# Pricing Neighborhood Amenities: A Proxy-Based Approach[*]

Alex Bell [†] Sophie Calder-Wang[‡] Shusheng Zhong[§]

December 31, 2023

## Abstract

Understanding how housing markets price neighborhood amenities is key to unpacking residential socioeconomic disparities. Yet prior approaches to amenity pricing have suffered from the key confounder of unmeasured neighborhood quality, often leading to wrong-signed estimates. In this paper, we develop a novel proxy-based method that allows us to estimate the price of housing amenities in the midst of unobserved housing quality. Our method shows, as long as one can find at least one relevant "proxy" variable for housing quality that satisfies conditional independence, the price of amenity can be identified as the ratio of two coefficients. Using detailed migration data, we construct an innovative measure of locational desirability–Geographic PageRank–to use as the proxy variable for quality. We show that this new approach can successfully correct the "wrong-signed" problem in the amenities valuation literature when applied to a standard measure of environmental air quality. The estimated amenity prices are key inputs to evaluating returns on investment in local public goods or environmental policies, including their role in reducing housing disparities.

*Keywords:* Amenities, Omitted Variable Bias, PageRank

*JEL Code*: R32, C39, Q51

[*] Calder-Wang thanks the Zell/Lurie Real Estate Center at Wharton for financial support. Code to implement the methods proposed in the paper is available as a Stata package in the following github repository https://github.com/ZhongShusheng/proxy_stata_package.

[†] University of California Los Angeles *alexbell@ucla.edu*

[‡] The Wharton School, University of Pennsylvania *sophiecw@wharton.upenn.edu*

[§] Kellogg School of Management, Northwestern University *shusheng.zhong@kellogg.northwestern.edu*

# 1 Introduction

Understanding the prices of housing and neighborhood amenities is key to answering several important economic questions in local public finance, housing inequality, and geographic mobility. Housing can be thought of as a bundle of amenities: besides the usual house-level attributes such as the number of bedrooms, number of bathrooms, and lot size, we are also concerned about the valuation of local amenities such as crime, congestion, school quality, and environmental (dis)amenities such as air pollution, fire, and flood risk. The prices of these neighborhood amenities provides us with guidance towards how public funds should be allocated locally and how to better mitigate environmental externalities. Furthermore, given that households have heterogeneous preferences for these amenities, credible estimates of the trade-offs imposed by the housing market would also help researchers unpack how much observed housing inequality is due to different preferences as opposed to other factors such as discrimination or income.

Despite the broad economic relevance and significance of housing consumption, traditional methods based on observational studies are often unsatisfactory in providing credible estimates of the prices of amenities. The challenges are so extreme that the price of neighborhood amenities based on regression methods is prone to the "wrong-signed" problem (Chay and Greenstone 2005). In Figure 1, we show the binscatter of home price against the air quality in the U.S., where we find a mostly upward-sloping pattern: places with worse air quality have higher prices, which generates a wrong-signed positive price estimate of air pollution. Given that people do not desire air pollution, it is logical to conclude that there must be some sort of desirability that is not presently quantified but also characterizes these high-priced but polluted neighborhoods. To address the bias from unmeasured quality, the usual approach is to control for the omitted variables that may have contributed to or be correlated with that unobserved quality, for example, by controlling for household income. We call such control variables "proxies" for quality.

While the dotted line in Figure 1 shows that controlling for income appears to move estimates in what might naturally be termed the "right direction," the wrong-signed problem persists. As stylized as it may appear, this simple example highlights the fundamental challenge faced by approaches centered on observational data: it is simply not feasible to include sufficient controls for quality to construct the apples-to-apples comparison one would hope to show from the data. As a result, it is not surprising that researchers have since departed from using observational data but moved on to finding natural experiments that could provide quasi-random variations in the assignment of amenities. However, this approach is not without limitations: Although such natural experiments tend to provide more credible estimates, they are inherently narrow in scope and incomplete in coverage.

The key contribution of this paper is to develop a novel estimation procedure that allows us to identify amenity prices, namely how prices vary by amenity while holding all else equal, in observational, non-experimental data: Rather than needing to fully observe all aspects of housing quality, our approach simply requires us to find at least one relevant "proxy" for housing quality that satisfies an appropriate conditional independence assumption. Then, the price of amenity can be identified by taking the ratio of the rate at which the proxy variable on average varies with amenity, to the rate at which the proxy variable varies with prices.

Motivated by classic sorting models of Rosen (1974) and Roback (1982), we enrich them with two additional strands of literature that illustrate how both horizontal amenity differentiation (Bayer, Ferreira and McMillan, 2007) and vertical quality segmentation (Epple and Sieg, 1999; Epple, Quintero and Sieg, 2020) can matter for the housing market. By applying our approach to the trade-off homebuyers face for environmental air quality, we illustrate how—contrary to popular wisdom—observational data on the housing market can yield reliable estimates of amenity trade-offs. In the case of air quality, we find that the trade-off homebuyers face for homes with better air quality is similar to quasi-experimental estimates of the causal effects of isolated changes in air quality on home prices (Chay and Greenstone, 2005; Currie et al., 2015). We draw identification from the revealed-preference methodology made popular in labor economics (Sorkin, 2018; Fogel, 2021) by constructing a novel proxy for place quality based on the PageRank algorithm. Lastly, our approach in this paper builds on one of the authors' work in a companion paper on labor market amenities (Bell, 2022), which is already proving its usefulness for labor economists looking to estimate wage-amenity tradeoffs (Folke and Rickne, 2022; Burbano et al., 2022).[1] Econometrically, our estimation framework is conceptually connected to the selection-correction approach of Altonji, Elder and Taber (2005) and Oster (2019), but with a totally different context and setting.

Concretely, we develop a novel proxy-based approach that allows us to estimate amenity prices using observational data. Two key assumptions built on insights from the real world and the literature are required for identification, which are adapted as follows from Bell (2022). First, the housing market must be well approximated by our *Single Index Assumption.* Second, there must exist a relevant proxy variable $H$ for quality that satisfies *Conditional Independence.*

*The Single Index Assumption* is an economic assumption about how prices paid and amenities chosen

[1] In addition to studying housing rather than job amenities, the present paper has made several improvements on the conceptual framework of Bell (2022), including additional proofs of identification (Appendix A), the derivation of a bias formula under the case of missing amenities (Appendix C) and developing the equivalence of using supply-side vs. demand-side proxies when quality plays a salient role in markets (Appendix B and D).

are determined in a given market. Formally, the single index assumption states that market prices are determined by the combination of observable amenities plus a *single* index of quality. Consequently, for any product in the market, we model that there exists a unique quality index $\Phi$ that explains its price $P$ and amenity $Z$. In the housing context, this assumption requires that the price-relevant characteristics of a house include observable amenities plus a single unobserved component of quality.

*The Conditional Independence Assumption*, on the other hand, is an econometric assumption about the property of a particular proxy variable for quality. Formally, conditional independence assumes that, conditioning on being in a particular quality segment $\Phi$, the proposed proxy variable $H$ is not correlated with amenity: $H \perp Z \,|\, \Phi$. In other words, we need a proxy variable that is relevant for the unobserved quality,[2] but is *not* correlated with the amount of amenities chosen within that given level of quality.

With these two assumptions, we construct the estimator for an amenity price based on the ratio of how the predicted proxy varies by the level of amenity, compared to how it varies by price,

$$\frac{\partial p}{\partial z} = -\frac{\partial \hat{H}}{\partial z} \Big/ \frac{\partial \hat{H}}{\partial p} \tag{1.1}$$

where the predicted proxy is simply defined by the expectation operator $\hat{H}(p,z) \equiv \mathbb{E}[H|P=p, Z=z]$.

Intuitively, the price of an amenity in dollar terms is the rate of change in $P$ with respect to a change in $Z$ within the level set of true unobserved quality $\Phi$. It turns out that when $H$ satisfies conditional independence, the predicted proxy $\hat{H}$ generates the identical level sets as $\Phi$. As a result, the price of the amenity can be estimated by examining how price $P$ varies with amenity $Z$ by just holding the predicted proxy $\hat{H}$ constant, without the need to explicitly hold the (poorly observed) underlying quality constant.[3]

In a model where price is a linear function of amenity and quality, then, given a proxy variable that satisfies linear conditional independence, the price of the amenity can be represented neatly as a simple ratio or regression coefficients $\hat{\beta}_z = -\frac{\delta_z}{\delta_p}$ , where $\delta_z$ and $\delta_p$ are the coefficients of a linear regression of the proxy variable $H$ on $Z$ and $P$, respectively. We also offer a variety of extensions for

[2] "Relevance" is a technical condition that requires the proxy variable to be at least on average related to the underlying quality. We formalize it as monotonicity between the predicted proxy function $\mathbb{E}[H|\Phi]$ and $\Phi$. While there may exist idiosyncratic deviations in how well the proxy variable represents the true unobserved quality, the predicted proxy $\mathbb{E}[H|P,Z]$ averages out the idiosyncratic components and has to change in tandem with true underlying quality.

[3] In Section 3, we illustrate the intuition behind the estimator, which is also shown graphically in Figure 3, through the lens of the multivariate implicit function theorem. The gradient of the predicted proxy function $\hat{H}$ represents the direction of the steepest ascent in the predicted proxy, shown by the blue arrow. The amenity's price in dollar terms is the rate of change in $P$ with respect to a change in $Z$ within the level set of $\hat{H}$, which is represented by the slope of the tangent vector of the level set, shown by the green arrow. The tangent vector of the level set is, by definition, perpendicular to the gradient of the function.

the estimator when there are multiple amenities and multiple plausible proxies.

We believe that both these two assumptions could be reasonably satisfied within typical empirical settings. In terms of the Single Index Assumption, while seemingly restrictive, we believe it is entirely plausible utility-relevant features of the housing market includes both amenities and quality. Moreover, the Single Index Assumption does not place any specific restrictions on the pricing setting process, and thus can be satisfied under a variety of price setting processes as long as certain regularity conditions are met. In terms of the Conditional Independence Assumption, finding a variable that satisfies "conditional indepedence" essentially boils down to finding a variable that has to be a "good" control for quality in the typical OLS: it is not otherwise correlated with amenity once controlled for quality.[4] While OLS requires finding all the controls, the key benefit of the proxy method is that identification can be obtained as long as the researcher can find just *one* such good control.

Empirically, we apply the proxy-based method to the problem of estimating neighborhood amenities in the housing context. We construct a novel proxy for the quality of a place, the Geographic PageRank (GPR), which enables us to estimate amenity prices.[5] It is adapted from the well-known PageRank algorithm to the housing context. Specifically, we use migration flows as the basis, and places relate to each other with a recursive measure of desirability: locations that draw migration are considered desirable, whereas locations that draw migration from other desirable locations are considered even more desirable. Using county-to-county migration data from the IRS, we compute the Geographic PageRank for all U.S. counties. In addition, we also compute measures of locational desirability at more granular geographic units (ZCTA) using individual-level migration data from Infutor.

Then, we illustrate the performance of the proxy-based method on the problem of pricing air pollution. In particular, we believe that the proposed Geographic PageRank serves as a reasonable proxy for quality because much of migration is likely driven by households moving towards overall more desirable places, so it is unlikely that the ranking, which is a measure of central tendencies, is particularly correlated with one's preference for better or worse air, even if specific households may have moved due to their preference for air quality. While the naïve OLS and OLS with controls perform poorly in terms of recovering even the correct sign for the price of air pollution, we find that the proxy-based valuation of air pollution does recover the correct sign. Moreover, the proxy-based estimate does fall in the reasonable range of price elasticities compared to the existing literature, providing some validation for the proposed method.

[4] See Section 2.E of Bell (2022) for more on the relation of the proxy assumption to good controls.

[5] In contemporaneous unpublished work, Fogel (2021) also applies the algorithm to examine residential desirability, though his approach focuses on providing an estimate of city value and not on how it may be used as a proxy for amenity pricing.

Overall, compared to the existing literature, our proxy-based approach has several major advantages. First, a crucial advantage of our proxy-based estimation procedure is that researchers are no longer required to come up with a complete list of controls for all components of unobservable quality. Rather, the focus is shifted towards finding an appropriate "proxy" variable that can act as a "shifter" for the unobservable quality. Importantly, such a "proxy" has to be conditionally independent with the amenity once true quality is controlled for, or more intuitively, it has to be a "good" control for quality. Moreover, because the method is based on observational data alone, it can be widely applied to a variety of settings without being constrained by the availability of specific natural experiments. Furthermore, the proxy-based method can produce more nuanced prices that can vary by the quality segments of the market, which allows for amenity prices to differ for different household income levels. Lastly, the implementation of the proxy-based method is also computationally inexpensive.

While the proxy approach offers a clear advantage over existing observational methods for analyzing amenity prices, limitations remain. Despite the strength of our approach in holding fixed a single index of housing quality, we cannot control for or price additional amenities that the researcher does not observe. The extent whether this is a limitation depends entirely on the equilibrium quantity the researcher is seeking to estimate. Our approach will not pin down a hypothetical price of an amenity holding all other amenities and quality constant (and to our knowledge, there is no such experiment or other approach that promises to do this for an arbitrary set of amenities across a whole market). Instead, our method does estimate the price a home buyer would pay to obtain the additional amenity, factoring in the existing correlation between the observed amenity and unobserved amenities as well as unobserved quality in the market. To the extent one cares about the actual amount home-buyers would need to pay to attain a certain additional amenity in the current market, our estimator is designed to identify such object. Regardless, even with missing amenities, the proxy estimator is still superior to what an OLS estimator could generate, because the proxy method still remains robust to the presence of unobserved quality.

The rest of the paper is organized as follows. Section 2 places this work in relation to the broader literature on housing amenities and selection-correction methodologies. Section 3 formalizes the theory behind the proxy approach to estimating amenity prices. We first prove that amenity prices can be consistently estimated in the general case with an appropriately-defined proxy variable. We then proceed with the special linear case, where we show that the amenity price can be estimated as the ratio of two regression coefficients. We also describe a set of relevant extensions to multiple amenities and multiple plausible proxies. In Section 4, we provide an empirical application of the proxy method in estimating the price of environmental amenities. Section 4.1 describes the data sources. In Section 4.2, we construct the Geographic PageRank as a proxy variable for quality using migration flows. In Section 4.3, we apply the Geographic PageRank to the problem of estimating the price of air pollution. Section 5 concludes.

# 2 Relation to Prior Literature

## 2.1 Relation to Literature on Valuing Amenities

Classic literature has used hedonic regression to estimate the price of amenities in the housing market. The theoretical foundation of hedonic models is developed in Rosen (1974) and Roback (1982). Under their spatial equilibrium model, housing price is a function of housing characteristics ("amenities"), and the prices of these amenities are jointly determined by the heterogeneous preference of perfectly mobile residents and the technology for procuring these amenities. The implicit price of housing amenities can thus be inferred using the hedonic method. Our key improvement is to break from the traditional spatial arbitrage assumption by making explicit vertical market segmentation.

However, in empirical applications, hedonic regression, even with controls, often reveals a weak relationship between amenities and housing prices, and sometimes the estimated price of the amenity might even be "wrong-signed" compared to common belief. This suggests that the spatial equilibrium assumptions might be too strong to allow for credible price estimates in the hedonic method. One main concern is that households face different budget sets and, therefore may choose different levels of housing quality. A partial step towards resolving this issue is to include household socio-economic characteristics (usually income) as controls for the housing quality chosen (Palmquist 1984; Black 1999), yet this approach is at best going to partially capture the confounding effect from unobserved housing quality.[6] Our proxy method does not require all components of housing quality to be observed and, therefore, resolves the issue more systematically.

One of our structural assumptions of the housing market, namely, the single index assumption, is motivated by the recent structural approaches toward residential sorting.[7] Our work combines the key elements of two popular structural approaches to modeling housing as a good: The first approach (Epple and Sieg 1999; Epple, Quintero and Sieg 2020) models housing as a vertically-differentiated good, while the second (Bayer, Ferreira and McMillan 2007) focuses on modeling housing as a horizontally-

[6] There is a parallel literature in labor economics on the "wrong-signed" problem as well. There, it frequently arises when using hedonic regression to estimate the price of job amenities with wage data (Brown 1980; Kniesner et al. 2012). Similarly, there were attempts to control for worker's unobserved productivity using observed worker characteristics as controls (Lucas 1977; Mas and Pallais 2017).

[7] See **?** for a comprehensive review of recent structural approaches.

differentiated good. One key element of Epple and Sieg (1999) is that there is a single latent variable—housing quality—that comprehensively summarizes all utility-relevant amenities of a house, and households commonly agree on the ranking of all houses by this measure of quality. We model this vertical dimension of the housing market with a similar single-index assumption about quality. However, we also incorporate household preference heterogeneity and product differentiation, which are the key elements of horizontal differentiation models. This flexibility has the key advantage of allowing us to estimate the trade-off residents face to access homes with more of a particular housing amenity (e.g., pollution, flood, or crime), which are often of important policy interest.

Our paper also contributes to the literature that estimates the effect of local environmental factors on housing values. Many papers in this literature, like Chay and Greenstone (2005) and Currie et al. (2015), combine quasi-experimental settings with hedonic methods to derive internally valid estimates of marginal willingness to pay for environmental amenities. Our approach complements this literature by providing an estimation method that can be applied to more general settings where quasi-experiments might not be readily available.

Finally, our paper builds on the nascent literature on PageRank that uses flow data to construct a "revealed-preference-based" measure of desirability. Sorkin (2018) applies this method with data on worker flows to rank firms by their attractiveness, and Fogel (2021) uses the same method with migration flow data to measure the value of American cities. Our paper extends this method by constructing Geographic PageRank at a much finer geographic level and develops econometric methods to fully utilize the constructed PageRank in the estimation of housing amenities.

## 2.2 Relation to Selection Correction Literature

Our estimation strategy can, to some extent, be re-cast in the language of the literature on coefficient corrections based on selection, as put forward by Altonji, Elder and Taber (2005) and Oster (2019). The key difference relative to this literature lies in the context to which we apply the correction. Historically, this literature has focused on inferring treatment effects when the structural equation for a single endogenous treatment variable is partially observed. For instance, Altonji, Elder and Taber (2005) study the effect of Catholic schools on earnings, with an eye toward holding fixed the multitude of socioeconomic factors that jointly determine both Catholic school attendance and earnings.[8] In contrast, we are interested in estimating price parameters that arise from a set of simultaneous equa-

[8] Analogously, we would call these observed factors proxies for that latent variable that jointly determines schooling and income.

tions.[9] We put forward our proxy framework to exploit the natural symmetry of the amenity pricing problem: (observed) prices paid and amenities chosen are jointly functions of (unobserved) housing quality and household preferences. Section 3.5.3 provides further detail on the relation of our exact assumptions to those of the selection correction literature.

# 3 Model and Methodology

## 3.1 Motivation for the Proxy Method: "Alice and Bob"

An illustrative example helps to explain the motivation and intuition for the proxy-based method as an approach towards pricing amenities.

First, consider the naive approach towards estimating the amenity price using observational data by simply regressing prices on amenities. Figure 2 Panel (a) illustrates this approach. In this example, there are two individuals: Alice and Bob. Alice lives in an expensive townhouse in downtown Los Angeles with poor air quality. Bob lives in an inexpensive townhouse in Susanville, California, with pristine air. A regression of price on amenity would find that the higher price of Alice's house compared to Bob's house must be justified by the difference in air quality, concluding a positive price for air pollution, which clearly has the wrong sign.

So, why does this naive regression of price on amenity fail? To the extent that there are other unobserved quality differences between Alice and Bob's houses, it would all be erroneously attributed to the price of air quality. In other words, Alice's house is expensive probably because it is of much higher quality than Bob's house, illustrated by Figure 2 Panel (b), where Alice's house sits on a higher quality segment than Bob's house.[10]

Naturally, a sensible approach to amend this problem of missing quality is to find other homes with comparable quality to Alice's house. For example, Figure 2 Panel (c) shows that Ashley's house in Orange County, California, might be a suitable candidate: it sits on the same quality segment, but it has better air than Alice's house in downtown LA and is also more expensive. As a result, by comparing the price difference between Ashley's house and Alice's house, where both of them sit on the *same* quality segment, we are able to correctly deduce the (negative) price of air pollution.

While the conventional approach would recommend the inclusion of additional control variables in

[9] Because this model is largely isomorphic to that of supply and demand, much of our discussion here carries over to how one might use a proxy instead of an instrument to identify that model.

[10] In this example, Susanville was the subject of a 2007 documentary entitled *Prison Town USA*, which chronicles residents struggling to get by as their town turns to the prison industry to save it from mounting economic and social problems.

the OLS to address the missing quality in the regression, it often becomes untenable because obtaining correct inference would then require the researcher to be able to control for essentially *all* components that contribute to quality, which is generally impossible.

The core innovation of our paper is that we can eschew the need to include all quality-relevant controls by the usage of a "proxy" variable: all we need is just *one* relevant proxy variable that satisfies the "conditional independence" assumption, namely, the proposed "proxy" variable is relevant in terms of being informative of the quality segment, but *not correlated* with preferences for amenities within a given quality segment. Then, instead of having to condition on the unobserved quality, we can condition on the expected level of the proxy variable, which allows us to consistently estimate the price of the amenity. In the next section, we provide a more formal account of the theory.

## 3.2 Theory of the Proxy Method

In this section, we state the the assumptions underlying the proxy approach and the main theorem of this paper. We develop the theorems in both a general (non-parametric) framework and in a parametric linear form. We illustrate the use of both frameworks in our empirical application in Section 4.

### 3.2.1 Theory of the Proxy Method (General Case)

**Assumption 3.1.** ***(Single Index Assumption)*** *The price of a home $P$ is determined by its amenity $Z$ and a* ***single*** *index of housing quality $\Phi$, where the pricing relationship $P = P(Z, \Phi)$ is invertible in both $Z$ and $\Phi$.*

Equivalently, the *Single Index Assumption* implies that there exists a *single* index of housing quality $\Phi = \phi(P, Z)$ that fully explains the variations in price $P$ and the amenity $Z$. For differentiable functions, a sufficient condition is that the mapping $P(\cdot, \cdot)$ is continuously differentiable and has non-zero derivative. Then, applying the Inverse Function Theorem, the *Single Index Assumption* implies that the triplet $(P, Z, \Phi)$ has the property where knowing any two of them determines the third element:

$$P = P(Z, \Phi) \tag{3.1}$$

$$\Phi = \phi(P, Z) = P^{-1}(P, Z) \tag{3.2}$$

$$Z = Z(P, \Phi) = P^{-1}(P, \Phi). \tag{3.3}$$

Eq (3.1) means that variations in price $P$ is fully explained by variations in amenity $Z$ and the quality index $\Phi$; Eq (3.2) means that variations in the quality index $\Phi$ is fully explained by variations in price $P$ and amenity $Z$; Eq (3.3) means that variations in amenity $Z$ is fully explained by variations in price $P$ and the quality index $\Phi$.

**Assumption 3.2.** ***(Conditional Independence)*** A proxy variable $H$ is conditionally independent of $Z$ if

$$H \perp Z \,|\, \Phi. \tag{3.4}$$

In other words, a proxy variable $H$ satisfies the conditional independence assumption if it is *not* informative about the level of amenity $Z$ once conditioned on a particular level of quality $\Phi$.

Note that the *Conditional Independence* assumption is an econometric assumption about a specific proxy variable $H$, whereas the previous *Single Index Assumption* is an economic assumption about the underlying structure of how prices are determined.

**Assumption 3.3.** ***(Strict Monotonicity)***

$$\mathbb{E}[H|\Phi] \text{ is strictly monotone in } \Phi. \tag{3.5}$$

The monotonicity assumption stipulates that the proxy variable under consideration $H$ is relevant: it is on average informative about the quality index $\Phi$.

**Lemma 3.4.** *In a market that satisfies the Single Index Assumption, if a proxy variable $H$ satisfies Conditional Independence, then the predicted proxy function $\hat{H}(\cdot,\cdot)$ satisfies*

$$\hat{H}(P=p, Z=z) \equiv \mathbb{E}[H \,|\, P=p, Z=z] = \mathbb{E}[H \,|\, \Phi=\phi] \equiv \hat{H}(\phi). \tag{3.6}$$

*where $\phi = \phi(p,z)$.*

*Proof.* See Appendix A.1. □

In other words, the lemma stipulates that the value of predicted proxy $\hat{H}(p,z)$ can be solely determined by the predicted proxy $\hat{H}(\phi)$ based on the corresponding quality $\phi = \phi(p,z)$. This lemma forms the basis of our main theorem.

**Theorem 3.5.** ***(The price of the amenity is identified using a proxy variable.)*** *Consider a market that satisfies the single index assumption as described in Assumption 3.1. If there exists a proxy variable $H$ that satisfies conditional independence (Assumption 3.2) and strict monotonicity at*

*$(p, z)$ (Assumption 3.3), then the price of the amenity is identified and can be consistently estimated as*

$$\left.\frac{\partial P}{\partial Z}\right|_{(p,z)} = -\frac{\partial \hat{H}(p,z)}{\partial Z} \Bigg/ \left.\frac{\partial \hat{H}(p,z)}{\partial P}\right|_{(p,z)} \tag{3.7}$$

*where $\phi = \phi(p, z)$, and $\hat{H}(\cdot, \cdot)$ refers to the predicted proxy function*

$$\hat{H}(p, z) = \mathbb{E}[H \mid P = p, Z = z]. \tag{3.8}$$

*Proof.* See Appendix A.1. □

The intuition of our main theorem is as follows. By definition, the price of the amenity at any given observed $(p_0, z_0)$ is determined by how prices change in response to a change in amenity, holding its unobservable quality $\phi_0 = \phi(p_0, z_0)$ fixed. Thus, that the amenity price can be represented by the tangent direction of the level set of $\phi(p, z)$ valued at $\phi_0$, as illustrated in Figure 3. When there exists a proxy variable $H$ that satisfies Conditional Independence, Lemma 3.4 says that the predicted proxy function satisfies $\hat{H}(p, z) = \hat{H}(\phi)$. This means that the level set of the function $\hat{H}(\cdot, \cdot)$ valued at $\hat{H}(\phi_0)$ is *identical* to the level set of $\phi(\cdot, \cdot)$ when valued at $\phi_0$

$$\Big\{(p, z) : \hat{H}(p, z) = \hat{H}(\phi_0)\Big\} = \Big\{(p, z) : \phi(p, z) = \phi(\phi_0)\Big\}, \tag{3.9}$$

thus producing the same tangent vector. Therefore, the amenity price can be estimated as the tangent vector of the level set of the predicted proxy function, namely,

$$\left.\frac{\partial p}{\partial z}\right|_{p_0,z_0} = -\frac{\partial \phi}{\partial z} \Big/ \left.\frac{\partial \phi}{\partial p}\right|_{p_0,z_0} = -\frac{\partial \hat{H}}{\partial z} \Big/ \left.\frac{\partial \hat{H}}{\partial p}\right|_{p_0,z_0}. \tag{3.10}$$

Moreover, the tangent direction is by definition perpendicular to the gradient of the predicted proxy function $\nabla \hat{H} = [\partial \hat{H}/\partial Z, \partial \hat{H}/\partial P]$, where the gradient represents the direction of increasing quality.

Therefore, even though we cannot fully observe the true quality $\phi$, because $\hat{H}(\cdot, \cdot)$ and $\phi(\cdot, \cdot)$ have identical level sets, it allows us to identify the correct amenity price as long as we can estimate $\hat{H}$, which is the key intuition for why the proxy method can work in the face of unobserved quality $\phi$. Moreover, it also suggests that an equivalent estimator is to simply control for $\hat{H}$ directly.

### 3.2.2 Theory of the Proxy Method (Linear Case)

In this section, we provide the formulation of the main theorems when amenity $Z$ and the quality index $\Phi$ enters linearly into the price.

**Assumption 3.6.** ***(Linear Single Index Assumption)*** There exists a single linear latent index of housing quality that fully explains the variations in price $P$ and amenity $Z$. Without loss of generality, let the true model of price be determined by $Z$ and $\Phi$ in a linear model as follows

$$P = \beta_0 + \beta_z Z + \Phi. \tag{3.11}$$

Notice that Assumption 3.6 represents the special linear case of Assumption 3.1 where $\phi(P, Z) = P - \beta_0 - \beta_z Z$. In addition, the true price of the amenity is $\beta_z$.

**Assumption 3.7.** ***(Linear Conditional Independence)*** A proxy variable $H$ satisfies linear conditional independence if

$$cov(H, Z^{\perp\Phi}) = 0 \tag{3.12}$$

where $Z^{\perp\Phi}$ denotes the residual of the linear projection from $Z$ to $\Phi$.

**Assumption 3.8.** ***(Relevance)*** The proxy variable is relevant:

$$cov(H, \Phi) \neq 0. \tag{3.13}$$

**Definition 3.9.** Define the linear projection operator $\mathbb{E}^*[Y|X]$ as follows:

$$\mathbb{E}^*[Y|X] = \beta_0 + \beta_x X \tag{3.14}$$

where $\beta_x$ denote the OLS coefficients.

**Theorem 3.10.** ***(The price of the amenity is identified by the "ratio of regression coefficients" in a linear model.)*** *Consider a market that satisfies the linear single index assumption as described in Assumption 3.6 as follows*

$$P = \beta_0 + \beta_z Z + \Phi. \tag{3.15}$$

*If there exists a proxy variable $H$ that satisfies linear conditional independence (Assumption 3.7) and is* *relevant (Assumption 3.8), then the price of the amenity is identified and can be consistently estimated as the ratio of two regression coefficients*

$$\beta_z = -\frac{\delta_z}{\delta_p} \tag{3.16}$$

*where $\delta_z$ and $\delta_p$ are OLS regression coefficients of $H$ on $Z$ and $P$ respectively:*

$$\mathbb{E}^*[H|Z, P] = \delta_0 + \delta_z Z + \delta_p P. \tag{3.17}$$

*Proof.* See Appendix A.2 □

Intuitively, the "ratio of regression coefficients" estimator for the linear estimator is simply a special case of applying the general theorem, where the general expectation operator $\mathbb{E}[H|P, Z]$ is replaced with the linear expectation operator $\mathbb{E}^*[H|P, Z]$. Nonetheless, it is important to note that the price of the amenity in the general case is defined locally for every triplet $(p, z, \phi)$, whereas the price of the amenity in the linear case is defined globally across the entire domain due to the linear single index assumption.

## 3.3 Estimation Procedure

Building on the results of Theorem 3.5 and Theorem 3.10, one can formalize an estimator for both the general case and for when the single latent index enters linearly.

### 3.3.1 Estimation Procedure for the General Case

The main Theorem 3.5 above not only provides us with the appropriate assumptions for when the amenity price is identified with the proxy method, but it also suggests a natural method to obtain an estimate of the price in practice. Concretely, to estimate the price of the amenity in the neighborhood, one can first divide the observations into quantiles $Q_i^z$ of $z$ (indexed by $i$) and quantiles $Q_j^p$ of $p$ (indexed by $j$ ). Next, we can compute the predicted proxy by taking the average of the proxy variable for all the observations of $(p, z)$ in each of the cell among the Cartesian-product of quantiles: $\hat{H}_{Q_j^p, Q_i^z} = \mathbb{E}[H|p \in Q_j^p, z \in Q_i^z]$.[11] Then, one may approximate the partial derivatives by their discrete

[11] Our approach of binning by quantiles is essentially the three-dimensional analogue of the binscatter non-parametric approach made popular by work such as Chetty, Friedman and Rockoff (2014) and formalized by Stepner (2013).

analogues:

$$\Delta\hat{H}/\Delta Z\Big|_{(p,z)} = (\hat{H}_{Q^p_j,Q^z_{i+1}} - \hat{H}_{Q^p_j,Q^z_i})/(\hat{z}_{Q^z_{i+1}} - \hat{z}_{Q^z_i}) \tag{3.18}$$

$$\Delta\hat{H}/\Delta P\Big|_{(p,z)} = (\hat{H}_{Q^p_{j+1},Q^z_i} - \hat{H}_{Q^p_j,Q^z_i})/(\hat{p}_{Q^p_{j+1}} - \hat{p}_{Q^p_j}) \tag{3.19}$$

where $(p \in Q^p_j, z \in Q^z_i)$ and $\hat{p}$ and $\hat{z}$ denote the means in the corresponding quantile. Hence, by Theorem 3.5, the price of the amenity can be estimated as

$$-\frac{\Delta\hat{H}/\Delta Z}{\Delta\hat{H}/\Delta P}\Big|_{(p,z)}. \tag{3.20}$$

### 3.3.2 Estimation Procedure for the Linear Case

By Theorem 3.10, the price of the amenity in the linear model is estimated by the ratio of two regression coefficients

$$-\frac{\delta_z}{\delta_p} \tag{3.21}$$

where $\delta_z$ and $\delta_p$ denote the coefficients from a linear regression of $H$ on $P$ and $Z$:

$$\mathbb{E}^*[H|P,Z] = \delta_0 + \delta_p P + \delta_z Z. \tag{3.22}$$

It is also interesting to note that a numerically equivalent estimator is to view the linear regression of $H$ on $P$ and $Z$ in Equation (3.22) above as a "first stage". The predicted values from the first stage, $\hat{H}$, are then included as a control in the original hedonic regression as follows:

$$\mathbb{E}^*[P|Z,\hat{H}]. \tag{3.23}$$

Intuitively, the reason that the "second-stage regression" in (3.23) is equivalent to the "ratio-of-coefficients" method in recovering the correct price could be viewed through the lens of the implicit function theorem: our main results in Theorem 3.10 shows that we can use the tangent vector of the quality level set as the price of the amenity, which motivates the ratio-of-coefficients specification. On the other hand, controlling for $\hat{H}$ is akin to estimating the tangent vector by directly controlling for being on the same quality level set through the inclusion of $\hat{H}$ as part of the control.

## 3.4 A Discussion of the Assumptions

Given that the applicability of the proxy-based method depends on the plausibility of various assumptions, in this section, we discuss some of the intuitions behind the key assumptions. We argue that the assumptions used in this paper are consistent with the body of academic scholarship on modeling housing markets as vertically and horizontally differentiated goods as well as the intuition about the way that home-buyers make choices subject to constraints.

### 3.4.1 The Single Index Assumption

Underlying the proxy approach is a structural assumption that there is a *single* latent index of vertical quality for which the researcher wants to control. Is that a reasonable assumption? What kind of data generating process might produce a market that satisfies the Single Index Assumption?

To micro-found this assumption is to consider a housing market where the utility-relevant features of each home includes a vector of horizontal amenities $Z$ and a single measure of vertical quality $\Phi$, in addition to price, $U_i(P_j, Z_j, \Phi_j)$.

Although this assumption does little to restrict the classes of utility functions one can consider, for a concrete example, one may consider a typical model of demand for differentiated goods where the utility is derived from both amenity $Z$ and quality $\Phi$:

$$U_i(P_j, Z_j, \Phi_j) = \alpha_i P_j + \beta_i Z_j + \Phi_j + \epsilon_{i,j}, \tag{3.24}$$

where $i$ indexes households, $j$ indexes housing units, $\beta_j \neq 0$, and $\epsilon_{i,j}$ are i.i.d. Then, as long as the supply side satisfies basic regularity conditions (such as amenities are not in infinite supply at zero price), then these utility relevant amenity $Z$ and quality $\Phi$ will enter into prices. In addition, to the extent that search and matching frictions may also contribute to both the price paid and the amenity chosen, without loss of generality, our model simply recasts it to be part of the unobserved quality.

As such, the Single-Index-Assumption does not necessarily impose any restriction on the supply-side of the market. To make the DGP concrete, in Appendix Section B and Section D, we provide simulations of the market clearing process with either competitive supply or fixed supply amenities. In the first case, we assume a model of the market with Cobb-Douglas utility and competitive supply of amenities. In the second case, we assume a model of the market where amenities and qualities are exogenous determined. We show that in both cases the Single Index Assumption are satisfied.

In addition, a useful intuition from the first model is that the expenditure on housing amenity and quality will be proportional to household's budgets. To the extent that the preference for quality is vertical (i.e., same across all households) as opposed to horizontal (i.e., heterogeneous across all households), there exists a one-to-one correspondence between household budget and housing quality purchased. As such, in empirical work, in addition to direct measures of neighborhood quality on the supply side, factors on the demand side that lead to better or worse household budgets as diverse as income, earnings potential, education, parental resources, personality traits could also be plausibly considered as proxies for housing quality.

Of course, in general, this strict one-to-one correspondence between household budget and housing quality need not hold. That said, in an alternative DGP with fixed supply of amenities, in Appendix Section D, we show that the correspondence between household budget and housing quality naturally emerges when one's preference for housing quality is relatively strong compared to their preference for amenities.

### 3.4.2 The Conditional Independence Assumption

While conceptually distinct from IVs, econometrically, one may interpret the Conditional Independence assumption as a form of exclusion restriction that aids identification.

Compared with the canonical IV methods which attempt to find an instrument that shifts with endogenous treatment but otherwise uncorrelated with the unobservable error, the proxy method attempts to find an "instrument" that shifts with the unobserved quality but is otherwise uncorrelated with the observable amenity. In some sense, the proxy method shifts the burden of identification from finding the most comprehensive set of controls to finding an appropriate proxy, much the same way that traditional IV methods shift the burden of identification from controlling for all correlated residuals towards finding an appropriate instrument that satisfies the exclusion restriction.

The main difference is that traditional IV is there to deal with an endogenous *observable* variable, whereas the proxy method is to address the presence of an endogenous *unobservable* variable.

## 3.5 Additional Properties of the Proxy Estimator

### 3.5.1 Extensions to Multiple Amenities

The estimation strategy can be readily extended to allow for multiple amenities. To examine market trade-offs between $P$ and a set of amenities $Z_1, \ldots, Z_K$, the conditional independence assumption must hold for each amenity:

$$\forall Z_k \in \{Z_1, \ldots, Z_K\} : \, H \perp Z_k \,|\, \Phi. \tag{3.25}$$

Intuitively, the proxy must contain no further information about an individual's choices over any given amenity condition on their choice of quality $\Phi$.

Estimation of the multiple-amenity model, again, follows either the ratio-of-coefficients approach or the predicted-values approach as described in Section 3.3.2. In particular, we can estimate the amenity price for each amenity as the ratio of coefficients as follows:

$$\mathbb{E}^*[H|P, Z_1, ..., Z_K] = \delta_0 + \delta_P P + \delta_1 Z_1 + ... + \delta_K Z_K \tag{3.26}$$

$$\forall k = 1, \ldots, K : \hat{\beta}_k = -\frac{\delta_k}{\delta_P} \tag{3.27}$$

Equivalently, the amenity price can be estimated by performing a second stage regression of price on amenities controlling for the predicted proxy

$$\mathbb{E}^*[P|Z_1, ..., Z_K, \hat{H}] = \beta_0 + \beta_1 Z_1 + ... + \beta_K Z_K + \beta_H \hat{H} \tag{3.28}$$

where the predicted proxy is analogously defined as $\hat{H}(P, Z_1, \ldots, Z_K) \equiv \mathbb{E}^*[H|P, Z_1, ..., Z_K]$.

A natural concern that arises here is how to address the issue of missing amenities. In other words, what if there are multiple utility-relevant amenities, but the researchers are only able to observe a subset of them? First, note that the issue of missing amenities is common to all observational methods, affecting both OLS estimates and proxy-based estimates. That is, the presence of missing amenities will bias both OLS and proxy estimates.

It is worth noting whether the coefficient is "biased" or not depends on what the researcher is interested in. Typically, for OLS estimates, even with perfect controls, missing amenities still lead to a "biased" estimate of the amenity price depending on the correlation between observed and unobserved amenities. To the extent that two amenities typically come bundled, then the estimates from these observational methods correctly encapsulates the notion in terms of how much a household has to pay to obtain it in the current market, because these amenities cannot be sold piece-by-piece. However, the estimates is biased from the point of view of an external intervention where one amenity is added without affecting the other. To provide a more concrete example, suppose school quality and library quality are perfectly correlated in the data, but library quality is missing from the econometric specification. As such, the estimated coefficients for both OLS and proxy method will estimate the price one has to pay for both school and library quality that come as a bundle. We believe that it may still be of great interest to policy makers, especially when it comes to the issue of measuring housing inequality, as one cannot "buy" better schools without "buying" the other amenities that typically come as a bundle but otherwise unobserved to the econometrician.

Regardless, we also derive a "bias" formula for the proxy estimate, which is distinct from the OLS bias formula, detailed in Appendix Section C. In the case of two amenities $Z_1, Z_2$ where $Z_2$ is missing, we show that the extent of bias is affected by the relationship between the unobserved amenity and observed quantities:

$$\hat{\beta}_1^{\text{proxy}} = -\frac{\delta_1 + \delta_2\,\psi_1}{\delta_p + \delta_2\,\psi_p} \tag{3.29}$$

where $\psi_1$ and $\psi_p$ is defined by the linear regression of $Z_2$ on $P$ and $Z_1$, $\mathbb{E}^*[Z_2|P, Z_1] = \psi_0 + \psi_1 Z_1 + \psi_p P$. Moreover, we also show in Appendix Section C that, even in the presence of missing amenities, the proxy estimator remains robust to unobserved quality as long as a viable proxy can be found, whereas the OLS estimator can become significantly more biased due to unobserved quality.

Conceptually, the issue of missing amenities is less problematic if we believe that the set of observed amenities plus a vertical index of quality are the main drivers of housing utility, and thus approximates the Single Index Assumption reasonably well. As such, while it may seem onerous to have to observe *all* utility relevant amenities, in practice, this may be reasonably feasible. In fact, approximating all amenities as only a single index has been a popular assumption of recent literature. In the housing context, Diamond (2016) uses principal components analysis to distill amenities down to a single dimension to be used in analysis. In the labor context, Sorkin (2018) analogously treats job amenities as a single-dimensional concept. In relation to the selection-on-observables literature, our single index assumption closely mirrors the baseline specification of Altonji, Elder and Taber (2005) in assessing the impact of Catholic school attendance on student outcomes. While the single index assumption cannot hold perfectly in the presence of unobserved independently varying amenities, it seems to be a reasonable enough approximation for tractability that is implicitly or explicitly made in the contemporary literature.

### 3.5.2 Extensions to Multiple Proxies

When multiple valid proxies are thought to exist for latent quality, the set of options available to the researcher is comparable to the case when multiple instruments are available for assessing the effect of an endogenous treatment on an outcome of interest.

In our paper, we take the approach of showing how the price estimate changes when different candidate proxies are used. The extent to which the estimates are similar means that either all proxies satisfy the identification condition, or else all are biased in a similar way, and the likelihood of rejecting this hypothesis can be quantified with a $J$-statistic.

A complementary approach to leveraging over-identification would be to combine the information from all proxies into a single more precise estimate with an optimal weighting given by a generalized methods of moments (GMM) estimator identified by our linearized conditional independence assumptions.

### 3.5.3 Relationship with Treatment Effect Estimators

While our proxy-based method is primarily developed to estimate amenity prices in the presence of a single index of latent quality, to the extent it is a method that deals with the endogenous unobservables, we spell out in this section how it is related to methods developed to estimate treatment effects in the presence of omitted variables such as Oster (2019) and Altonji, Elder and Taber (2005).

Our key assumption of *Conditional Independence* (Assumption 3.2) can be viewed as analogous to Oster's Assumption 1 of "*equal selection*."[12] In the treatment-effect setting, Oster explains this assumption as "the unobservables and observables are equally related to the treatment." In our setting, equal selection translates to the need for the observed proxy to be unrelated to the amount of amenity chosen conditional on the true quality. We formalize this statement using conditional independence notation.

Within a context of equal selection, Oster gives the following formula for calculating the true parameter $\beta^*$:

$$\beta^* = \tilde{\beta} - [\mathring{\beta} - \tilde{\beta}]\frac{R_{max} - \tilde{R}}{\tilde{R} - \mathring{R}}$$

where $\tilde{\beta}$ and $\mathring{\beta}$ are the coefficients from regressions with and without the observed control, respectively, and likewise $\tilde{R}$ and $\mathring{R}$ are the corresponding R-squared values. The only parameter not given by the data, which must be chosen by the researcher, is $R_{max}$. In the context of treatment effects, this parameter represents the R-squared from a "hypothetical regression of the outcome on treatment and both observed and unobserved controls." Altonji, Elder and Taber (2005) set $R_{max} = 1$, although Oster (2019) points out that this assumption is unlikely to hold in real-world data due to the likelihood that the outcome contains either idiosyncratic variation or measurement error.[13] In our case, $R_{max}$ should be thought of as the R-squared from a regression of home prices on the amenity in which we hold fixed the home's hypothetical true unobserved quality. Our *Single Index Assumption* (Assumption 3.1) is equivalent to setting $R_{max} = 1$. Intuitively, we assume that homes of the same quality that have the same amenities should have the same price. Put more simply, we must believe that, if we could observe true home quality, then having data on the amenities and home prices would allow us

[12] In the notation of Oster (2019), this is given by $\delta = 1$, where $\delta$ can be termed a coefficient of proportionality.

[13] See also Goldberger's 1984 critique of reverse regression for salary discrimination. In that case, Goldberger points out that substantial idiosyncrasies in wage determination (of the type later formalized by Mortensen) necessitate that $R_{max}$ when controlling for true productivity must be below 1.

pin down the trade-off that buyers face.

# 4 Empirical Application

## 4.1 Data

### 4.1.1 Geographic Mobility

Our measure of population mobility across geographic regions comes from two sources. The first data source is the 2018-2019 county-to-county migration data from the Internal Revenue Service (IRS) Statistics of Income (SOI) Division, which is "based on year-to-year address changes reported on individual income tax returns filed with the IRS."[14] The data is publicly available and covers the entire universe of population with Forms 1040 filings. The main caveats of the data are that it does not cover migration at finer geographic level or by demographics, and it does not include non-filers. Furthermore, flows between counties with fewer than 20 returns are suppressed for confidentiality protection. To measure geographic mobility at a finer resolution, we also use the Infutor data. The Infutor data contain individual-level mobility data across the United States that records the past ten residences at the street address level for each individual. Coverage is thought to be broad, though variable over time. For this version of the paper, we focus on a subsample of the data that contains all the individuals who have ever lived in the state of California to measure migration flows between California ZIP Code Tabulation Areas (ZCTA) in the year 2019.

### 4.1.2 Home Prices

Our house price data comes from the Zillow Home Value Index (ZHVI), which reflects the typical value for homes in the 35th to 65th percentile range. We collect monthly ZHVI for single-family residences with 1,2,3,4, and 5+ bedroom at the county-level in 2019 and take the monthly average over the year.

### 4.1.3 Geographic Amenities

The main amenity data we use for our analysis is the Air Quality Index (AQI) from EPA's Air Quality System, which is an aggregate measure of five pollutant levels. We collect daily county-level AQI data in 2019 from EPA Air Data and use the median across all days in the year. Since the AQI data is only available for 1063 counties out of the total 3033 counties, we interpolate the AQI using the inverse distance weighting method commonly used in the economics literature on air pollution (Neidell

[14] https://www.irs.gov/pub/irs-soi/1819inpublicmigdoc.pdf

2004; Currie and Neidell 2005).[15] For these counties, the interpolated AQI is calculated as the average of AQI readings from other counties weighted by the inverse of the squared distance between the counties.

#### 4.1.4 Other Data Sources

Finally, we complement our data with county-level socio-demographics from the 2019 American Community Survery (ACS) 5-year data. This includes household median income, percent high school educated or above, and percent college educated or above.

### 4.2 Geographic PageRank (GPR)

In this section, we propose and construct a novel measure of locational desirability called "Geographic PageRank (GPR)," which we later use as a proxy variable for the purpose of valuing geographic amenities.

#### 4.2.1 The Construction of Geographic PageRank

The key idea behind "Geographic PageRank" is to capture a ranking of geographic locations using a recursive logic based on migration: by revealed preference, when households move, they must on average move to a "better" place than their previous place, and places that draw people from "better" places are considered "best" places. As such, we use a matrix of migration to arrive at a ranking of places based on such recursive definition of common desirability. To implement the idea of a recursive-based ranking using flows, we use the famous PageRank algorithm introduced into labor economics by Sorkin (2018), which was invented by Larry Page (2001) and had been the foundational piece underlying much of Google's search rankings. Our approach to adapting PageRank from the labor to housing literature also compliments that of Fogel (2021), who was to our knowledge the first to implement PageRank at the geographic level, but his focus is to use it to show a null effect of local labor demand shocks on city value as measured by PageRank.

At a conceptual level, our proposed Geographic PageRank algorithm ranks a location by its share in the stationary distribution of household locations in accordance to the migration matrix. Specifically, let the migration matrix $M_{i,j}$ represent the fraction of population migrating from location $j$ to $i$. In other words, the matrix specifies the destination locations for everyone from location $j$, and $M_{i,j}$

[15] We are exploring other measures of air pollution, including those derived from satellite data used by Sullivan and Krupnick (2018). Results thus far are qualitatively robust to restricting the sample to non-interpolated counties.

represents a transition matrix (in a column format) such that every column sums to one

$$\forall j : \sum_i M_{i,j} = 1. \tag{4.1}$$

In its simplest form, the PageRank algorithm computes the column vector $v$ such that

$$Mv = v. \tag{4.2}$$

Equivalently, $v$ represents the eigenvector associated with the transition matrix $M$.[16] In this sense, the resulting eigenvector $v$ represents the stationary distribution of the population if they migrated according to the transition matrix $M$ indefinitely: namely, the transition matrix $M$ on the distribution $v$ recovers $v$. As a last step, we obtain the ordinal "rank" of location $i$ based on the rankings of the value of each element $v_i$.

To provide some intuition behind the PageRank algorithm, consider the following stylized example. In Panel (A) of Figure 4, there are three locations A, B, and C. Assume that they are equal-sized cities and we observe that 10% of C's residents migrated to A and another 10% migrated to B. Performing the PageRank algorithm on the resulting migration matrix $M$ results an equal ranking between A and B, which are both ranked above location C.

In Panel (B) of Figure 4, we also consider three locations A, B, and C. However, in this case, we find 5% of C's residents migrated to A, and 15% migrated to B. Moreover, we also find 5% of B's residents also migrated to A. In this case, the PageRank algorithm ranks A above B, and then above C. Notably, the PageRank algorithm contains more information than simple counts of net-migration: even though both location A and B see a net-in migration of 10% in both examples, the PageRank algorithm ranks A above B in the second example, as A is drawing residents from B, which is in turn more desirable than C.

Beyond its widespread application in computer science, Sorkin (2018) is the first to use the PageRank algorithm in economics and it is applied to ranking a set of U.S. firms in the labor market context. Indeed, Sorkin (2018) proceeds to use the ranking of firms as a way to measure how much wage dispersion is due to job amenities. We highlight that there is a natural parallel between the challenges of the empirical valuation of job amenities and the empirical valuation of neighborhood amenities, even though there has not been much overlap between the two literature in recent works. As such,

[16] In practice, the PageRank algorithm also allows for a damping factor $d$ such that every period a small fraction $1 - d$ from each node will be randomly redistributed to all other nodes to ensure numerical stability. We find that the ordinal rank of the nodes are not sensitive to the choice of the damping factor.

our work fits into this important theoretical and methodological gap.

### 4.2.2 A Ranking of U.S. Counties

To implement the Geographic PageRank algorithm at the national level, we use the outflow version of the IRS 2018-2019 county-to-county Migration Data, which records "the number of residents leaving a State or county and where they went." Table 1 shows the summary statistics for the IRS migration data. 2,824 out of the total 3,033 counties has more than 20 migration flows to another county. Within these counties, the mean total outflow is 2090.6 with a standard deviation of 6802.9, and the median total outflow is 348. There are 47,974 origin-destination county pairs out of all the 9,196,056 possible pairs with more than 20 migration flows. Within these county pairs, the mean migration flow is 123.1 with a standard deviation of 468.9, and the median migration flow is 41.

Next, we obtain the corresponding rankings of all U.S. counties using the Geographic PageRank algorithm described above.[17] Figure 5 illustrate the rankings on the map: green shades indicate higher-ranked (i.e., more desirable) counties and red shades indicate lower-ranked (i.e., less desirable) counties. We generally find higher-ranked counties along the coastal regions. Table 2 shows the top 20 counties by its Geographic PageRank. The top 10 cities are Phoenix, Houston, Los Angeles, Dallas, Chicago, Fort Worth, Seattle, Austin, Minneapolis, and San Antonio. We observe that many of them are in the Sun Belt, reflecting recent migratory trends.

In general, places that see population growth are on average ranked higher. The top panel of Figure 7 shows the binscatter between Geographic PageRank and net migration, clearly indicating a positive relationship. However, there remains significant dispersion for the same level of net population change, and the average relationship is also non-linear, as shown in the scatter plot in the bottom panel of Figure 7. The remaining dispersion suggests that the recursive nature of the PageRank algorithm synthesizes more information from the full migration matrix $M$, as opposed to the simple summation operation that computes the net migration alone.

### 4.2.3 More Granular Rankings from Infutor Data

Although considerably noisier at present, our dataset of residential mobility from Infutor also allows us to rank finer-grained geographies. Figure 6 shows a map of Geographic PageRank for California's ZCTA's, based on moves recorded by Infutor in 2019. In this version of the paper, we do not show additional results leveraging ZCTA-level moves due to the increased noise associated with small sample sizes at finer-grained geographies. However, we are in the process of cleaning and incorporating

[17] Our specific implementation assumes a damping factor of $d = 0.85$, which corresponds roughly to the average stay rate.

additional years and geographic coverage of the data, which will allow for more reliable estimates to study more granular mobility patterns.

#### 4.2.4 Limitations

Two key limitations of the application of PageRank to geographies are worth noting. First, there are reasons to suspect that migratory tendencies might not always be equatable with place quality. If people's preferences for amenities systematically change as they age, then migratory flows may be less indicative of general place quality, and more indicative of the presence of particular amenities.[18] In preliminary results not shown here, we have found evidence that PageRank tends to over-value warm climates more than other candidate proxies we have examined (such as income), which gives credence to the concern that migratory flows might not necessarily point in the direction of quality, but rather in the direction of amenities that become increasingly desirable as one ages. This limitation must be grappled with if one wishes to apply Geographic PageRank to pricing weather as an amenity, which Glaeser, Kolko and Saiz (2001) find to be among the most important amenities in driving home price differentials. For this reason, in future work, we will use the Infutor data to investigate the extent to which rankings of places are stable or change over individuals' life cycles by computing ranks based on specific age cohorts over time.

The second limitation of the PageRank algorithm when applied to geographies is that it is not invariant to aggregation of geographic units. To be more concrete, for example, New York City is divided into five counties (i.e., the five boroughs) in the data with Manhattan ranked at the 21st place. However, if we created a synthetic "New York City County" that includes all the boroughs, it would have ranked at the 5th place. One may take the problem of aggregation to its logical extreme by aggregating more and more counties to generate an artificially top-ranked "super" county. Therefore, in order to obtain a meaningful interpretation of the rankings, we need to start with locations that we believe are somewhat comparable in their geographic scope.

### 4.3 Results

In this section, we illustrate that the proxy method can be used to obtain sensible estimates of the price of air pollution. We demonstrate how it works in both the general non-linear case and the more specific linear case. We show that the proxy-based estimates are able to address the "wrong-signed" problems that seem to be pervasive in the estimation of similar environmental amenities. We also find that the estimates are in-line with other existing reduced-form estimates using natural experiments.

[18] In a related vein, as described by Sorkin (2018), correlated negative shocks (e.g., an economic recession) could lead migratory flows to tend toward worse rather than better places.

First, we obtain a correctly-signed negative price for air pollution in the general setting without assuming linearity of the index mapping function, as shown in Figure 8. Specifically, the estimation procedure goes as follows: First, we divide both price and amenities into quantiles, and in this case, we have divided them each into $N = 11$ equally sized quantiles. Then, the Cartesian product of each price quantile and each air pollution quantile generates a total of $N^2$ cells. For each cell, we collect all households who belong to that cell and compute the average Geographic PageRank of members of the cell, which captures $\mathbb{E}[H|P, Z]$, namely, the predicted proxy variable conditional on price and amenity. Hence, Figure 8 is the direct empirical analogue of our illustration for identification as shown previously in Figure 3.

The pattern in Figure 8 indicates a trade-off between home price and air quality. It shows a clear increase in the predicted Geographic PageRank in the diagonal direction: the best-ranked places (i.e. dark green) are in the upper-right hand corner and the worst-ranked places (i.e. dark red) are in the bottom-left corner. As such, the upward-sloping blue line in Figure 8 indicates the direction of expanding choice sets, whereas the perpendicular direction (i.e., the dark green line) shows the level sets: the set of cells all with the same average PageRank but with varying combinations of price and amenity. In other words, the direction perpendicular to the direction of choice-set expansion shows the implied trade-off between home price and clean air: places with cleaner air are priced higher than places with more polluted air, conditioning on being in the same housing quality segment, which are proxied by the predicted Geographic PageRank.

Next, if we take on the lens of linearity, then we can use the "ratio-of-coefficients" method to obtain a single coefficient estimate of the amenity price. Recall that linearity imposes the additional restriction that the index mapping function $\phi(p, z)$ must be linear in both price $p$ and amenity $z$. To visually benchmark the magnitude of our estimate in a linear model, Figure 9 compares the estimated price elasticity with known estimates in the literature, specifically, with Chay and Greenstone (2005). The shaded region indicates plausible point estimates of the housing price elasticity with respect to total suspended particulates (TSPs), which ranged from -0.20 to -0.35. The first three dots in Figure 9 shows that the conventional OLS or OLS with controls appear to generate wrong-signed and downward-biased estimates. The last dot in Figure 9 shows a significantly negative elasticity of -.40 estimated by our proxy-based approach, and the confidence interval overlaps with the range of point estimates in Chay and Greenstone (2005). Though it should not necessarily be expected ex ante that they are the same given the nuances in how price is conceptualized (discussed in Section 3.5.1), the fact that our estimate of the price of air pollution is indistinguishable from the range of estimates in the quasi-experimental literature lends some support to the credibility of our approach.

Building on the visual evidence of Figure 9 , Table 3 shows the estimated home price elasticity with respect to air pollution for different candidate proxy variables and estimation strategies. We find that using either income or PageRank as a proxy—but not as a control—we get an intuitively signed price for air pollution, indicating that buyers face higher home prices to live in cleaner areas. Column (1) shows the naive OLS estimates, which reproduces the wrong-signed problem, indicating a positive price for air pollution. Column (2) shows that the conventional method of adding average household income as an additional control reduces the amount of bias, but the coefficient still remains positive. In column (3), instead of using income as a control, we estimate the price elasticity using income as a proxy. We find that, using household income as a proxy, it produces a correctly-signed negative price for air pollution. Next, instead of income, we can use the county-level Geographic PageRank either as a control or proxy; column (4) shows that controlling for GPR also has the effect of reducing the OLS coefficient, but still exhibits a positive sign, whereas column (5) shows our main result. The estimated price elasticity using GPR proxy produces a correctly-signed negative price elasticity.

# 5 Conclusion

In this paper, we develop a novel proxy-based method that allows us to price a wide range of amenity tradeoffs in the housing market in the midst of unobserved quality. Our estimator can be applied to commonly used observational datasets, and proves that the typical confounder of market segmentation by vertical quality that is to blame for classic “wrong signed” results should no longer be viewed as endemic to observational approaches. These tradeoffs are identified under a notably lax set of assumptions, central to which is that the researcher has available to her a rough “proxy” for quality.

To investigate whether such a viable “proxy” for quality might exist, we construct a revealed-preference measure of desirability—“Geographic PageRank”—and illustrate its usefulness in pricing the salient amenity of air quality. Our proxy for quality recursively ranks places as more desirable when they draw in more migration, with particular weights placed on migration from other highly-ranked places. When we use our revealed-preference ranking of places as a proxy for quality, we estimate that American home-buyers on average must pay 4% more to achieve a 10% improvement in air quality. Unlike virtually any other estimate estimate one could have obtained from observational data—which tend to be oppositely signed— our estimate is similar to the findings of quasi-experimental approaches to measuring home price elasticities. This lends credibility to our proxy approach, and promise that the proxy approach can be used more widely in economics.

Having described and validated our approach, we next aim to connect our price estimates to policy. How do costly amenity tradeoffs contribute to—or mitigate—housing inequities by demographic groups such as race? Answering this question is key to motivating appropriate government intervention in the provision of local public goods. Though our analysis of air quality is instructive, we also plan to incorporate other amenities such as causal place effects on children estimated by by Chetty and Hendren (2018*a*,*b*), which will allow us to shine light on the extent to which people from already-disadvantaged social groups are "priced out" of living in higher-opportunity areas.

# References

**Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber.** 2005. “An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling.” *The Journal of Human Resources*, 40(4): 791–821. Publisher: [University of Wisconsin Press, Board of Regents of the University of Wisconsin System].

**Bayer, Patrick, Fernando Ferreira, and Robert McMillan.** 2007. “A Unified Framework for Measuring Preferences for Schools and Neighborhoods.” *Journal of Political Economy*, 115(4): 588–638. Publisher: The University of Chicago Press.

**Bell, Alex.** 2022. “Job Amenities and Earnings Inequality.” Available at SSRN: http://dx.doi.org/10.2139/ssrn.4173522.

**Black, S. E.** 1999. “Do Better Schools Matter? Parental Valuation of Elementary Education.” *The Quarterly Journal of Economics*, 114(2): 577–599.

**Brown, Charles.** 1980. “Equalizing Differences in the Labor Market.” *The Quarterly Journal of Economics*, 94(1): 113–134. Publisher: Oxford University Press.

**Burbano, Vanessa, Olle Folke, Stephan Meier, and Johanna Rickne.** 2022. “The Gender Gap in Meaningful Work: Explanations and Implications.” *CEPR Discussion Papers.* Number: 17634 Publisher: C.E.P.R. Discussion Papers.

**Chay, Kenneth Y., and Michael Greenstone.** 2005. “Does Air Quality Matter? Evidence from the Housing Market.” *Journal of Political Economy*, 113(2): 376–424. Publisher: The University of Chicago Press.

**Chetty, Raj, and Nathaniel Hendren.** 2018*a*. “The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects*.” *The Quarterly Journal of Economics*, 133(3): 1107–1162.

**Chetty, Raj, and Nathaniel Hendren.** 2018*b*. “The Impacts of Neighborhoods on Intergenerational Mobility II: County-Level Estimates*.” *The Quarterly Journal of Economics*, 133(3): 1163–1228.

**Chetty, Raj, John N. Friedman, and Jonah E. Rockoff.** 2014. “Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood.” *American Economic Review*, 104(9): 2633–2679.

**Currie, Janet, and Matthew Neidell.** 2005. “Air Pollution and Infant Health: What Can We Learn from California’s Recent Experience?*.” *The Quarterly Journal of Economics*, 120(3): 1003–1030.

**Currie, Janet, Lucas Davis, Michael Greenstone, and Reed Walker.** 2015. “Environmental Health Risks and Housing Values: Evidence from 1,600 Toxic Plant Openings and Closings.” *American Economic Review*, 105(2): 678–709.

**Diamond, Rebecca.** 2016. “The Determinants and Welfare Implications of US Workers’ Diverging Location Choices by Skill: 1980-2000.” *American Economic Review*, 106(3): 479–524.

**Epple, Dennis, and Holger Sieg.** 1999. “Estimating Equilibrium Models of Local Jurisdictions.” *Journal of Political Economy*, 107(4): 645–681. Publisher: The University of Chicago Press.

**Epple, Dennis, Luis Quintero, and Holger Sieg.** 2020. “A New Approach to Estimating Equilibrium Models for Metropolitan Housing Markets.” *Journal of Political Economy*, 128(3): 948–983. Publisher: The University of Chicago Press.

**Fogel, Jamie.** 2021. “Valuing American Cities: A Revealed Preference Approach.”

**Folke, Olle, and Johanna Rickne.** 2022. “Sexual Harassment and Gender Inequality in the Labor Market*.” *The Quarterly Journal of Economics*, 137(4): 2163–2212.

**Glaeser, Edward L., Jed Kolko, and Albert Saiz.** 2001. “Consumer city.” *Journal of Economic Geography*, 1(1): 27–50.

**Kniesner, Thomas J., W. Kip Viscusi, Christopher Woock, and James P. Ziliak.** 2012. “The Value of a Statistical Life: Evidence from Panel Data.” *The Review of Economics and Statistics*, 94(1): 74–87.

**Lucas, Robert E. B.** 1977. “Hedonic Wage Equations and Psychic Wages in the Returns to Schooling.” *The American Economic Review*, 67(4): 549–558. Publisher: American Economic Association.

**Mas, Alexandre, and Amanda Pallais.** 2017. “Valuing Alternative Work Arrangements.” *American Economic Review*, 107(12): 3722–3759.

**Neidell, Matthew J.** 2004. “Air pollution, health, and socio-economic status: the effect of outdoor air quality on childhood asthma.” *Journal of Health Economics*, 23(6): 1209–1236.

**Oster, Emily.** 2019. “Unobservable Selection and Coefficient Stability: Theory and Evidence.” *Journal of Business & Economic Statistics*, 37(2): 187–204.

**Page, Lawrence.** 2001. “Method for node ranking in a linked database.”

**Palmquist, Raymond B.** 1984. “Estimating the Demand for the Characteristics of Housing.” *The Review of Economics and Statistics*, 66(3): 394.

**Roback, Jennifer.** 1982. “Wages, Rents, and the Quality of Life.” *Journal of Political Economy*, 90(6): 1257–1278.

**Rosen, Sherwin.** 1974. “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition.” *Journal of Political Economy*, 82(1): 34–55. Publisher: The University of Chicago Press.

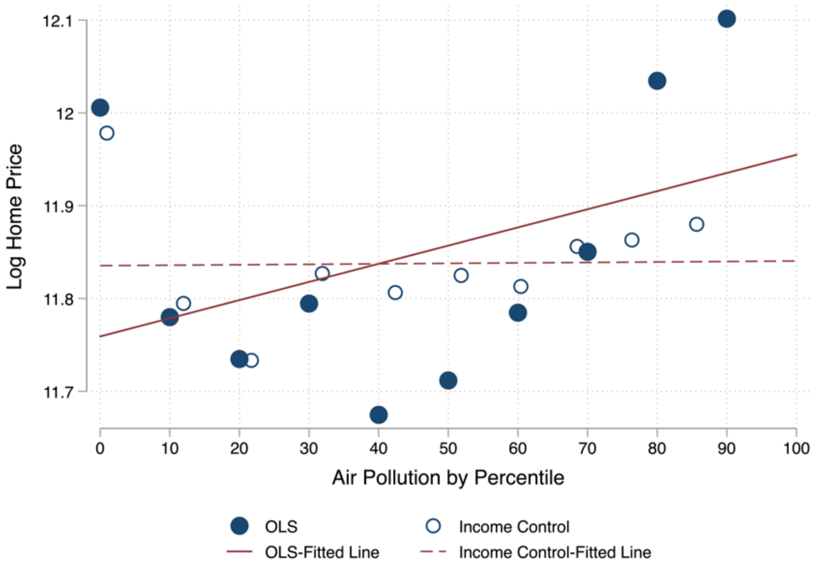
**Sorkin, Isaac.** 2018. “Ranking Firms Using Revealed Preference*.” *The Quarterly Journal of Economics*, 133(3): 1331–1393.

**Stepner, Michael.** 2013. "BINSCATTER: Stata module to generate binned scatterplots." *Statistical Software Components*. Publisher: Boston College Department of Economics.

**Sullivan, Daniel M, and Alan Krupnick.** 2018. "Using Satellite Data to Fill the Gaps in the US Air Pollution Monitoring Network."

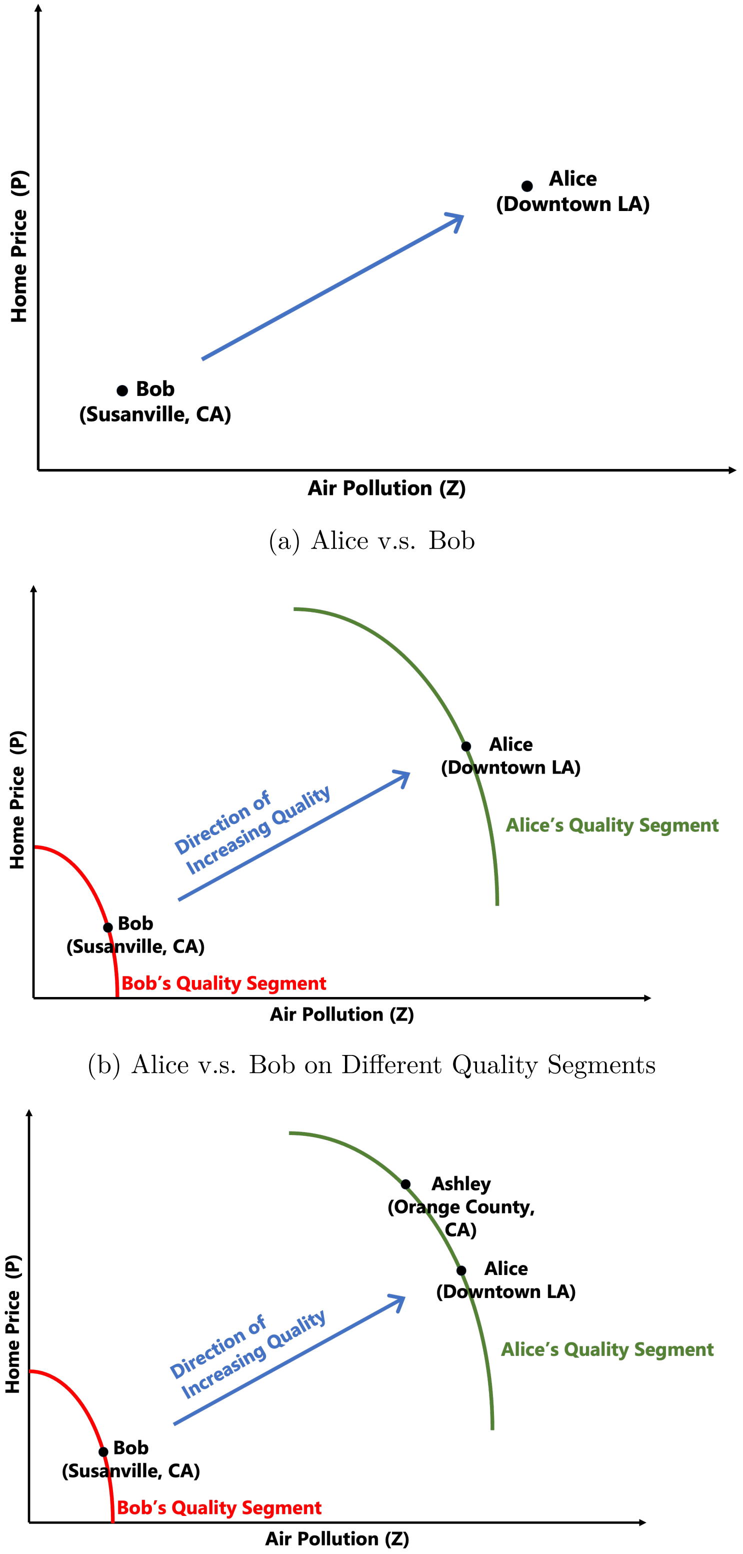
# 6 Figures

Figure 1: Illustration of the “Perverse Sign” Problem



*Notes*: The dark blue dots plot the binscatter with 2019 county-level deciles of median air quality index (aqi) on the x-axis and 2019 county-level log house price index (Zilow HPI for 3-bedrooms) on the y-axis. The white circled dots plot the same binscatter, but controlling for household median income from ACS. The red line plots the fitted line of the OLS regression of the home price on air quality, whereas the red dash line plots the fitted line of the same OLS regression with the addition of an income control. In both cases, the methods generates a positive price for air pollution, which is wrong-signed.

Figure 2: Illustration of the Identification Issue from Unobserved Quality in the Hedonic Model

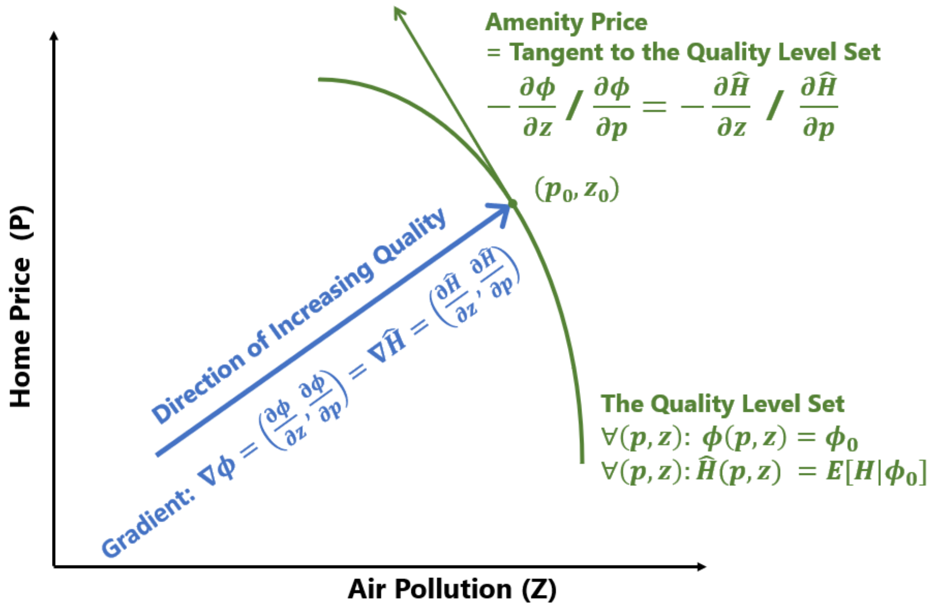


(a) Alice v.s. Bob

(b) Alice v.s. Bob on Different Quality Segments

(c) Alice and Ashley on the Same Quality Segment

Figure 3: Illustration of Identification Using the Proxy Approach

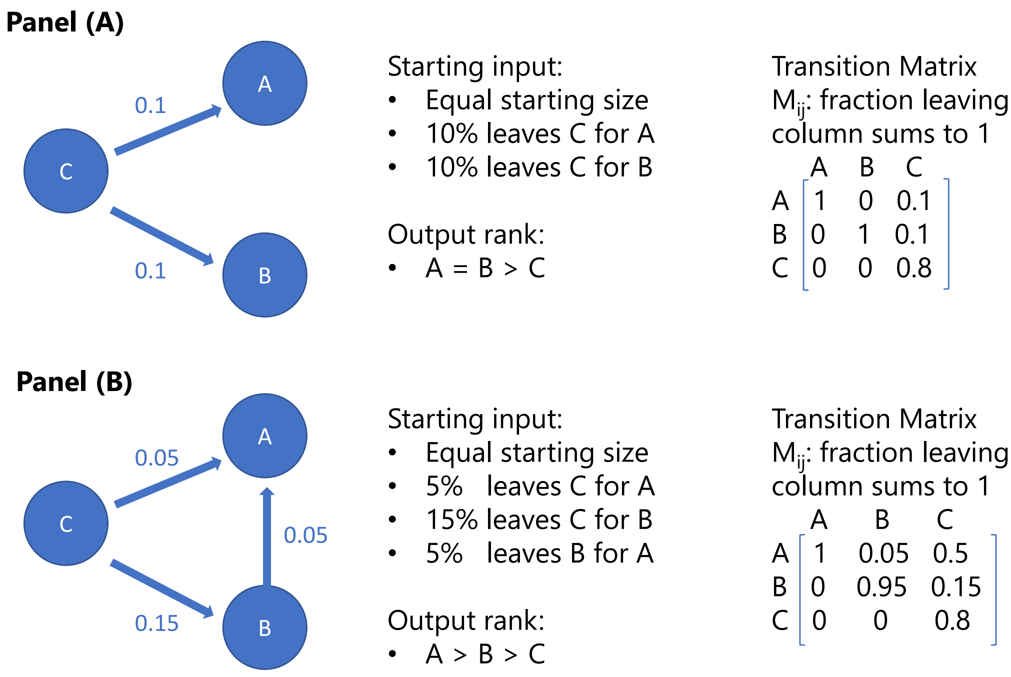

Amenity Price
= Tangent to the Quality Level Set
$-\frac{\partial \phi}{\partial z} / \frac{\partial \phi}{\partial p} = -\frac{\partial \hat{H}}{\partial z} / \frac{\partial \hat{H}}{\partial p}$
$(p_0, z_0)$
Direction of Increasing Quality
Gradient: $\nabla \phi = \left(\frac{\partial \phi}{\partial z}, \frac{\partial \phi}{\partial p}\right) = \nabla \hat{H} = \left(\frac{\partial \hat{H}}{\partial z}, \frac{\partial \hat{H}}{\partial p}\right)$
The Quality Level Set
$\forall (p, z)$: $\phi(p, z) = \phi_0$
$\forall (p, z)$: $\hat{H}(p, z) = E[H|\phi_0]$
Home Price (P)
Air Pollution (Z)

*Notes:* The figure above illustrates the intuition behind the identification strategy. By definition, the price of the amenity at any given observed $(p_0, z_0)$ is determined by how prices change in response to a change in amenity, holding its unobservable quality $\phi_0 = \phi(p_0, z_0)$ fixed. Thus, that the amenity price can be represented by the tangent direction of the level set of $\phi(p, z)$ valued at $\phi_0$, as shown above. When there exists a proxy variable $H$ that satisfies Conditional Independence, Lemma 3.4 says that the predicted proxy function satisfies $\hat{H}(p, z) = \mathbb{E}[H|\Phi = \phi(p, z)]$. This means that the level set of the function $\hat{H}(\cdot, \cdot)$ valued at $\mathbb{E}[H|\Phi = \phi_0]$ is *identical* to the level set of $\phi(\cdot, \cdot)$ when valued at $\phi_0$, thus producing the same tangent vector. Therefore, the amenity price can be estimated as the tangent vector of the level set of the predicted proxy function, namely,

$$\left.\frac{\partial p}{\partial z}\right|_{p_0, z_0} = -\frac{\partial \phi}{\partial z} \Big/ \left.\frac{\partial \phi}{\partial p}\right|_{p_0, z_0} = -\frac{\partial \hat{H}}{\partial z} \Big/ \left.\frac{\partial \hat{H}}{\partial p}\right|_{p_0, z_0}. \tag{6.1}$$

Moreover, the tangent direction is by definition perpendicular to the gradient of the predicted proxy function $\nabla \hat{H} = [\partial \hat{H}/\partial Z, \partial \hat{H}/\partial P]$, where the gradient represents the direction of increasing quality.

Figure 4: Example of PageRank Algorithm



**Panel (A)**

Starting input:

- Equal starting size
- 10% leaves C for A
- 10% leaves C for B

Output rank:

- A = B > C

Transition Matrix

$M_{ij}$: fraction leaving j for i; each column sums to 1

$$\begin{array}{cccc} & A & B & C \\ A & 1 & 0 & 0.1 \\ B & 0 & 1 & 0.1 \\ C & 0 & 0 & 0.8 \end{array}$$

**Panel (B)**

Starting input:

- Equal starting size
- 5% leaves C for A
- 15% leaves C for B
- 5% leaves B for A

Output rank:

- A > B > C

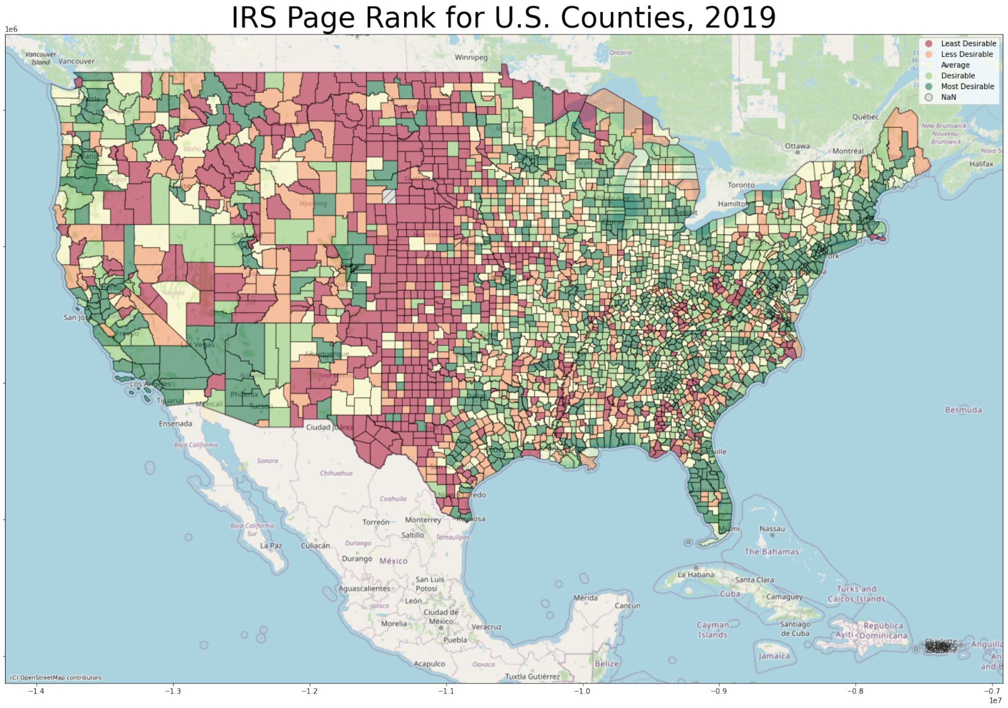
Transition Matrix

$M_{ij}$: fraction leaving j for i; each column sums to 1

$$\begin{array}{cccc} & A & B & C \\ A & 1 & 0.05 & 0.5 \\ B & 0 & 0.95 & 0.15 \\ C & 0 & 0 & 0.8 \end{array}$$

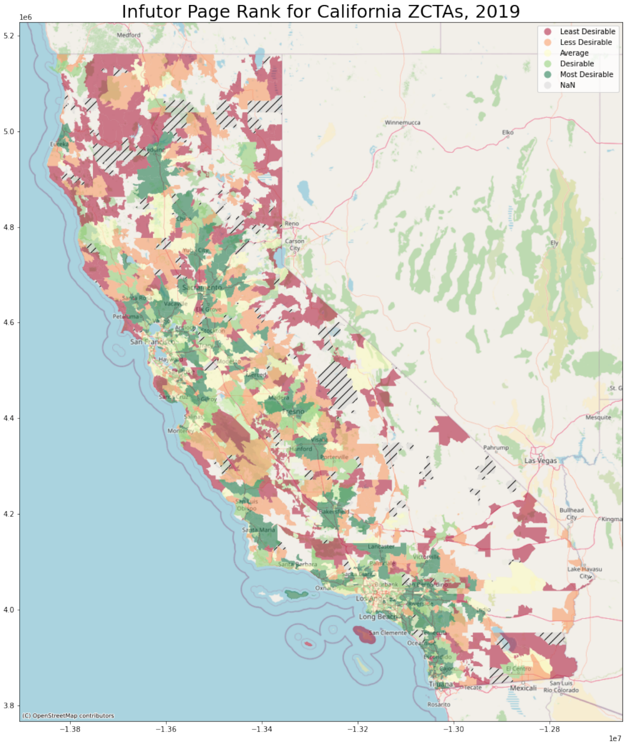
*Notes:* The figure above illustrates a simple example of the PageRank algorithm. In the top panel (A), there are three locations A, B, and C, where there is a net migration from location C to both A and B in equal proportions. The adjacency matrix $M_{i,j}$ indicates the transition probabilities in the column format, namely, $M_{i,j}$ represents the fraction departing from node $j$ for node $i$. As such, the normalization requires $\forall i : \sum_j M_{i,j} = 1$. The resulting rank is $A = B > C$. By contrast, the bottom panel (B) illustrates another example where there is not only a migration from C, but also some migration from B to A. Note that the net in-migration for both examples for location A and B are the same, namely, a net migration of 10%. However, because location A is drawing people from the more desirable location B, as opposed to C, the resulting rank has A being ranked above B, namely $A > B > C$.

Figure 5: County-level PageRank using IRS County Migration Data


IRS Page Rank for U.S. Counties, 2019

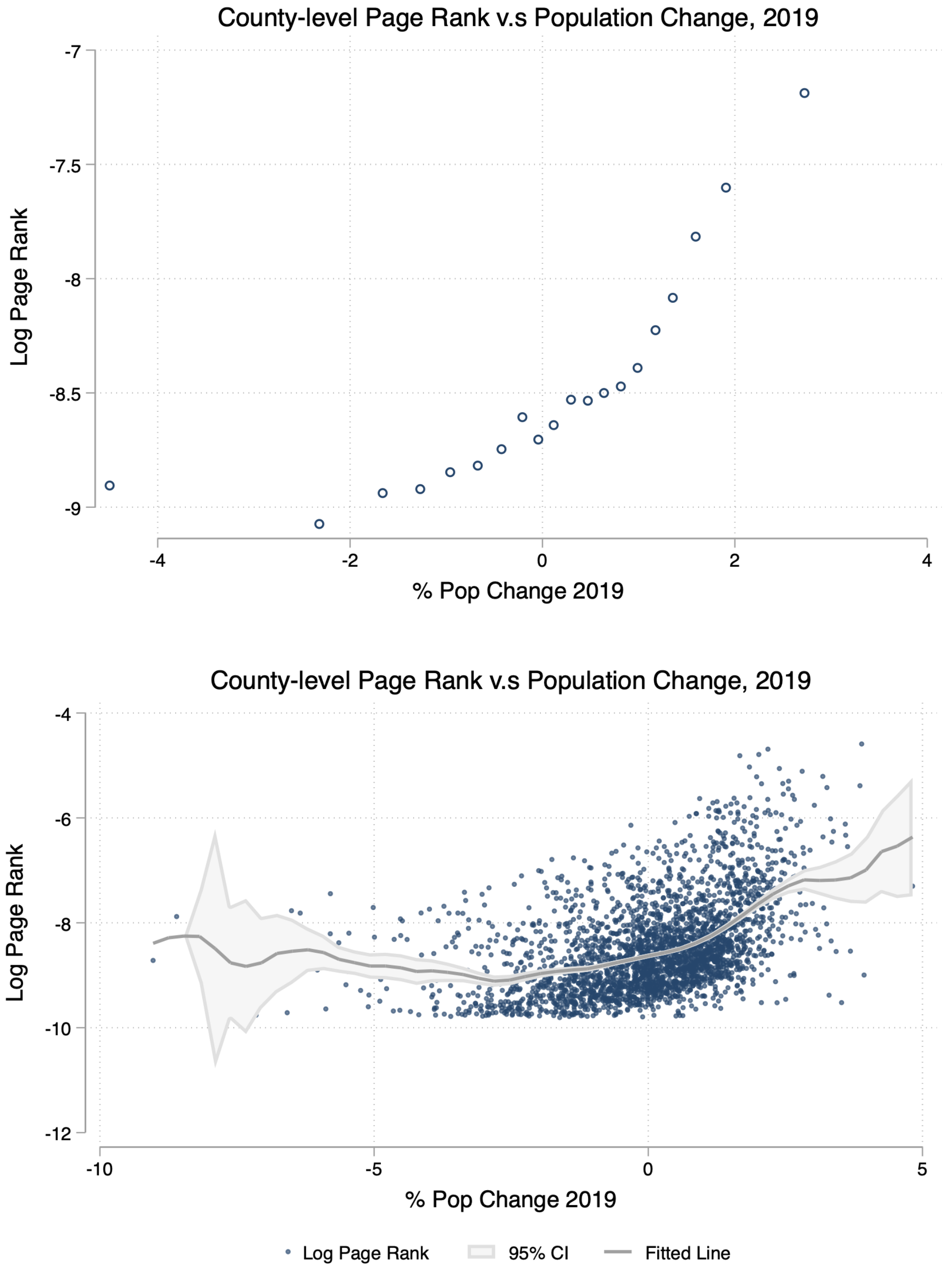
*Notes*: This map shows geographic PageRank at the county-level constructed using nationwide IRS county migration flow data in 2019. Green shades indicate higher-ranked (i.e., more desirable) counties and red shades indicate lower-ranked (i.e., less desirable) counties.

Figure 6: ZCTA-level PageRank using Infutor Migration Data in California



*Notes*: This map shows geographic PageRank at the ZCTA-level constructed using Infutor data in 2019. Green shades indicate higher-ranked (i.e., more desirable) ZCTAs and red shades indicate lower-ranked (i.e., less desirable) ZCTAs.

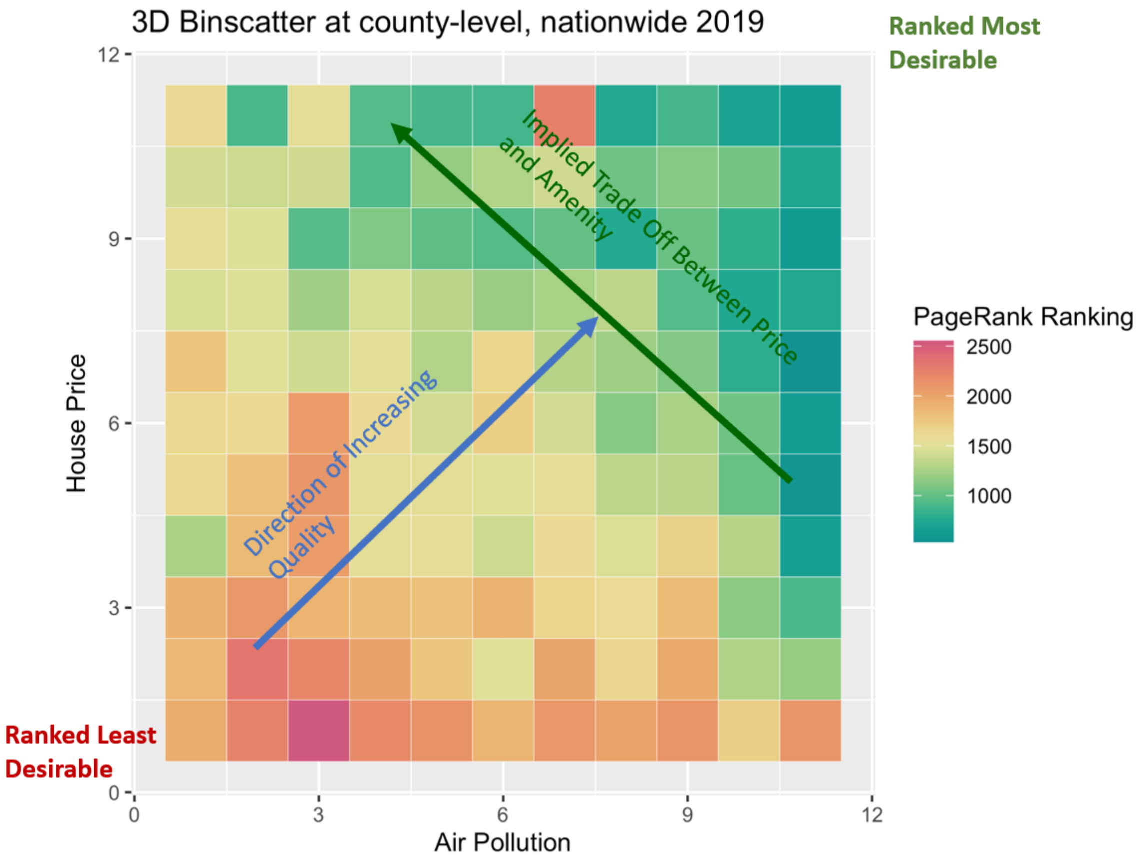
Figure 7: Geographic PageRank v.s. Net Migration


County-level Page Rank v.s Population Change, 2019

County-level Page Rank v.s Population Change, 2019

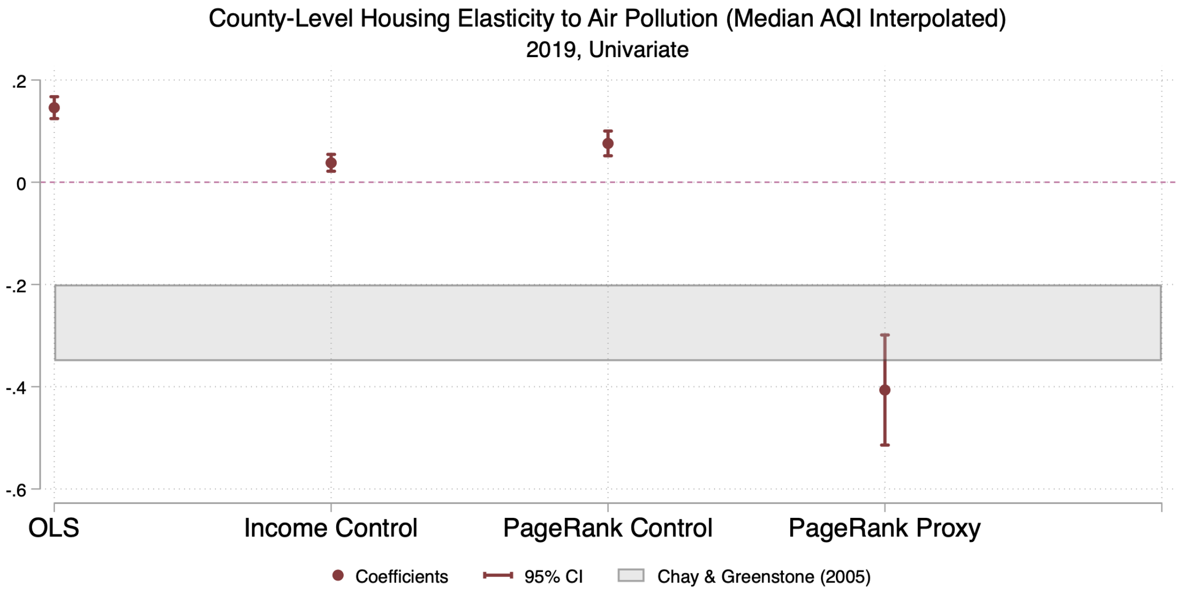
*Notes*: The binscatter (top panel) and the scatter plot (bottom panel) shows the relationship between Geographic PageRank and net migration. The y-axis represents the logarithm of the elements of the PageRank eigenvector. In general, places that see population growth are ranked higher. However, there remains significant dispersion for the same level of net migration and the average relationship is also non-linear.

Figure 8: 3D Binscatter of Predicted PageRank by Price and Air Pollution



*Notes*: This figure categorized the counties into 121 (11x11) cells. Each cell represents counties that lie within certain house price decile x air pollution decile. For example, the upper-right-most cell represents counties in top income quantile and top air pollution quantile. The color of each cell represents the median Geographic PageRank Ranking of all the counties in this cell. The PageRank algorithm ranks the most desired place as number 1. Green shades represent higher-ranked (i.e., more desirable) places and red shades represent lower-ranked (i.e., less desirable) places.

Figure 9: Pricing Air Pollution Using Proxy Approach



*Notes*: The estimated coefficients on this figures correspond to the estimated price elasticity to median AQI from regression results in Table 3.

# 7 Tables

Table 1: Summary Statistics for 2018-2019 IRS County-to-County Migration Data

| | County | | Origin-Destination Pairs |
|---|---|---|---|
| | Outflows | Inflows | Migration Flows |
| | (1) | (2) | (3) |
| Mean | 2090.62 | 2142.20 | 123.06 |
| SD | 6802.89 | 6415.80 | 468.86 |
| Min | 20 | 20 | 20 |
| 25th Percentile | 126.5 | 128.5 | 27 |
| Median | 348 | 362.5 | 41 |
| 75th Percentile | 1130.5 | 1211 | 81 |
| Max | 145204 | 111327 | 20908 |
| N | 2824 | 2756 | 47974 |

*Notes*: Migration flows between counties with less than 20 counts are suppressed in the data. Column (1) reports the county-level summary statistics for the total migration flows out of each county into other counties, and Column (2) reports the county-level summary statistics for the total migration flow into each county from other counties. Column (3) reports the summary statistics for migration flows at the level of origin county-destination pairs.

Table 2: Top Ranked Counties in the U.S. (2019)

| FIPS | Ranking | County Name | MSA Name |
|---|---|---|---|
| 4013 | 1 | Maricopa County, Arizona | Phoenix-Mesa-Scottsdale |
| 48201 | 2 | Harris County, Texas | Houston-The Woodlands-Sugar Land |
| 6037 | 3 | Los Angeles County, California | Los Angeles-Long Beach-Anaheim |
| 48113 | 4 | Dallas County, Texas | Dallas-Fort Worth-Arlington |
| 17031 | 5 | Cook County, Illinois | Chicago-Naperville-Elgin |
| 48439 | 6 | Tarrant County, Texas | Dallas-Fort Worth-Arlington |
| 53033 | 7 | King County, Washington | Seattle-Tacoma-Bellevue |
| 48453 | 8 | Travis County, Texas | Austin-Round Rock |
| 27053 | 9 | Hennepin County, Minnesota | Minneapolis-St. Paul-Bloomington |
| 48029 | 10 | Bexar County, Texas | San Antonio-New Braunfels |
| 13121 | 11 | Fulton County, Georgia | Atlanta-Sandy Springs-Roswell |
| 6073 | 12 | San Diego County, California | San Diego-Carlsbad |
| 32003 | 13 | Clark County, Nevada | Las Vegas-Henderson-Paradise |
| 48085 | 14 | Collin County, Texas | Dallas-Fort Worth-Arlington |
| 48121 | 15 | Denton County, Texas | Dallas-Fort Worth-Arlington |
| 8031 | 16 | Denver County, Colorado | Denver-Aurora-Lakewood |
| 37119 | 17 | Mecklenburg County, North Carolina | Charlotte-Concord-Gastonia |
| 12057 | 18 | Hillsborough County, Florida | Tampa-St. Petersburg-Clearwater |
| 39049 | 19 | Franklin County, Ohio | Columbus |
| 12095 | 20 | Orange County, Florida | Orlando-Kissimmee-Sanford |

*Notes*: A list of top ranked counties in the U.S. based on 2019 migration statistics using Geographic PageRank.

Table 3: Estimation of Home Price Elasticity to Median AQI

| | OLS | Income | | PageRank | |
|---|---|---|---|---|---|
| | (1) | (2)<br>Control | (3)<br>Proxy | (4)<br>Control | (5)<br>Proxy |
| Air Pollution (2019) | 0.146*** | 0.0380*** | -0.0794*** | 0.0759*** | -0.406*** |
| | (0.0109) | (0.00841) | (0.0121) | (0.0124) | (0.0550) |
| HH Median Income | | 1.380*** | 2.882*** | | |
| | | (0.0344) | (0.0710) | | |
| PageRank | | | | 0.210*** | 1.661*** |
| | | | | (0.0162) | (0.143) |
| Constant | 11.59*** | -3.809*** | -20.57*** | 13.26*** | 24.75*** |
| | (0.0213) | (0.384) | (0.790) | (0.124) | (1.164) |
| Room Fixed Effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 8322 | 8322 | 8322 | 8322 | 8322 |
| Adjusted $R^2$ | 0.410 | 0.692 | | 0.485 | |

*Notes*: This table shows the estimated price elasticity to median AQI at the county-level in 2019 using home price index data from Zillow and interpolated median AQI data from EPA. Column 1 shows the OLS result from simply regressing house price index on median AQI. Column 2 and Column 4 regress house price index controlling for household median income and geographic PageRank, respectively. Column 3 and Column 5 apply the proxy method, using household median income and geographic PageRank as the proxy, respectively. The geographic PageRank is the 2019 nationwide version constructed using IRS migration inflow data, and HH median incomes comes from ACS. Standard errors are clustered at the county level.

# A Proofs

This appendix provides the proofs for why a proxy variable that satisfies conditional independence can identify amenity prices. For the general (non-parametric) case, developed based on Section 3 of Bell (2022), we make a clear statement on how the price is formulated as the ratio of the two partial derivatives. The proof also clarifies the nature of local monotonicity required. For the special linear case, unlike Bell (2022), we provide an intuitive formulation that is based on the key properties of the linear projection residuals.

## A.1 Proof of Identification: General Case

**Theorem A.1.** ***(The price of the amenity is identified using a proxy variable.)*** *Consider a market that satisfies the single index assumption as described in Assumption 3.1. If there exists a proxy variable $H$ that satisfies conditional independence (Assumption 3.2) and strict monotonicity at $(p, z)$ (Assumption 3.3), then the price of the amenity is identified and can be consistently estimated as*

$$\left.\frac{\partial P}{\partial Z}\right|_{(p,z)} = -\left.\frac{\partial \hat{H}(p,z)}{\partial Z}\bigg/\frac{\partial \hat{H}(p,z)}{\partial P}\right|_{(p,z)} \tag{A.1}$$

*where $\phi = \phi(p, z)$, and $\hat{H}(\cdot,\cdot)$ refers to the predicted proxy function*

$$\hat{H}(p, z) = \mathbb{E}[H \mid P = p, Z = z]. \tag{A.2}$$

*Proof.* Denote $\phi = \phi(p, z)$. First, consider

$$\begin{aligned}
&\hat{H}(P = p, Z = z) && && \text{(A.3)}\\
&= \mathbb{E}[H|P = p, Z = z] && \text{by definition} && \text{(A.4)}\\
&= \mathbb{E}[H|P = p, Z = z, \Phi = \phi] && \text{by the Single Index Assumption} && \text{(A.5)}\\
&= \mathbb{E}[H|Z = z, \Phi = \phi] && \text{also by the Single Index Assumption} && \text{(A.6)}\\
&= \mathbb{E}[H|\Phi = \phi]. && \text{by the Conditional Independence Assumption} && \text{(A.7)}
\end{aligned}$$

The statement above proves Lemma 3.4. Next, take the total derivative of the LHS of Equation (A.3):

$$\frac{d}{dZ}\Big(\hat{H}(P = p, Z = z)\Big) = \left.\frac{\partial \hat{H}}{\partial P}\right|_{(p,z)} \left.\frac{\partial P}{\partial Z}\right|_{(z,\phi)} + \left.\frac{\partial \hat{H}}{\partial Z}\right|_{(p,z)} \tag{A.8}$$

Correspondingly, take the total derivative of the RHS of Equation (A.7):

$$\frac{d}{dZ}\Big(\mathbb{E}[H|\Phi = \phi]\Big) = 0 \tag{A.9}$$

Taken together, we have

$$\left.\frac{\partial P}{\partial Z}\right|_{(z,\phi)} = -\frac{\partial \hat{H}(p,z)}{\partial Z} \bigg/ \frac{\partial \hat{H}(p,z)}{\partial P}. \tag{A.10}$$

Note that the denominator

$$\frac{\partial \hat{H}(p,z)}{\partial P} = \frac{\partial \hat{H}(\phi)}{\partial \Phi} \times \frac{\partial \Phi(p,z)}{\partial P} \neq 0 \tag{A.11}$$

because the Strict Monotonicity Assumption ensures $\partial \hat{H}(\phi)/\partial \Phi \neq 0$ and the single index assumption ensure that $\partial \Phi(p,z)/\partial P \neq 0$. □

## A.2 Proof of Identification: Linear Case

**Theorem A.2.** *Consider a market that satisfies the linear single index assumption as described in Assumption 3.6 as follows*

$$P = \beta_0 + \beta_z Z + \Phi. \tag{A.12}$$

*If there exists a proxy variable $H$ that satisfies linear conditional independence (Assumption 3.7) and is relevant (Assumption 3.8), then the price of the amenity is identified and can be consistently estimated as the ratio of two regression coefficients*

$$\beta_z = -\frac{\delta_z}{\delta_p} \tag{A.13}$$

*where $\delta_z$ and $\delta_p$ are OLS regression coefficients of $H$ on $Z$ and $P$ respectively:*

$$\mathbb{E}^*[H|Z,P] = \delta_0 + \delta_z Z + \delta_p P. \tag{A.14}$$

*Proof.* First, consider the following regression of $H$ on $Z$ and $\Phi$:

$$\mathbb{E}^*[H|Z,\Phi] = \gamma_0 + \gamma_z Z + \gamma_\phi \Phi \tag{A.15}$$

Denote $\hat{H}_{Z,\Phi} = \gamma_0 + \gamma_z Z + \gamma_\phi \Phi$, then

$$H = \hat{H}_{Z,\Phi} + E_{Z,\Phi} \tag{A.16}$$

$$cov(Z, E_{Z,\Phi}) = 0 \tag{A.17}$$

Here, the coefficient on $Z$ is zero due to linear conditional independence (Assumption 3.7):

$$\gamma_z = \frac{cov(H, Z^{\perp\Phi})}{var(Z^{\perp\Phi})} = 0 \tag{A.18}$$

Thus, we have

$$\hat{H}_{Z,\Phi} = \gamma_0 + \gamma_\phi \Phi \tag{A.19}$$

Second, consider the following regression of $H$ on $P$ and $\Phi$:

$$\mathbb{E}^*[H|P,\Phi] = \theta_0 + \theta_p P + \theta_\phi \Phi \tag{A.20}$$

Denote $\hat{H}_{Z,\Phi} = \theta_0 + \theta_p P + \theta_\phi \Phi$, then

$$H = \hat{H}_{P,\Phi} + E_{P,\Phi} \tag{A.21}$$

$$cov(P, E_{P,\Phi}) = 0 \tag{A.22}$$

Here, notice that the linear single index assumption implies that

$$\begin{aligned} cov(H, P^{\perp\Phi}) &= cov(H, (\beta_0 + \beta_z Z + \Phi)^{\perp\Phi}) && \text{(A.23)} \\ &= cov(H, \beta_z Z^{\perp\Phi}) && \text{(A.24)} \\ &= 0 && \text{(A.25)} \end{aligned}$$

where the last step is due to the linear conditional independence (Assumption 3.7). Thus, the coefficient on $P$ is zero:

$$\theta_p = \frac{cov(H, P^{\perp\Phi})}{var(P^{\perp\Phi})} = 0 \tag{A.26}$$

Thus, we have

$$\hat{H}_{P,\Phi} = \theta_0 + \theta_\phi \Phi \tag{A.27}$$

Lastly, consider the following regression of $H$ on $\Phi$:

$$\mathbb{E}^*[H|\Phi] = \alpha_0 + \alpha_\phi \Phi \tag{A.28}$$

where

$$\hat{H}_\Phi = \alpha_0 + \alpha_\phi \Phi \tag{A.29}$$

$$H = \hat{H}_\Phi + E_\Phi \tag{A.30}$$

Hence, together, we have

$$\mathbb{E}^*[H|Z,\Phi] = H_{Z,\Phi} = \gamma_0 + \gamma_\phi \Phi \tag{A.31}$$

$$\mathbb{E}^*[H|P,\Phi] = H_{P,\Phi} = \theta_0 + \theta_\phi \Phi \tag{A.32}$$

$$\mathbb{E}^*[H|\Phi] = H_{\Phi} = \alpha_0 + \alpha_\phi \Phi \tag{A.33}$$

Because there exists only one unique best linear predictor of $H$ on $\Phi$, it follows that

$$H_\Phi = H_{Z,\Phi} = H_{P,\Phi} \tag{A.34}$$

$$E_\Phi = E_{Z,\Phi} = E_{P,\Phi} \tag{A.35}$$

Therefore, based on Eq (A.17) and Eq (A.22) we have

$$cov(E_\Phi, Z) = 0, cov(E_\Phi, P) = 0 \tag{A.36}$$

Together with the true model, we have

$$H = \alpha_0 + \alpha_\phi \Phi + E_\Phi \tag{A.37}$$

$$= \alpha_0 + \alpha_\phi(-\beta_0 - \beta_z Z + P) + E_\Phi \tag{A.38}$$

$$= (\alpha_0 - \beta_0 \alpha_\Phi) - \beta_z \alpha_\Phi Z + \alpha_\Phi P + E_\Phi \tag{A.39}$$

where Equation (A.38) follows from the Linear Single Index Assumption (Assumption 3.6).

Because $cov(E_\Phi, Z) = cov(E_\Phi, P) = 0$ and, again, there exists a unique best linear predictor, we have

$$\delta_0 = \alpha_0 - \beta_0 \alpha_\Phi \tag{A.40}$$

$$\delta_z = -\beta_z \alpha_\Phi \tag{A.41}$$

$$\delta_p = \alpha_\Phi \tag{A.42}$$

where the regression coefficients $\delta$ is defined as

$$\mathbb{E}^*[H|P,Z] = \delta_0 + \delta_z Z + \delta_p P. \tag{A.43}$$

Therefore, we have

$$-\frac{\delta_z}{\delta_p} = \frac{-\beta_z \alpha_\Phi}{\alpha_\Phi} = \beta_z \tag{A.44}$$

where the last step uses the relevance assumption (Assumption 3.8) $\alpha_\Phi = \frac{cov(H,\Phi)}{var(\Phi)} \neq 0$. □

# B An Illustrative DGP with Competitive Supply

In this section, we present a simple model of the housing market under Cobb-Douglas utility and competitive supply of amenities. Moreover, we show that there could be a natural one-to-one correspondence ("duality") between the household budget constraint and the housing quality index, which motivates us to use measures of the household budget constraints (e.g., household income) as the proxy variable for amenity pricing.

## B.1 Model Set-Up

- **Demand:** Households have Cobb-Douglas utility over housing amenity $Z$, housing quality $\phi$, and the consumption of a numeraire good, C

  $$U_i(Z_i, \Phi_i, C_i) = Z_i^{\theta_i^Z} \Phi_i^{\theta^\phi} C_i^{\theta_i^C} \tag{B.1}$$

  where $\theta_i^Z + \theta^\phi + \theta_i^C = 1$. The preference for amenity is heterogeneous, measured by $\theta_i^Z$. The preferences for quality is $\theta^\phi$. Moreover, we assume that the preference for quality is vertical, namely $\theta^\phi$ is the same across all households. The preference for the outside consumption good $\theta_i^C$.

- **Supply:** There is one price-taking representative firm on the supply side with a cost function for producing amenity and quality: $TC(Z, \Phi)$. For simplicity, we assume linear cost function, then we have

  $$TC(Z, \Phi) = c_z Z + c_\Phi \Phi \tag{B.2}$$

- **Market Clearing:** Let $P(Z, \Phi)$ denote the price of a home with amenity $Z$ and quality $\Phi$. Then, in a competitive equilibrium, marginal cost equals to marginal revenue. So, the cost function determines the price of amenity and quality respectively:

  $$\frac{\partial P}{\partial Z} = \frac{\partial TC}{\partial Z} = c_z \tag{B.3}$$

  $$\frac{\partial P}{\partial \Phi} = \frac{\partial TC}{\partial \Phi} = c_\phi \tag{B.4}$$

  Therefore, the price of a home is

  $$P(Z, \Phi) = \beta_z Z + \beta_\phi \Phi \tag{B.5}$$

  where $\beta_z = c_z$ and $\beta_\phi = c_\phi$. Without loss of generality, we assume the intercept to be zero (i.e., zero profit for the producers).

## B.2 Correspondence Between Housing Quality and Household Budget

Given the model set-up, each household solves its utility maximization problem subject to market prices of the homes and its own budget constraint. We denote their budget constraint by $\eta_i$:

$$\max_{Z,\phi,C} U_i(Z,\Phi,C) \qquad \text{s.t.} \quad \beta_z Z + \beta_\phi \Phi + C \leq \eta_i. \tag{B.6}$$

Solving the utility maximization yields that the expenditure on amenity is a $\theta_i^Z$ fraction of the household's budget and the expenditure on quality is a $\theta^\phi$ fraction of the household's budget:

$$Z_i^* = \frac{\theta_i^Z}{\beta_z}\eta_i \tag{B.7}$$

$$\Phi_i^* = \frac{\theta^\phi}{\beta_\phi}\eta_i \tag{B.8}$$

Since the preference for quality $\Phi$ is vertical (i.e., not heterogeneous by household), there is a one-to-one correspondence between the amount of housing quality chosen $\Phi_i^*$ and the household's budget constraint $\eta_i$. As a result, instead of trying to control for unobserved housing quality, it would be equivalent to control for the household budget set $\eta_i$, for which we can find a proxy.

In other words, while there may be various contributors to one's budget constraints that are unobservable to the econometrician such as one's wealth, inheritance, debt obligations etc., the proxy approach allows us to still estimate the amenity price as long as we can observe a proxy variable for the budget set. For instance, one may consider household income to be relevant for a household's budget; yet, conditional on the total budget, the variations in household income may be thought of plausibly orthogonal to the preference for amenity $\theta_i^Z$, and thus orthogonal to the amount of amenity chosen $Z_i^*$.

# C “Missing” Amenities

In Section 2, we noted that when there are missing amenities, the amenity price estimated by our proxy method may differ from the price one would estimate from a quasi-experiment.

Our proxy method estimates the price a home buyer would pay to obtain the additional amenity, factoring in the existing correlation between the observed amenity and unobserved amenities as well as unobserved quality in the market. This may differ from the impact on home prices due to an exogenous and isolated change in the same amenity, as found in quasi-experiments. Although the latter is useful when comparing the costs of producing such amenities, we believe the former is more pertinent for assessing housing inequality, specifically, the actual amount home-buyers would need to pay to attain a certain amenity level.

To compare our estimate with the amenity price derived from quasi-random variations, we introduce an analogous “bias” formula for our proxy methods. Unlike the OLS methods, we demonstrate that the “bias” of the proxy estimator remains stable even in the presence of unobserved quality. This suggests that our proxy method can offer a more reliable price estimate compared to traditional OLS methods, even when there are ”missing” amenities.

## C.1 Set-Up

Consider a market where prices are determined by two amenities $Z_1$, $Z_2$ and quality $\Phi$. We focus on the linear case where the pricing equation is

$$P = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_\phi \Phi \tag{C.1}$$

where $\beta_1, \beta_2, \beta_\phi \neq 0$, so the Single Index Assumption is satisfied. Further, consider that a proxy variable $H$ exists and satisfies linear conditional independence:

$$cov(H, Z_1^{\perp\Phi}) = 0, \qquad cov(H, Z_2^{\perp\Phi}) = 0. \tag{C.2}$$

Then, the amenity price could be consistently estimated by the proxy method using the ratio of coefficients formula derived in Theorem 3.10 if all amenities are observed:

$$-\frac{\delta_1}{\delta_p} = \beta_1 \tag{C.3}$$

where the coefficients are based on the linear regression

$$\mathbb{E}^*[H|Z_1, Z_2, P] = \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P. \tag{C.4}$$

## C.2 Formula for Missing Amenity Bias - OLS

To start, we first derive the bias using OLS while controlling for $H$

$$\mathbb{E}^*[P|Z_1, H] = \beta_0^{\text{ols}} + \beta_1^{\text{ols}} Z_1 + \beta_H^{\text{ols}} H.$$

We derive the OLS bias formula when quality is imperfectly controlled for

$$\hat{\beta_1}^{\text{ols}} = \frac{cov(P, Z_1^{\perp H})}{var(Z_1^{\perp H})} \tag{C.5}$$

$$= \frac{cov(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_\phi \Phi, Z_1^{\perp H})}{var(Z_1^{\perp H})} \tag{C.6}$$

$$= \beta_1 + \beta_2 \frac{cov(Z_2, Z_1^{\perp H})}{var(Z_1^{\perp H})} + \beta_\phi \frac{cov(\Phi, Z_1^{\perp H})}{var(Z_1^{\perp H})} \tag{C.7}$$

where the first bias component $cov(Z_2, Z_1^{\perp H})$ is driven by the correlation between observed and unobserved amenity and the second bias component $cov(\Phi, Z_1^{\perp H})$ is caused by the fact that $H$ is an imperfect control of $\Phi$.

Similarly, we can sign the direction of the bias based on the signs of the correlation terms. To provide some intuition, suppose that we have positive priced amenities $\beta_1 > 0$ and $\beta_2 > 0$ that are positively correlated $cov(Z_2, Z_1^{\perp \Phi}) > 0$. When $H$ perfectly measures $\Phi$, the OLS estimator is biased upward $\hat{\beta_1^{ols}} > \beta_1$. Moreover, if $H$ does not measure $\Phi$ well, for example, in the extreme case when $H$ is not correlated with $\Phi$ at all, then the additional bias term $cov(\Phi, Z_1^{\perp H})$ will likely to be positive.

As an illustration, suppose $Z_1$ represents “clean air” and $Z_2$ represents “clear roads”, and to the extent that they are positively correlated when controlling for housing quality $\Phi$, then the OLS estimator provides a positively biased estimator. To interpret this “bias”, it essentially means that a home-buyer cannot live in a neighborhood with clean air without also having to pay for the fact that these are less congested neighborhoods, and that is because the presence of these amenities in the market is positively correlated even when conditioning on quality.

### C.3 Formula for Missing Amenity Bias - Proxy

Next, we derive the “bias” formula for the proxy estimator when there are missing amenities. Suppose $Z_2$ is missing, then, the proxy estimator becomes

$$\hat{\beta}_1^{\text{proxy}} = -\frac{\gamma_1}{\gamma_p} \tag{C.8}$$

where the coefficients are based on the linear regression of $H$ on $P$ and $Z_1$

$$\mathbb{E}^*[H|Z_1, P] = \gamma_0 + \gamma_1 Z_1 + \gamma_p P. \tag{C.9}$$

Recall the residual term of a linear regression is uncorrelated with the regressor

$$\mathbb{E}^*[H|Z_1, Z_2, P] = \delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P \tag{C.10}$$

$$cov(\epsilon_{Z_1,Z_2,P}, Z_1) = 0 \tag{C.11}$$

$$cov(\epsilon_{Z_1,Z_2,P}, P) = 0. \tag{C.12}$$

We can rewrite the numerator of the proxy estimator as follows:

$$\begin{aligned}\gamma_1 &= \frac{cov(H, Z_1^{\perp P})}{var(Z_1^{\perp P})} &\text{(C.13)}\\ &= \frac{cov(\delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P + \epsilon_{Z_1,Z_2,P},\ Z_1^{\perp P})}{var(Z_1^{\perp P})} &\text{(C.14)}\\ &= \delta_1 + \delta_2 \frac{cov(Z_2, Z_1^{\perp P})}{var(Z_1^{\perp P})} &\text{(C.15)}\end{aligned}$$

Similarly, we can rewrite the denominator of the proxy estimator as follows:

$$\begin{aligned}\gamma_p &= \frac{cov(H, P^{\perp Z_1})}{var(P^{\perp Z_1})} &\text{(C.16)}\\ &= \frac{cov(\delta_0 + \delta_1 Z_1 + \delta_2 Z_2 + \delta_p P + \epsilon_{Z_1,Z_2,P},\ P^{\perp Z_1})}{var(P^{\perp Z_1})} &\text{(C.17)}\\ &= \delta_p + \delta_2 \frac{cov(Z_2, P^{\perp Z_1})}{var(P^{\perp Z_1})}. &\text{(C.18)}\end{aligned}$$

Note that the covariance term $cov(\epsilon_{Z_1,Z_2,P}, Z_1^{\perp P}) = 0$ and $cov(\epsilon_{Z_1,Z_2,P}, P^{\perp Z_1}) = 0$, because the residual term $\epsilon_{Z_1,Z_2,P}$ is by construction uncorrelated with $P$ and $Z_1$. Putting the two together, the proxy estimator can be written as

$$\hat{\beta}_1^{\text{proxy}} = -\frac{\gamma_1}{\gamma_p} = -\frac{\delta_1 + \delta_2 \times \psi_1}{\delta_p + \delta_2 \times \psi_p} \tag{C.19}$$

where $\psi_1$ and $\psi_p$ is defined by the linear regression of $Z_2$ on $P$ and $Z_1$:

$$\mathbb{E}^*[Z_2|P, Z_1] = \psi_0 + \psi_1 Z_1 + \psi_p P. \tag{C.20}$$

Given that $-\frac{\delta_1}{\delta_p} = \beta_1$, we have that

$$\text{if} -\frac{\psi_1}{\psi_p} = \beta_1, \quad \text{then} \quad \hat{\beta_1}^{\text{proxy}} = \beta_1. \tag{C.21}$$

The derivation of the proxy bias in Equation (C.19) also clarifies that the bias due to missing amenities in the proxy estimator is *not* sensitive to how well the quality $\Phi$ itself is observed. In particular, the proxy "bias" is determined by the regression coefficients of $Z_2$ on $P$ and $Z_1$,[19] which does not contain $H$. The "bias" is also affected by $\delta_2$, which comes from the regression of $H$ on $P$, $Z_1$, and $Z_2$ and is not sensitive to the measurement errors in the outcome variable.[20] Therefore, even though neither the proxy method nor the OLS method is inherently designed to address the issue of

[19] One may also note that this ratio $-\psi_1/\psi_p$ can be viewed as the proxy estimator for the price of amenity when we use $Z_2$ as the proxy variable. Of course, this model is likely misspecified unless the corresponding conditional independence assumption happens to hold true. Namely, $cov(Z_2, Z_1^{\perp \tilde{\Phi}}) = 0$ where $\tilde{\Phi} = \beta_2/\beta_\phi Z_2 + \Phi$ measures an adjusted quality. Equivalently, the proxy estimator coincides the quasi-random estimate when $cov(Z_1, Z_2) = cov(Z_1, \tilde{\Phi})cov(Z_2, \tilde{\Phi})/var(\tilde{\Phi})$.

[20] Note that more measurement error in $H$ does not change the estimate of $\delta_2$ because $H$ is on the LHS as the outcome variable in the regression, which is a key difference from the OLS case where $H$ is included in the RHS as a regression

missing amenities, the proxy estimate remains robust to unobserved qualities even in the presence of missing amenities.

Moreover, the bias term $-\frac{\phi_1}{\psi_p}$ for the proxy estimator encapsulates both the correlation between the unobserved amenity $Z_2$ and the observed amenity $Z_1$, but also the correlation between the unobserved amenity $Z_2$ with the unobserved quality, indirectly through its impact on price $P$.

As an illustration, suppose again $Z_1$ represents "clean air" and $Z_2$ represents "clear roads". They are positively correlated when controlling for housing quality $\Phi$, but they are also positively correlated with housing quality, as richer households tend to consume more of both amenities and housing quality. Then, to interpret the proxy "bias", it means that to access cleaner air, a home buyer can achieve this through a combination of purchasing more "clear roads" and shifting to a segment of higher housing quality.

## C.4 Simulation

In this section, we illustrate the property of both OLS and the proxy method when there is an unobserved amenity by running simulations using a DGP based on Cobb-Douglas preferences and competitive supply as outlined previously in Section **??**.

### C.4.1 DGP for Simulation

The DGP is based on Cobb-Douglas demand over two amenities $Z_1$ and $Z_2$.

$$U_i(Z_1, Z_2, \Phi, C) = Z_1^{\theta_i^{Z_1}} Z_2^{\theta_i^{Z_2}} \Phi^{\theta^\phi} C^{\theta_i^C} \tag{C.22}$$

where $\theta_i^{Z_1} + \theta_i^{Z_2} + \theta^\phi + \theta_i^C = 1$. The true price for the amenities is $(\beta_1, \beta_2) = (0.3, 0.3)$. The first amenity $Z_1$ is observed by the researcher, but $Z_2$ is not.

Households budget sets $\eta_i$ are simulated as follows: The set of all possible values of household budgets $\eta_i$ is $\mathcal{E} = \{10n | n = 1, 2, ..., 100\}$. For each value in $\mathcal{E}$, there are exactly 50,000 households with this amount of budget, so there are a total of 5,000,000 households in the simulation. We simulate under both the case when $\eta_i$ is perfectly observed, and when $\eta_i$ is observed with a random noise as $h_i = 0.7\eta_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 100)$ is the random noise.

The preference parameter for quality, $\theta^\phi$, is 0.5. The preference parameter for the amenities, $\theta_i^{Z_1}$ and $\theta_i^{Z_2}$, are drawn from a joint distribution, independently of $\eta_i$. Each has a uniform marginal distribution on $[0, 0.25)$, and the correlation between them is $corr(\theta_i^{Z_1}, \theta_i^{Z_2}) = \rho$. The correlation $\rho$ is a key determinant of the direction of missing amenity bias, since it captures the co-movement of the demand for the two amenities as the expenditure on each amenity is driven by one's preferences. Therefore, we examine the results under various correlation $\rho$ in our simulations. Finally, $\theta_i^C = 1 - \theta_i^{Z_1} - \theta_i^{Z_2} - \theta^\phi$.

### C.4.2 Simulation Results

In the first set of simulations, we simulated the data by varying the correlation between the preference for the observed and unobserved amenity to examine how the proxy and OLS results respond to the change in $\rho$.

The results from the simulations are presented in Figure C.1. The x-axis represents the correlation between preference for amenities in each simulation, while the colored series depicts the estimated price for amenity one using different methods.

First, let's consider a scenario where all household budget sets, and therefore the amount of housing quality chosen, is perfectly observed. In this case, the solid blue and green lines indicate the estimated price for amenity using OLS and proxy, respectively. Notice that in the absence of correlation between preferences, the OLS method (solid blue) restores the original price $\beta_1$. The proxy method (solid green) also restores the correct price but at a different point when there is some positive correlation in preference such that $-\psi_1/\psi_p = \beta_1$. Moreover, these two lines intersect at both ends, where the preferences for the two amenities are perfectly correlated. This is to be expected in the case of $\rho = \pm 1$, because the price of $Z_2$ cannot be separately identified from $Z_1$, and both methods are reduced to estimating the price of the single combined amenity.

Next, when the household budget sets are observed with noise, the amount of housing quality chosen is measured with noise. In this case, the OLS estimator yields a highly biased result (dotted blue line), while the proxy estimator remains unaffected and robust to components of unobserved quality (dotted green line).

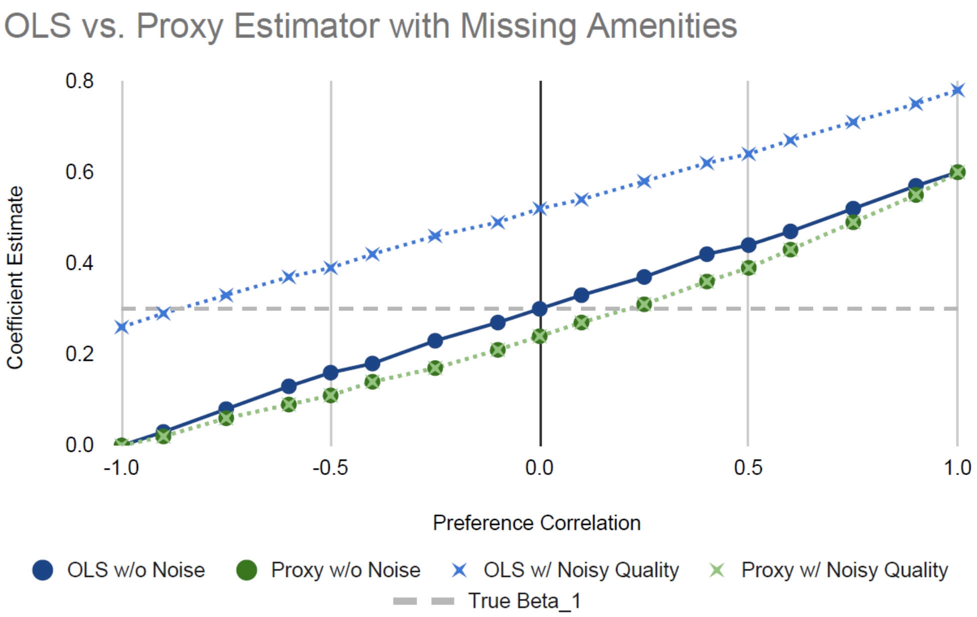
Furthermore, when we explore how sensitive our estimate is to the relationship between the proxy $h$ and the budget set $\eta$, we simulate the data by holding the total variance of $\eta$ fixed while varying the signal to total ratio of $h$.[21] If the variance of $h$ is significantly higher than the variance of $\epsilon$, then most of the variation in $\eta$ is captured in $h$, and we will have a high signal-to-total ratio, and vice versa.

Figure C.2 shows that as the signal-to-total ratio decreases (going from the right of the graph to the left), the bias from OLS with proxy control increases steadily, while the bias for the proxy method is unchanged across different levels of signal to total ratio.[22] This simulation result is consistent with our derivation where the missing amenity "bias" of the proxy estimator is robust to varying levels of signal-to-total ratio of the proxy variable, which is a property that OLS does not have.

[21] To simulate this, households have a random draw of both proxy and noise, with $h_i \sim \mathcal{N}(0, 100)$ and $\epsilon_i \sim \mathcal{N}(0, 100)$, and the budget set is given as $\eta_i = \lambda h_i + (1-\lambda)\epsilon_i$, where $\lambda$ denotes varying levels of how informative $h$ is.

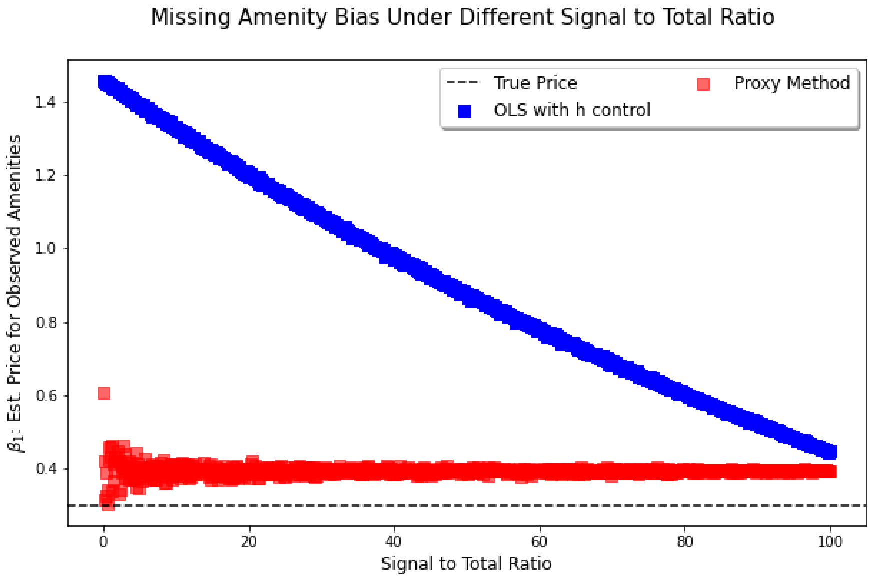
[22] When the signal-to-total ratio is very close to zero, the proxy method suffers from a "weak" proxy problem, analogous to the weak instrument problem of an IV estimator, where the denominator in the ratio-of-coefficients becomes close to zero and leading to larger standard errors.

Figure C.1: Missing Amenities Bias v.s. Preference Correlation



*Notes*: This figure compares the OLS and proxy estimator with a missing amenity under varying levels of preference correlations for the two underlying amenities. OLS estimates are shown in blue lines, and the proxy estimates are shown in green lines. Both methods can suffer from the missing amenity bias. However, in the presence of imperfectly measured quality, the bias in the OLS estimate (dotted blue line) is substantial, whereas the proxy estimator is not affected (dotted green line).

Figure C.2: Missing Amenity Bias v.s Signal to Total Ratio



*Notes*: This figure shows that as the signal-to-total ratio decreases (going from the right of the graph to the left), the bias from OLS with proxy control increases steadily, while the bias for the proxy method is unchanged across different levels of signal-to-total ratio.

# D An Illustrative DGP with Fixed Supply

In this section, we provide a simulation exercise to illustrate the DGP for prices when the supply of amenity and quality is fixed. In this case, the prices are pinned down by the market clearing condition. We show the analogous trade-offs between price and amenity at varying levels of underlying quality. We also show that, when the household preference for quality is sufficiently strong, the consumption of housing quality is roughly segmented by household budgets.

## D.1 Simulation Set-Up

- **Demand:** There are $N$ households in the market indexed by $i$. Each must choose exactly one unit of housing. Each household has an endowment of budget $\eta_i$ from a random draw with distribution $U(0,1)$, which they will allocate between housing and a numeraire good. Households have Cobb-Douglas preferences over housing characteristics and the numeraire good. The preference parameter for housing quality, $\theta^{\phi}$, is uniform across all households, while households have heterogeneous preference parameter for housing amenities, $\theta_i^Z$, drawn randomly from $U[0, 1-\theta^{\phi})$. Therefore, the utility that household $i$ derives from choosing the housing type $(Z, \Phi)$ is

$$U_i(Z,\Phi) = (Z+1)^{\theta_i^Z}(\Phi+1)^{\theta^{\phi}}(\eta_i - P(Z,\Phi))^{1-\theta_i^Z-\theta^{\phi}}. \tag{D.1}$$

- **Supply:** There are $N$ housing units in the market indexed by $j$. Each house $j$ is characterized by a triplet of housing attributes $(Z_j, \Phi_j, P_j)$, where $Z_j \in \{0,1\}$ is a binary (positive) amenity, $\Phi_j \in \{0,1,2\}$ measures the quality of housing, and $P_j$ is price of the house. The current housing stock is fixed and exogenously determined. Specifically, there are $\frac{N}{6}$ houses for each particular housing type in the market.

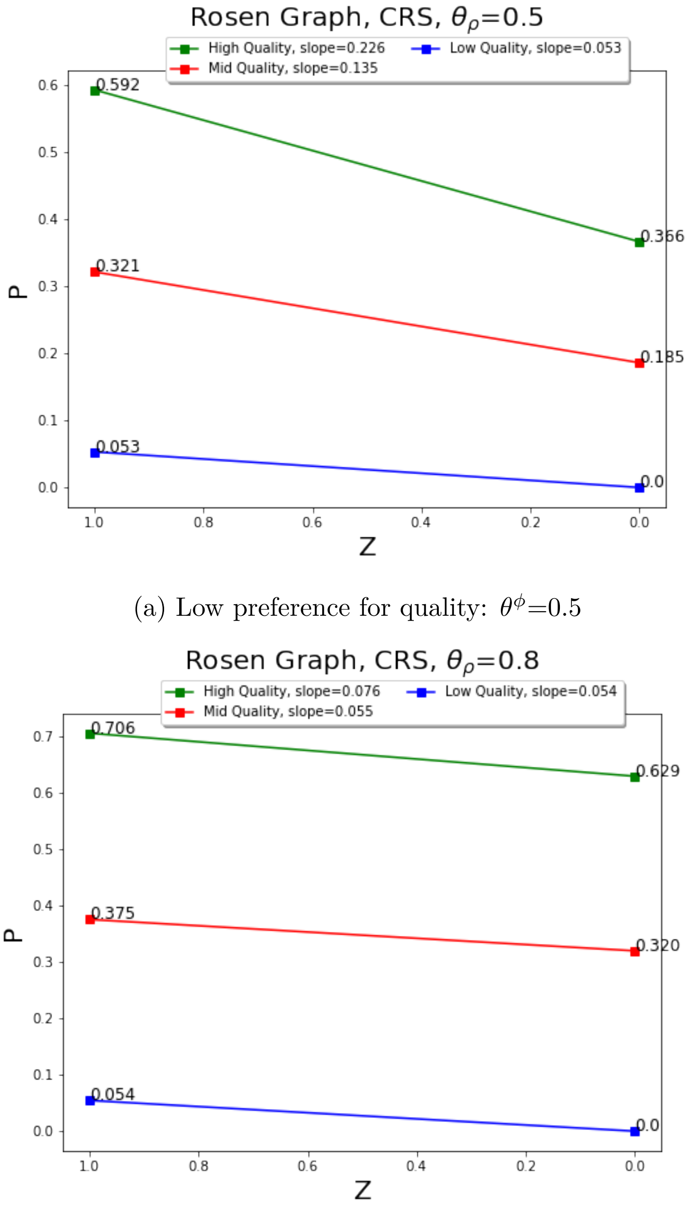
- **Market Clearing:** In equilibrium, the price of each housing type is then determined by the market clearing condition, under which the allocation of households to housing types by their optimal choice perfectly matches the distribution of the fixed housing stock. We also require that prices be measurable with respect to quality and amenity $\Big(\text{i.e, } (Z_j,\Phi_j) = (Z_{j'},\Phi_{j'}) \Rightarrow P_j = P_{j'}\Big)$. The price for the housing type $(Z=0, \Phi=0)$ is normalized to be 0. Therefore, the equilibrium price vector consists of five prices $P_{Z,\Phi}$ for each of the remaining housing types.

## D.2 Simulation Results

We perform two simulations using the models described above. In the first simulation, we set $\theta^{\phi} = 0.5$ to reflect the case when households have a moderate preference for housing quality, and in the second simulation, households have a strong preference for housing quality with $\theta^{\phi} = 0.8$. In both simulations, the total number of households is 8000. Figure D.1a and D.1b show the equilibrium price for each housing type in both simulations. The different colors mark the quality of the housing types, the x-axis indicates the amenity, and the y-axis shows the equilibrium price. Housing types with the same quality are connected by a line to reflect the Rosen frontier of specific quality segments. In other words, a trade-off exists within each segment of the market: for a given housing quality, buyers will have to pay more for their home if they also want to have clean air. This pattern mirrors that of Figure 2.

Figure D.2 graphically displays the allocation of housing types by underlying preference parameters and the budget sets in the simulations. The color of the dots in the graph represents the choice of housing type by individual households, and the x- and y-axis represent respectively their endowed budget set $\eta_i$ and preference for amenity $\theta_Z$. In Figure D.2a, when household preference for quality is weaker, it is not necessarily the case that households with higher budgets will necessarily choose houses with higher quality. However, when the household preference for housing quality is sufficiently high, as in Figure D.2b, we observe a nearly one-to-one relationship between the household's budget and the housing quality chosen, and hence, the market is segmented accordingly.

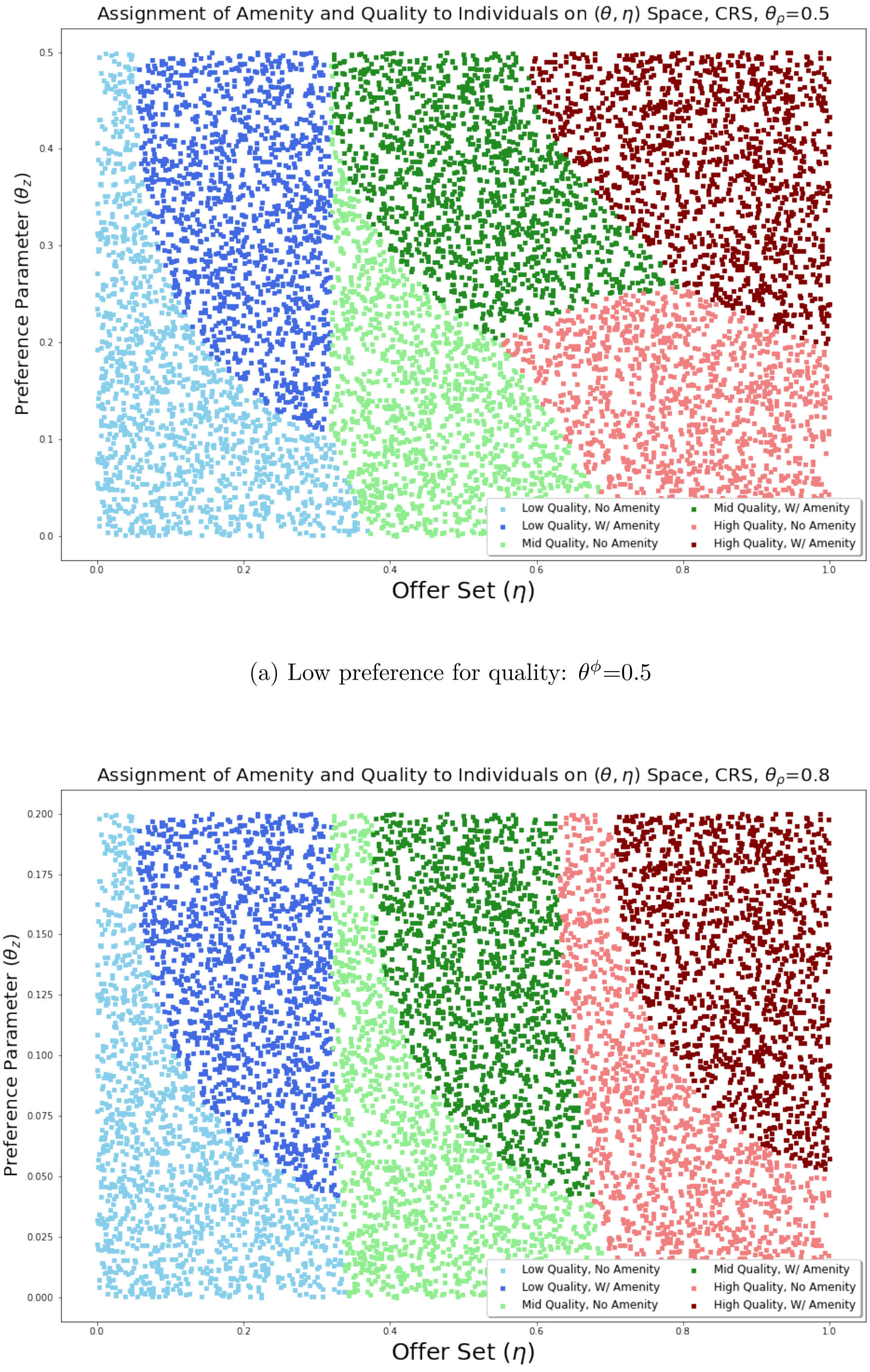
Figure D.1: Simulated Rosen Frontier For Each Quality Segment



(a) Low preference for quality: $\theta^{\phi}$=0.5

(b) Low preference for quality: $\theta^{\phi}$=0.8

Figure D.2: Simulated Allocation of Housing Types



(a) Low preference for quality: $\theta^{\phi}$=0.5

(b) High preference for quality: $\theta^{\phi}$=0.8