivity of $\hat{S}_2, \hat{Q}_2, \hat{\tau}$ to the choices of c, γ.

We use the filtered data set employed in Table 5. The filtered data set does not include vacant houses, rental properties, or agent-owned transactions by listing agents with more than three agent-owned transactions during the study period. Thus, the filtered data is a subsample of the data used to create the dictionaries detailed in Table 3.

The results in Table C1 show that the penalty parameters affect the number of tokens ($\hat{Q}_2$) selected for the sufficient dictionaries ($\widehat{S}_2^{zip}$) but do not have a material impact on the agent-owned estimates ($\hat{\tau}$). Given the sensitivity of $\hat{S}_2, \hat{Q}_2$ to the tuning parameters, we caution against interpreting the dictionary as the *true* real estate dictionary as $\hat{S}_2$ most likely 1) includes irrelevant tokens and 2) omits relevant tokens with small coefficients. However, identifying the true dictionary in a high-dimensional setting is most likely unrealistic (Buhlmann and Van De Geer, 2011). Moreover, our primary interest is to use tokens in the dictionary as controls when estimating τ and not to correctly identifying S_2. Despite imperfect variable selection, asymptotic inference on τ is still valid provided $\hat{S}_2$ includes the most important tokens (Belloni et al., 2014). In summary, regardless of the penalty imposed during the variable selection process, the agent-owned estimates reported in Table C1 are similar to the estimates reported in column 2 of Table 5.

Table C1: Agent-owned estimates with varying penalty parameters

Penalty		Atlanta (2007-2016)		Phoenix (2000-2013)	
c	γ	$\hat{Q}_2$	$\hat{\tau}$	$\hat{Q}_2$	$\hat{\tau}$
1.05	0.01	603	0.017	652	0.016***
1.10	0.01	586	0.017	618	0.015***
1.25	0.01	525	0.018	553	0.015***
1.05	0.10	660	0.016	708	0.016***
1.10	0.10	638	0.016	685	0.015***
1.25	0.10	578	0.017	611	0.015***
1.05	0.25	687	0.016	760	0.016***
1.10	0.25	664	0.016	716	0.016***
1.25	0.25	603	0.017	649	0.016***
Controls		✓	✓	✓	✓
Time x Zip FE		✓	✓	✓	✓
Tokens			$\widehat{S}_2^{zip}$		$\widehat{S}_2^{zip}$

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

C.2 Out-of-sample agent-owned estimates

The bulk of the empirical analysis focuses on in-sample estimation. In this section, we investigate the extent to which the textual analysis framework we provide can be used for out-of-sample estimation. To do so, we randomly split the full sample into equally sized training (A) and testing (B) subsamples. Next, we run the double-selection LASSO procedure using only the data in A to create market-specific real estate dictionaries ($\hat{S}_2^A$). Then, we estimate the agent-owned premium using only the transactions in the testing sample (B), while controlling for the attributes in $\hat{S}_2^A$. The results, which we present in columns 1 to 3 of Table C2, are nearly identical to Table 4. Finally, 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 results, which we present in columns 4 to 6 of Table C2, are similar to the estimates in columns 1 and 2 of Table 5.

Table C2: Out-of-sample agent-owned estimates

Panel A: Atlanta (2007-2016)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.039^{**}	$0.019^{\dagger}$	0.029^{*}	0.032^{*}	0.016	0.025
	(0.016)	(0.011)	(0.015)	(0.016)	(0.013)	(0.016)
Vacant	-0.121^{***}	$-0.077^{***\dagger}$	-0.101^{***}			
	(0.017)	(0.009)	(0.013)			
Vacant x Agent-owned	0.052^{*}	$0.009^{\dagger}$	0.008			
	(0.030)	(0.016)	(0.027)			
Rental	−0.034	−0.016	-0.044^{**}			
	(0.022)	(0.012)	(0.018)			
Rental x Agent-owned	0.007	−0.003	−0.005			
	(0.031)	(0.030)	(0.030)			
N	52,679	52,679	52,679	33,622	33,622	33,622
P	1,109	1,693	1,160	1,082	1,666	1,133
R^2	0.811	0.857	0.829	0.822	0.862	0.836
Panel B: Phoenix (2000-2013)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.042^{***}	$0.018^{***\dagger}$	$0.032^{***\dagger}$	0.037^{***}	$0.016^{***\dagger}$	0.029^{***}
	(0.005)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Vacant	-0.047^{***}	$-0.026^{***\dagger}$	-0.040^{***}			
	(0.006)	(0.004)	(0.006)			
Vacant x Agent-owned	-0.024^{***}	-0.015^{**}	-0.028^{***}			
	(0.008)	(0.006)	(0.008)			
Rental	-0.066^{***}	$-0.028^{***\dagger}$	-0.055^{***}			
	(0.010)	(0.006)	(0.009)			
Rental x Agent-owned	-0.040^{***}	$-0.011^{*\dagger}$	-0.032^{***}			
	(0.012)	(0.006)	(0.012)			
N	137,413	137,413	137,413	82,230	82,230	82,230
P	1,816	2,291	1,867	1,757	2,232	1,808
R^2	0.902	0.928	0.909	0.914	0.937	0.920
Controls	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓	✓	✓	✓
Tokens		$\hat{S}_2^A$	L&S		$\hat{S}_2^A$	L&S

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

C.3 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 C3 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 C3: Agent-owned estimates by token type

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Atlanta	0.029**	0.017	0.019	0.023*	0.031**	0.018	0.020*	0.026**
	(0.014)	(0.011)	(0.011)	(0.013)	(0.013)	(0.011)	(0.011)	(0.013)
Phoenix	0.036***	0.015***†	0.019***†	0.029***†	0.032***	0.014***†	0.017***†	0.026***†
	(0.004)	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)	(0.002)
Controls	✓	✓	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓	✓				
Time x Tract FE					✓	✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	$\hat{S}_2^{zip+}$	$\hat{S}_2^{zip-}$		$\hat{S}_2^{tract}$	$\hat{S}_2^{tract+}$	$\hat{S}_2^{tract-}$

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

Note: † denotes coefficient estimates in columns 2 to 4 (6 to 8) that are statistically different than the baseline estimate in column 1 (5) at the 10 percent level. Standard errors clustered at the time, location, and agent level are reported in brackets.

C.4 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 C4 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 client-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 pricing model helps control for idiosyncratic nature of the housing stock.

Table C4: Agent-owned estimates across heterogeneous subsamples

	Income				Heterogeneity					
	Low		High		Low		Medium		High	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Atlanta	0.022	0.012	0.041***	0.015†	0.017*	0.011	0.034**	0.023	0.038**	0.019
	(0.016)	(0.015)	(0.009)	(0.008)	(0.009)	(0.008)	(0.017)	(0.017)	(0.016)	(0.014)
Phoenix	0.046***	0.022***†	0.033***	0.012***†	0.028***	0.013***†	0.044***	0.020***†	0.040***	0.016***†
	(0.008)	(0.005)	(0.005)	(0.003)	(0.006)	(0.003)	(0.006)	(0.004)	(0.004)	(0.004)
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

Note: † denotes coefficient estimates in even columns that are statistically different than their corresponding estimate in odd columns at the 10 percent level. Standard errors clustered at the time, location, and agent level are reported in brackets.

C.5 Sample selection correction

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 C5 are similar to those reported using a specification without the IMR.

Table C5: Agent-owned estimates with sample correction

	(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^{***}	$0.017^{***\dagger}$	$0.029^{***\dagger}$	0.032^{***}	$0.015^{***\dagger}$	$0.025^{***\dagger}$
	(0.004)	(0.003)	(0.003)	(0.003)	(0.002)	(0.002)
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

Note: † denotes coefficient estimates in columns 2 and 3 (5 and 6) that are statistically different than the baseline estimate in column 1 (4) at the 10 percent level. Standard errors clustered at the time, location, and agent level are reported in brackets.

C.6 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 C6 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 tokens in the $\hat{S}_2$ dictionary. Column 3 and 6 use the dictionary described in Levitt and Syverson (2008). The estimates in Table C6 are similar to those reported using an OLS specification without controlling for time-on-market.

Table C6: 2SLS agent-owned estimates

	(1)	(2)	(3)	(4)	(5)	(6)
Atlanta	0.031**	0.017	0.023*	0.033***	0.019	0.025**
	(0.014)	(0.012)	(0.013)	(0.012)	(0.011)	(0.011)
Phoenix	0.037***	0.016***†	0.028***†	0.035***	0.013***†	0.027***†
	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)	(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

Note:† denotes coefficient estimates in columns 2 and 3 (5 and 6) that are statistically different than the baseline estimate in column 1 (4) at the 10 percent level. Standard errors clustered at the time, location, and agent level are reported in brackets.

C.7 Short holding periods

Following Levitt and Syverson (2008), we drop houses that sold twice within a three year period. The filter removes houses that were likely rehabbed and flipped by investors, while also allowing a direct comparison of our results with the extant literature. In this section, we rerun the dictionaries and empirical analysis using *only* houses that sold twice within a three year period. The agent-owned coefficient estimates reported below are comparable to those reported in Table 4.

Table C7: Agent-owned estimates for short holding periods by tenure

Panel A: Atlanta (2007-2016)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.036^{**}	0.020	0.029^{**}	0.036^{**}	0.015	0.030^{*}
	(0.015)	(0.014)	(0.014)	(0.016)	(0.018)	(0.016)
Vacant	-0.124^{***}	$-0.087^{***\dagger}$	-0.107^{***}	-0.104^{***}	$-0.070^{***\dagger}$	-0.088^{***}
	(0.016)	(0.010)	(0.013)	(0.015)	(0.009)	(0.011)
Vacant x Agent-owned	0.054^{**}	$0.008^{\dagger}$	0.015	0.072^{***}	$0.021^{\dagger}$	$0.031^{\dagger}$
	(0.022)	(0.018)	(0.024)	(0.017)	(0.014)	(0.019)
Rental	0.005	0.004	−0.016	0.029^{*}	0.019	$0.005^{\dagger}$
	(0.017)	(0.009)	(0.013)	(0.016)	(0.012)	(0.014)
Rental x Agent-owned	0.005	−0.018	0.001	0.026	0.009	0.019
	(0.039)	(0.033)	(0.042)	(0.035)	(0.037)	(0.040)
N	25,415	25,415	25,415	25,415	25,415	25,415
P	1,115	1,406	1,164	4,894	5,194	4,943
R^2	0.779	0.841	0.808	0.853	0.894	0.875
Panel B: Phoenix (2000-2013)						
	(1)	(2)	(3)	(4)	(5)	(6)
Agent-owned	0.043^{***}	$0.019^{***\dagger}$	$0.032^{***\dagger}$	0.038^{***}	$0.018^{***\dagger}$	$0.027^{***\dagger}$
	(0.004)	(0.003)	(0.004)	(0.004)	(0.003)	(0.004)
Vacant	-0.037^{***}	$-0.020^{***\dagger}$	-0.033^{***}	-0.031^{***}	$-0.017^{***\dagger}$	-0.028^{***}
	(0.004)	(0.003)	(0.004)	(0.003)	(0.002)	(0.003)
Vacant x Agent-owned	-0.018^{***}	$-0.011^{**\dagger}$	-0.021^{***}	-0.008^{*}	-0.007^{*}	-0.012^{**}
	(0.005)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
Rental	-0.064^{***}	-0.024^{***}	-0.048^{***}	-0.054^{***}	-0.020^{***}	-0.039^{***}
	(0.012)	(0.005)	(0.010)	(0.010)	(0.004)	(0.008)
Rental x Agent-owned	-0.035^{***}	−0.017	-0.024^{*}	−0.023	−0.009	−0.015
	(0.013)	(0.011)	(0.013)	(0.015)	(0.013)	(0.014)
N	145,803	145,803	145,803	145,803	145,803	145,803
P	1,787	2,261	1,838	10,875	11,384	10,926
R^2	0.900	0.927	0.909	0.932	0.948	0.937
Controls	✓	✓	✓	✓	✓	✓
Time x Zip FE	✓	✓	✓			
Time x Tract FE				✓	✓	✓
Tokens		$\hat{S}_2^{zip}$	L&S		$\hat{S}_2^{tract}$	L&S

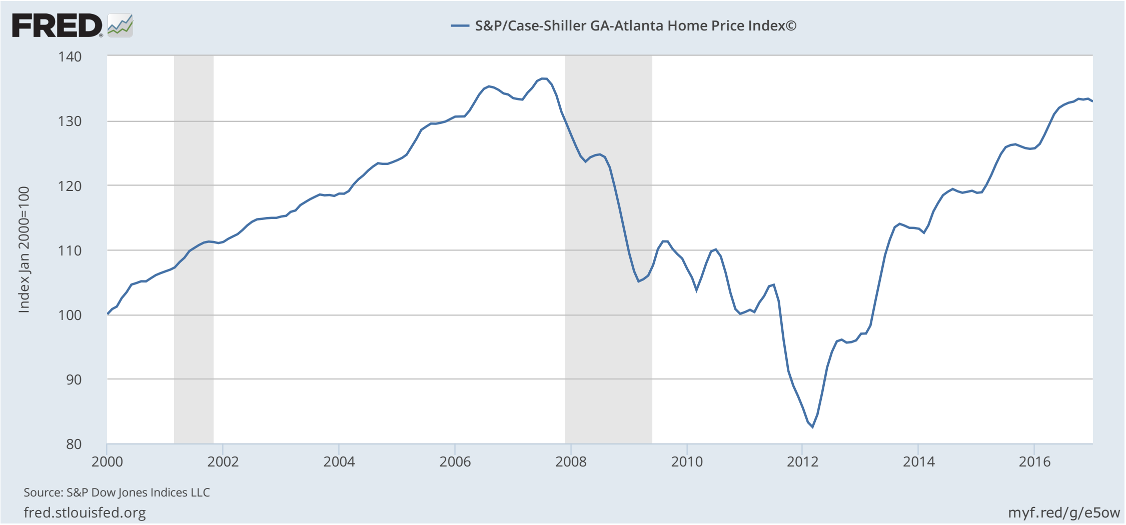
*p<0.1; **p<0.05; ***p<0.01

D Miscellaneous (Internet Appendix)

D.1 HPI for subperiod analysis

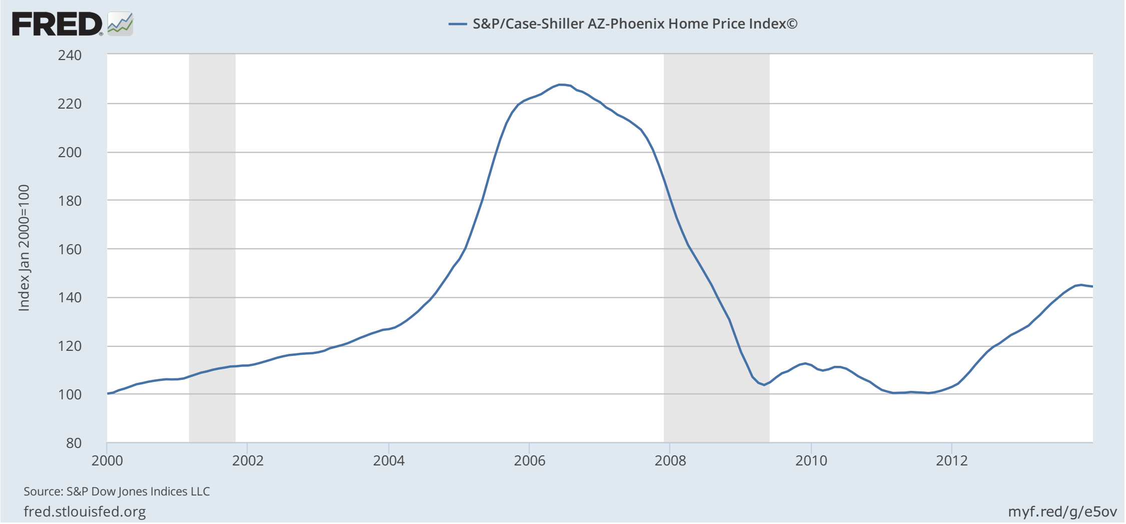
We use the Case-Shiller Repeat-sales Indices for Atlanta and Phoenix to identify the appropriate cutoffs for the subperiod analysis. The subperiods in Atlanta represent bust (2007-2011) and recovery (2012-2016) periods. The subperiods in Phoenix represent pre-boom (2000-2003), boom (2004-2006), bust (2007-2009), and recovery (2010-2013) periods. The repeat-sales indices displayed in Figure D1 and Figure D2 were downloaded from the St. Louis Federal reserve website.

Figure D1: Atlanta Case-Shiller Repeat-sales Index



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

Figure D2: Phoenix Case-Shiller Repeat-sales Index



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