Chapter 2: Using Property Tax Assessments and Business Filings to Identify Landowners and Operationalize their Characteristics

Chapter Abstract

Property owners play central roles in many urban sociological theories, but empirical analysis of these actors has frequently been stymied by a lack of data. Few surveys collect detailed information on landowners and using administrative data sources presents multiple challenges, including generally sparse information, the difficulty of accurately operationalizing rental properties, and, most importantly, the fact that property owners frequently obscure their identities through corporate structures. This paper presents a data construction pipeline for creating linked, longitudinal datasets describing urban properties and the people and companies that own them using widely available tax assessment records and business filings. The author implements this approach in four metropolitan areas — Boston, Massachusetts, Baltimore Maryland, Miami, Florida, and Houston, Texas — between 2005 and 2020, demonstrating the adaptability of the method to areas with differing levels of data quality. The construction pipeline draws on four methodological innovations. First, it uses internal validation and external data harmonization to transform tax assessment records into accurate representations of urban housing markets. Second, it presents a network-based entity reconciliation methodology better suited than existing methods to the sparse but linked data contained in the source records. Third, it presents a flexible and comprehensive method for operationalizing landowners' corporate networks. Finally, it operationalizes multiple measures describing landowners' sociological characteristics, including their organizational formality, geographic locations, scales of operations, and racial identities, and it estimates potential bias in these measures. Steps in the data construction pipeline are validated internally and externally and resulting constructs are compared to those derived from alternate approaches.The paper concludes by demonstrating the range of empirical analyses this methodology opens.

Introduction

Urban sociological theories have long argued for the importance of urban landowners. Speculators and developers shape patterns of neighborhood turnover and change (Schwirian 1983; Molotch and Logan 1987), with the largest forming coalitions with local politicians and corporate executives to affect entire cities' growth trajectories (Molotch et al. 2000; Hackworth 2006). Landlords' decisions about whom to rent to, when to evict tenants, what to charge for rent, and whether to maintain their properties shape residential segregation (Massey and Denton 1993; Pager and Shepherd 2008), residential precarity (Desmond 2012; Garboden and Rosen 2019), cost of living (Desmond and Wilmers 2019; Leung et al. 2020), and disinvestment patterns (Sternlieb 1966; Travis 2019), respectively. Landowners are an extremely heterogeneous

group, ranging from small-scale "mom and pops" to massive publicly traded corporations (Mallach 2007), with similarly varied repertoires of behavior (Shiffer-Sebba 2020). Studies suggest that in recent decades urban landowners have become increasingly large-scale and financialized (Fields 2018; August 2020), making it all the more important to understand these social actors.

However, the individuals and companies who own American cities have remained empirically under examined due to a paucity of quantitative data. Surveys rarely ask questions about respondents' landlords or sample landowners directly, and those that do rarely collect detailed information, sufficiently sample the largest owners, or link landlords to the properties they own. This lack of data has made a range of important questions difficult if not impossible to answer. Basic empirical facts, like the proportion of rental properties owned by large- versus small-scale landlords and the differences in landlord characteristics between majority-White and majority-Black neighborhoods, remain largely unknown.

One potential solution is the growing number of administrative records describing land ownership, since these data describe full populations of landowners and can be linked to a range of auxiliary datasets. However, these administrative records, created for purposes other than research, present multiple challenges: Most importantly, they give only the names of the often-anonymous corporate entities through which landowners hold properties, concealing the true identities of owners and the connections between them. They also contain only sparse information about landowners, impeding entity reconciliation efforts and the operationalization of relevant characteristics. Finally, they often contain inaccuracies and biases that must be corrected before they can be useful representations of urban land markets. Methodologies to solve these problems have been an active area of methodological research within urban studies

(Hangen and O'Brien 2024; Messamore 2024) but have not yet developed a comprehensive methodology that solves each of these problems.

In this paper, I detail a methodology that overcomes these challenges and creates linked, longitudinal datasets describing landowners, the properties they own, and events at those properties. Drawing on over 87 million tax assessment, business filing, and eviction filing records, I apply this data construction pipeline to four metropolitan areas — Boston, Massachusetts, Baltimore, Maryland, Miami, Florida, and Houston, Texas — between 2005 and 2019. The pipeline consists of four methodological subtasks: First, I use internal validations and harmonization with external datasets to address the biases within tax assessment records and make them accurate representations of urban land and housing markets. Second, I develop a novel network-based entity reconciliation method for linking records that have limited, but networked, information about each individual. Third, I develop a corporate network identification methodology to uncover the complex configurations of people and companies that constitute individual landlords. Finally, I develop operationalizations of several sociologically important landlord characteristics, (including their organizational formality, racial identities, home locations, corporate locations, and scales of operations), and, critically, I assess these measures' potential biases. Throughout the process, I ensure that the datasets are accurate, reproducible, and transparent by conducting internal and external validations, creating multiple implementations of key variables, and making all data and code publicly available online (Nelson 2019).

I conclude the paper by demonstrating some of the new analytic possibilities that this methodology opens. These data provide the first comprehensive descriptive portrait of landowner characteristics across multiple metropolitan areas and time periods, allowing detailed analyses of

differences in landlord composition between neighborhoods and changes in property ownership over time. I find that although LLC ownership has increased dramatically in the past 15 years, this may simply indicate a change in the legal form through which properties are owned, rather than a change in types of owners, since more sophisticated measures of organizational formality show only muted increases. Likewise, I find that increases in landlord scale have been smaller than many studies suggest. However, I find large differences between neighborhoods in landlord characteristics, with poorer and more non-White areas showing more large-scale, organizationally formal owners. Finally, because the datasets contain a full population of owners, they allow for identification of the small numbers of elite owners whose behaviors are most consequential for urban life, and I conclude with a brief demonstration of this type of small-N analysis. Because all data and code are published publicly online and the necessary administrative datasets are widely available across the United States, scholars can use this methodology to develop landowner measures in other geographic areas and conduct similar analyses.

Background

Existing Approaches to Studying Landlords and Landowners

Few extant datasets contain detailed information on landowner characteristics or behaviors. Urban-focused surveys rarely sample landowners directly or contain questions about respondents' landlords, and housing-focused surveys contain only limited information about owners' characteristics, frequently just their legal forms.[4] As a result, until recently much of the

[4] The Property Owners and Managers Survey contains more detailed information but was fielded nearly thirty years ago, does not provide property-level data that can be linked to other datasets, and omits many important measures, such as landlord races for properties owned through shell companies.

scholarship on landowners has been theoretical, has used qualitative or case study methods, or has made do with imperfect quantitative data.

For example, studies of the largest urban landowners and the urban regimes they form with other elite actors have typically been theoretical works (Molotch and Logan 1987; Feagin and Parker 1990), have used qualitative methods to study a single case (Warner and Molotch 1995; Gotham 2000), or have used indirect proxies for landowner characteristics (Lyon et al. 1981; Schneider 1992). Likewise, in studies of gentrification the roles of speculators and developers has frequently been elaborated in theory (Smith 1979; Marcuse and Madden 2016), noted but not made the empirical focus, or analyzed through qualitative methods focused on a single case (Lees 2003; Freeman 2006).

Similarly, much of the research on landlord behavior has drawn on qualitative and ethnographic methods (Desmond 2016; Garboden et al. 2018). These studies have been highly productive in elaborating and analyzing the relationships between landlords and tenants (Balzarini and Boyd 2020; Rosen and Garboden 2020), the ways in which landlords frame and make management decisions (Shiffer-Sebba 2020; Garboden et al. 2018), the effects of those decisions on tenants (Leung et al. 2020; Desmond 2016), and other interpretive, processual topics for which qualitative methods are best suited. Quantitative studies have frequently made use of indirect indicators for landlord characteristics, such as the presence of large multi-family properties (Gilderbloom and Appelbaum 1987), which suffer from identification problems. Other quantitative works have drawn on relatively uncommon datasets, such as the Rental Housing Finance Survey's restricted data (Desmond and Wilmers 2019), surveys they fielded themselves (Desmond and Gershenson 2016; Decker 2023), or audits (Fischer and Massey 2004; Pager and Shepherd 2008).

The review presented above is not intended to criticize the extant work, but merely to note that those analyses that require large-scale, comparative quantitative data on full populations of owners have been rare. Specifically, topics like the aggregate composition of landlords, the characteristics and behaviors of the largest owners, the relationships between owners, and the varied behaviors of landlords remain underexplored.

The Possibilities and Challenges in Using Big Data to Study Landowners

However, in the past two decades cities have increasingly made available digitized administrative records (Connelly et al. 2016; Salganik 2017), which, along with other forms of big data, have been used to operationalize many important sociological topics, including physical disorder (Hwang and Naik 2023; O'Brien et al. 2015), political and economic ideologies (Rule et al. 2015; Goldenstein and Poschmann 2019), and social protests (Zhang and Pan 2019). Several studies have used the increasing availability of administrative data to study landowners (Travis 2019; Gomory 2022), typically drawing on tax assessment records and business filings, two data sources with considerable potential for operationalizing landowners. For example, studies using these data sources have examined differences in landlord eviction behaviors (Raymond et al. 2018; Immergluck et al. 2019), maintenance practices (Travis 2019) and investment patterns (Seymour and Akers 2020; Seymour and Shelton 2023). However, because tax assessments and business filings were created for purposes other than research, they present several challenges (Salganik 2017), and overcoming these difficulties constitutes an active and unresolved area of methodological research (Hangen and O'Brien 2024; Messamore 2024).

Biases and inaccuracies in tax assessment records: Tax assessment records frequently have missing data, inconsistent identifiers between years, and other inaccuracies created by changes in data collection (Salganik 2017). Studies have addressed these issues through ad hoc data cleaning (Travis 2019; Gomory 2022), but no systematic cleaning and validation methodologies exist. Tax assessment records also rarely distinguish rental from non-rental properties, making it difficult to identify the former, particularly for single-unit properties. Studies have addressed this by drawing on owner-occupancy flags and rental registries (separate datasets identifying rentals), but using owner-occupancy likely overstates single-unit rentals and few cities collect registries (Preis 2024).

Linking people and companies within and between datasets: Entity reconciliation, or linking the instances where a person or company is mentioned within and between datasets, is a common and often essential task in computational social science (Elmagarmid et al. 2007; Enamorado et al. 2018). However, most approaches rely on a broader range of information (e.g., birth dates, genders) than is provided in tax assessments and business filings and cannot readily leverage the networked information these records contain. Landlord studies that have used these records have typically used ad hoc entity reconciliation procedures or methods that do not fully take advantage of the networked nature of the data (Hangen and O'Brien 2024; Gomory 2022).

Operationalizing corporate networks: Perhaps most importantly, tax assessments detail only the legal owners of properties, and landowners frequently own properties through shell companies and complex corporate configurations that obscure their true identities (Travis 2019; Messamore 2024). Identifying all of the people and companies that constitute a single landowner is essential

for accurately operationalizing sociologically relevant characteristics, like their scales of operations and degrees of corporate formality, but doing so is very difficult.

Extant studies have overcome the problem of corporate obscurity in several ways. First, some studies analyze only the direct owning entities, limiting their analyses to landlord characteristics like liability protection or contact addresses that do not require identification of the full corporate network (Travis 2019; Preis 2024). Other studies link owning entities using only the names and addresses in tax assessments, but this data is very sparse and is likely to produce a large number of both false positives and negatives (Gomory and Desmond 2023). A third approach is to limit analyses to only a handful of very large owners, whose corporate networks can be identified through publicly available records and qualitative methods (August 2020; Charles 2020), but the time and data necessary for this approach make it inapplicable to the broader population of landowners.

A fourth approach to landlord obscurity is to draw on state-level business filings, which detail the officers of the anonymous companies found in tax assessment records. However, even with business filing data, it is often unclear when to link particular people and companies together. Although several studies have made use of these records, none has provided a methodology that draws on a sufficiently wide array of potential links or constructs the links with sufficient sophistication to avoid false negatives and false positives, both of which are essential to accurately operationalizing landowners' corporate networks (Gomory 2022; Hangen and O'Brien 2024; Messamore 2024).

Sparse data and difficulty operationalizing landowner characteristics: Because tax assessments and business filings were not created for the purpose of research, they do not contain many of the

sociological characteristics of greatest interest to landowner scholars, such as measures of organizational formality, racial identity, and geographic location. Accordingly, landowner studies have frequently used scale of ownership as the primary measure through which owners are understood, but this measure is only indirectly associated with many of the concepts of greatest importance to scholars. Further work is necessary to leverage the existing data to create measures with greater theoretical purchase and to ensure that these constructs are unbiased.

The data construction pipeline below addresses each of these problems, producing validated, accurate longitudinal datasets describing landowners and the properties they own.

Data Sources

I used two primary data sources, tax assessment records and state-level business filings. Tax assessment records are collected by local tax assessor offices[5] and detail the land usage, number of residential units, square footage of living space, lot size, year built, owner's name and address, and owner-occupancy tax exemption for all parcels of land in each year. Of the original 63.1 million parcel-years of data, I dropped 1.5 million (or 2.3%) because they were missing an owner name or parcel address and 0.7 million (or 1.1%) because they were duplicates, leaving 61.0 million. Parcel shapefiles, linked to the tax assessment records, were also collected.[6]

Business filings are collected by state Secretaries of Commerce and detail the principal address, legal type, officers, officers' titles and addresses, resident agent, and resident agent's address for all companies registered to do business within a state, including those originating in a

[5] These records are made available at the municipal level in Massachusetts, the county level in Baltimore and Texas, and the state level in Florida.
[6] 98.9%, 97.9%, 99.2%, and 93.6% of parcel-years in Boston, Baltimore, Miami, and Houston metro areas were linked to records in the parcel shapefiles, respectively.

different state or country. Two exceptions are banks, which file at the national level, and trusts, which are not companies and thus not required to submit filings.

These records are collected in a variety of relational database formats, with a range of variables and variable values, so before beginning data construction I rearranged these into a single uniform format.[7] Table 1 shows the years and number of records, from each source, after reformatting.

Table 1: Input Datasets

Geography	Years	N	Geography	Years	N
Tax assessment records			*Business filing records*		
Boston metro			Massachusetts	2000-2019	1,214,217
Boston city	2003-2018	2,434,957	Maryland	2000-2020	1,574,679
MA cities	2010-2019[8]	1,393,307			
Baltimore metro			Florida	pre-2019[9]	7,888,735
Baltimore City County	2005-2020	1,423,189	Texas	pre-2019	5,994,099
Baltimore County	2005-2020	2,075,310	*Eviction filing records*		
Anne Arundel County	2005-2020	1,465,521	Boston metro	2001-2016	132,556
Harford County	2005-2020	676,506	Miami metro	1999-2017	629,042
Miami metro			Houston metro	2001-2016	768,273
Florida[10]	2008-2018	24,668,262	*Miscellaneous datasets*		
Houston metro			Census/ACS data at the tract level for 2000,		

[7] Appendix A.1. details the cleaning and reformatting process. Once in a uniform format, the same data construction pipeline was applied for each set of records, so if new tax assessments and business filings were cleaned to conform to these uniform standards, the data pipeline could be run using them. I have made code publicly available to check that cleaned datasets conform to the uniform standard.

[8] In Massachusetts municipalities other than Boston, the tax assessment records are only available for particular years, which differ by municipality. In Maryland, tax assessment records are made available by a third-party contractor for a fee. To reduce costs, only the years 2005, 2007, 2009, 2011, 2014, 2017, and 2020 (2008 instead of 2009 in Baltimore city) were purchased.

[9] Year data is missing for Miami and Houston, so all filings before 2019 were included.

[10] State-wide data was subsetted to Miami-Dade, Broward, and Palm Beach Counties.

Harris County	2005-2020	22,035,748	2010, and 2020, and 2015-2020
Galveston County	2007-2020	1,989,625	Census data at the block level for 2000, 2010, and 2020
Montgomery County	2007-2020	2,756,684	
Fort Bend County	2015-2019	1,848,420	Street addresses associated with parcels in Boston City, MA
Brazoria County	2007-2019	1,823,907	

Note: Table 1 lists the datasets used as inputs in the data construction methodology presented in this chapter and used in the subsequent chapters.

Creating a Longitudinal Parcel Dataset

Tax assessment records have many advantages as a data source for describing rental properties: (1) They are complete populations, rather than samples, which is necessary for many analyses, such as identification of the largest property owners (Salganik 2017:17). (2) They are detailed to the property level, rather than the block or tract, which is essential for studying ownership. (3) They contain real addresses and names, rather than proxies, allowing the incorporation of auxiliary datasets like business filings and eviction filings. (4) They are typically collected yearly, allowing for longitudinal analysis. (5) They are unmediated measures, rather than self-reports of properties owned, reducing potential self-report bias (Jerolmack and Khan 2014; Salganik 2017). (6) Cities frequently make them available for free or at a low cost, allowing them to be gathered at scale.

However, there are also several challenges in using tax assessment records (Preis 2024; Gomory 2022). (1) They exhibit "drift," meaning unique identifiers and variable values may change between years due to administrative changes made by the collecting agency (Salganik 2017:33). (2) Like many administrative records, they are "dirty," meaning variables may be missing or inaccurate for particular geographies and years (Salganik 2017:37). (3) Yearly data is unavailable in some cases. (4) Tax assessment records rarely contain sufficient information to definitively determine which residential properties are rentals, rather than owner-occupied,

vacant, or otherwise not for rent, particularly for single-unit properties (Salganik 2017:24; Preis 2024).[11] I used a range of imputations, validation tests, and corrections, detailed below, to address these challenges. I focused on ensuring accurate data for unit counts, unique identifiers, and land usages, as these are the variables that are essential for measuring ownership.

Identifying Instances of Administrative "Drift" Using Internal Validation

Over time, tax assessment agencies change their data collection practices, such as how they categorize land usages or identify parcels, creating inaccuracies (Salganik 2017). I identified and corrected these issues using internal validation—calculating the proportion of parcels within each city and year that no longer existed or changed land usage in the subsequent year, and identifying city-years with high values that suggested an administrative artifact. For example, in the original data, all but a handful of parcels in Galveston County, Texas, were classified incorrectly as single-family in 2009, which I identified by noting that nearly half changed land usage between 2009 and 2010. I corrected inaccurate identifiers and land usages, where possible, using values from subsequent or previous years (for more details, see Appendix B). After making these corrections, among all parcels, fewer than 1% changed land usage or ceased to exist between years, and the city-year with the highest yearly change had fewer than 10% of parcels change land usage or cease to exist.[12]

[11] I consider condominiums to be single-unit properties, in this context, because even if a condominium is in a multi-unit building, similar data ambiguities arise as with single-family properties.
[12] I chose 10% as a threshold because at this point the vast majority of such instances were not due to data problems but instead small cities where a large development had been built.

Imputing Missing and Inaccurate Parcel Data

In addition to instances of administrative drift, variables were simply inaccurate or missing for particular areas and years (Salganik 2017). Unit data was missing for all large multi-family properties in Fort Bend, Galveston, and Montgomery counties in the Houston metro, and in Boston city proper. I imputed unit data for these records by drawing on auxiliary datasets of street addresses, the property's square footage, and Census counts of housing units in the same block. Owner-occupancy data was missing for the cities in Massachusetts outside of Boston, and I imputed owner-occupancy based on the owner's name and contact address. For more details see Appendix C.

Interpolating Between Years

Data for every year between 2005 and 2020 was not available for every parcel. In the municipalities surrounding Boston proper, data was only collected for intermittent years between 2010 and 2019; in Baltimore metro, yearly tax assessment rolls were prohibitively expensive and I obtained records at three-year intervals; and in other cases data was not available as early as 2005. To create a balanced longitudinal dataset, I assumed that a parcel existed previously and continued to exist if it was in the first or last year of available data, respectively, and I interpolated values between years by using the most recent value for categorical variables, and doing linear interpolation for continuous variables.

Table 2 shows the proportion of variables that were originally nonmissing, were imputed, and were interpolated, for both the tax assessment records and the resulting longitudinal dataset. As discussed above, relatively few observations were missing unit, land usage, or owner-occupancy data, and all but a trivial number of those missing were imputed. Year built

and parcel area were missing most often in the original records, at 77% and 97%, respectively. These are not key variables used for estimating counts of rental units, though they are useful in particular analyses.

Table 2: Variable Missingness and Imputation in Tax Assessment Records

	Records			*Parcel dataset (2005-2015)*
Variable	Non-missing	Imputed	Missing	Missing
Units	0.982	0.008	0.010	0.006
Land usage	0.978	0.022	0.000	0.000
Owner-occ	0.978	0.022	0.000	0.000
Year built	0.773	0.000	0.227	0.141
Parcel area	0.974	0.000	0.026	0.087
Total value	1.000	0.000	0.000	0.000
Land value	1.000	0.000	0.000	0.000

Note: Table 2 shows the proportion of observations in the input datasets (records) that were initially non-missing, were imputed, and remained missing, for seven key variables. The final column shows the proportion that remained missing in the constructed longitudinal dataset.

Operationalizing Rental Properties

Finally, tax assessment records rarely contain direct identifiers specifying which properties and units are for rent (Salganik 2017:24; Preis 2024). Typically, scholars have drawn on owner-occupancy tax exemptions, labeling single-unit properties without exemptions rental properties and deducting one unit from owner-occupied multi-family properties (Gomory 2022; O'Brien and Hangen 2024). However, single-unit properties without exemptions may be non-rentals (e.g., investment vehicles, vacation homes, or owner-occupied homes that did not file an exemption); single-unit properties with exemptions may be rentals (incorrectly filing an exemption); multi-unit properties with an exemption may not rent out the remaining units; and

some large multi-unit residential properties' may not be for rent because they are under construction or otherwise lying vacant. Accordingly, further efforts are necessary to ensure accurate rental unit counts (Preis 2024).

To ensure the tax assessment records had accurate rental unit counts, I calculated the number of residential and rental units in properties of different sizes for each county-year and tract-year and compared them to analogous counts from the ACS/Census.[13] As Figure 1 shows, the accuracy of the total unit counts is very high, with correlations of 0.99 at the county-year level for single-family units, multi-family units, and all units, and analogous correlations of 0.91, 0.89, and 0.90 at the tract-year level. However, the tax assessment records overcount rental units overall, driven largely by five of the 18 counties. This appears to be due to overcounting rentals in condominiums and single-family properties. I estimated a regression predicting the census rental unit counts using the counts for each land usage from the tax records (see Appendix D). This showed that tracts with more condominiums and single-family rental units tended to have much lower counts in the census, suggesting many of these rental units were actually not for rent. The counts for multi-family units appeared largely accurate.

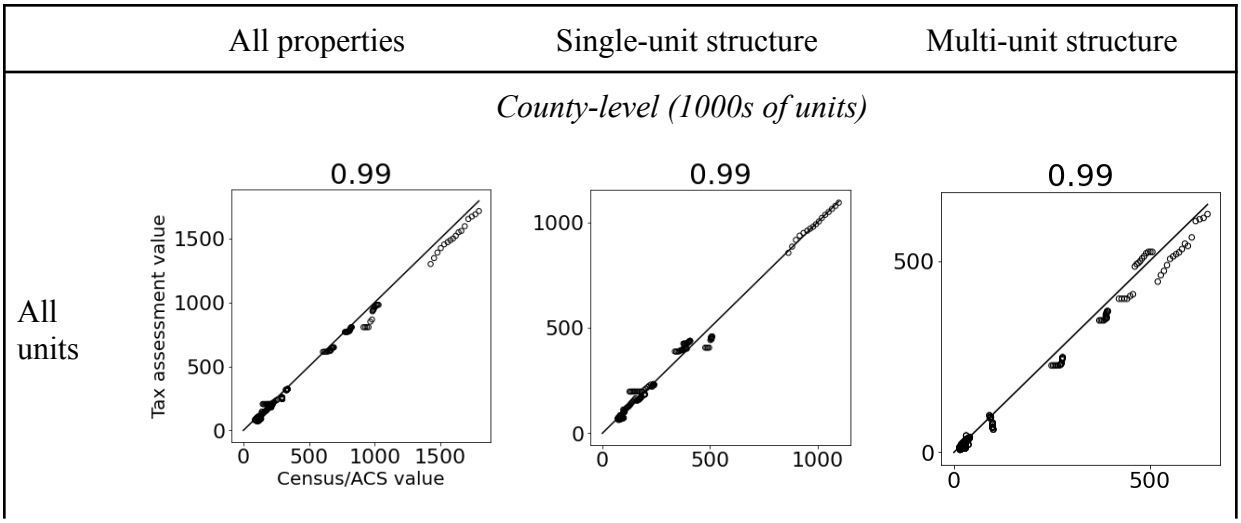
To account for these potential inaccuracies, I created alternate unit values adjusted to match the ACS/Census data. I grouped by granular property size categories and either inflated or deflated the counts to match those from the Census, such that the total adjusted rental units in each property type and tract-year matched analogous values from the ACS/Census.[14] That the

[13]The census reports unit counts based on the total units in the structure in which they exist, whereas tax assessments only detail the units by ownership. Accordingly, a condominium in a high-rise complex would be reported as being in a multi-unit property by the census, but as only a single-unit property in the tax assessment records. To account for this, I grouped condominiums by their street address, without the unit, to calculate the number of units in the structure. I used census data from 2000, 2010, and 2020, and ACS data from 2015 to 2019, linearly interpolating counts between years.

[14] For example, if a tract-year had 100 single-family rental units in my data, but 80 in the ACS/Census, an adjusted rental unit variable was created with 0.8 rental units for each single-family rental in that tract-year. I also created alternate operationalizations that grouped by larger and smaller bins of property sizes, for example comparing the counts for two-unit properties in the tax records to counts for two-unit properties in the Census/ACS, rather than

vast majority of dots in the scatterplots in Figure 1 are quite close to the identity line indicates that these adjustments were quite small, overall. Adjusting units ensures that tract-year level estimates account for bias created by inaccurate identification of single-family rentals, large multi-family developments under construction, and even high vacancy rates[15] for particular types of properties.[16] For example, since small-scale landlords are more likely to own small-scale properties, the overestimation of single-family rental units in the tax assessment records would downwardly bias estimates of landlord scale, but adjusting to match the ACS/Census addresses this potential bias.

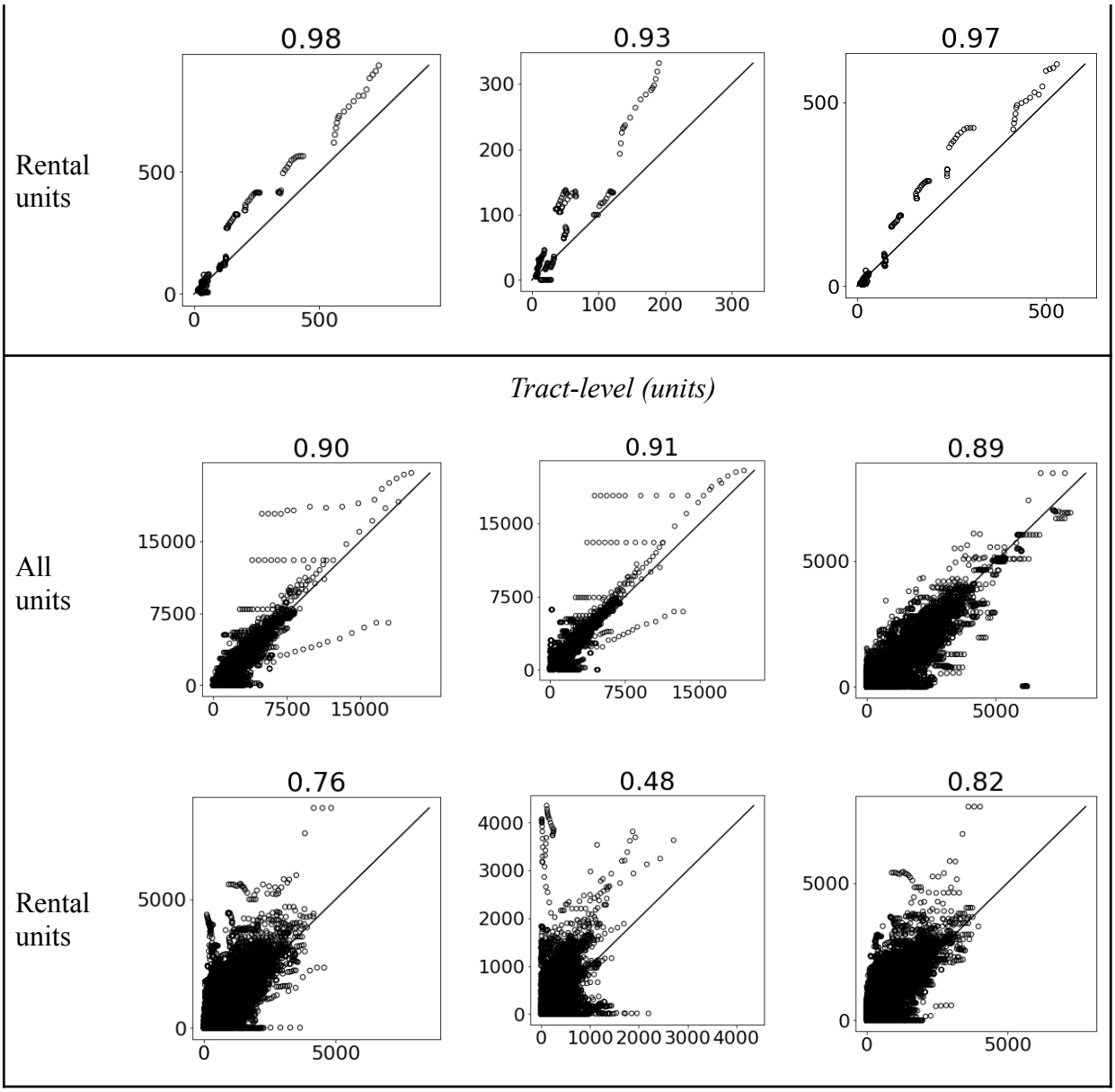
Figure 1: Comparison of Tax Assessment Unit Counts to Census/ACS Unit Counts



grouping all multi-unit properties together. Also, for each adjusted variable I created a second value that only adjusted halfway to match the Census/ACS data. This create five rental unit variables: the original count, a fully adjusted count based on coarse groups, a half-way adjusted count based on large groups, a fully adjusted count based on small groups, and a halfway-adjusted count based on small groups.

[15] This adjustment accounts for vacancies in different tracts and property types because the ACS/Census counts pertain to occupied rental units. Accordingly, if a particular tract has a large number of vacancies in single-family properties in a particular year, that would be reflected in the adjusted unit count.

[16] These adjustments account for bias only insofar as ACS/Census records are accurate, and the value being estimated at the tract-year level (e.g., a landlord characteristic) is not correlated with inaccuracies in unit counts conditional on tract, year, and property type.



Note: Figure 1 shows comparisons of unit counts from the unadjusted tax assessment records (x-axis) to counts from the ACS/Census (y-axis). In the first six plots, each dot represents a county-year of data and in the remaining six each point represents a tract-year. The rows distinguish counts of all units from counts of rental units alone, and the columns distinguish between types of properties. The number above each plot is the correlation between the two measures. Alternate unit counts were constructed (as discussed above) to make the tax assessment counts match those from the ACS/Census.

String Processing: Cleaning, Parsing, and Categorizing Names and Addresses

Cleaning and parsing names and addresses is key to many steps in the data construction process.

I distinguished names of companies and government agencies from those of people using a list of

keywords. For business names, I standardized the spellings of common words (e.g., development, properties) and corporate signifiers (e.g., corporation, partnership), and parsed the names into constituent parts. For example, "Foster Bates Realty LLC" was the owner of 9 Foster Street, a two-unit rental in Arlington, but the company may appear in other records as just "Foster Bates Realty," and there may be other affiliated companies with names like "Foster Bates Holdings" or "Foster Bates Properties." To identify connections to those entities, I created sub-names without business signifiers like "LLC" and "Incorporated," and without generic words like "Properties," "Holdings," and "Realty."

For person names, I removed honorifics and suffixes and corrected misspellings of common first names.[17] I then distinguished first and last names based on the structure of the name (e.g., "John R James" is likely to be first name, middle initial, last name, while "John, James R" is likely to be last name, first name, middle initial), a list of first name frequencies from the Census, and the predominant format in the remaining cases.

I split each address string into its constituent parts—the apartment number, street number, directional prefix, street name, directional suffix, city name, zip code, and country—and I identified instances where the address was incomplete. After linking entities (discussed below), I used a network-based address cleaning method that used information from all of the addresses that an entity is connected to to fill in missing information and make corrections (for details see Appendix E).

[17] In some cases, multiple people are listed as a single owner, such as "John and Jane Doe." These are first split into distinct names.

Reconciling Entities Between and Within Datasets

The tax assessment records provide substantial information about the physical and economic characteristics of parcels, but relatively little about ownership—only the legal owner's name and contact address. In about half of rental units in the data, the legal owner was not a person—for example, 9 Foster Street's owner in the tax assessment records was "Foster Bates Realty LLC"—giving little insight into who actually owns the property and what other properties they own.

I traced through anonymous shell companies like "Foster Bates Realty LLC" using business filings, which detail each company's officers, as well as other information like the principal address and registered agent.[18] Business filings have many of the same advantages as tax assessment records—they are complete populations of data,[19] contain real names and addresses, detail information at the company level, and are widely publicly available. However, they do not contain unique identifiers that allow the companies to be connected to the tax assessment records. Despite this problem, the business filings have an advantage that no other data set has, that they detail ownership for otherwise anonymous companies.

To incorporate the business filing data and more broadly to identify all of the instances where the same person or company was mentioned within the datasets required developing a unique identifier for each person and company. Commonly known as "entity reconciliation," this is a common challenge in using administrative records and an active area of research among data scientists (Elmagarmid et al. 2007; Enamorado et al. 2018). However, existing approaches were not well-suited to my data, because they typically require multiple fields on which to match entities and the tax assessments and business filings provide only a name and address. However, the records contain rich networked information, such as connections between people, companies,

[18] Officers of companies are not necessarily owners, although in most cases they are (see Appendix G.1)
[19] As mentioned above, two exceptions are banks and non-company, non-person entities like trusts.

and parcels, which existing methods cannot as readily take advantage of (for details see Appendix F.1).

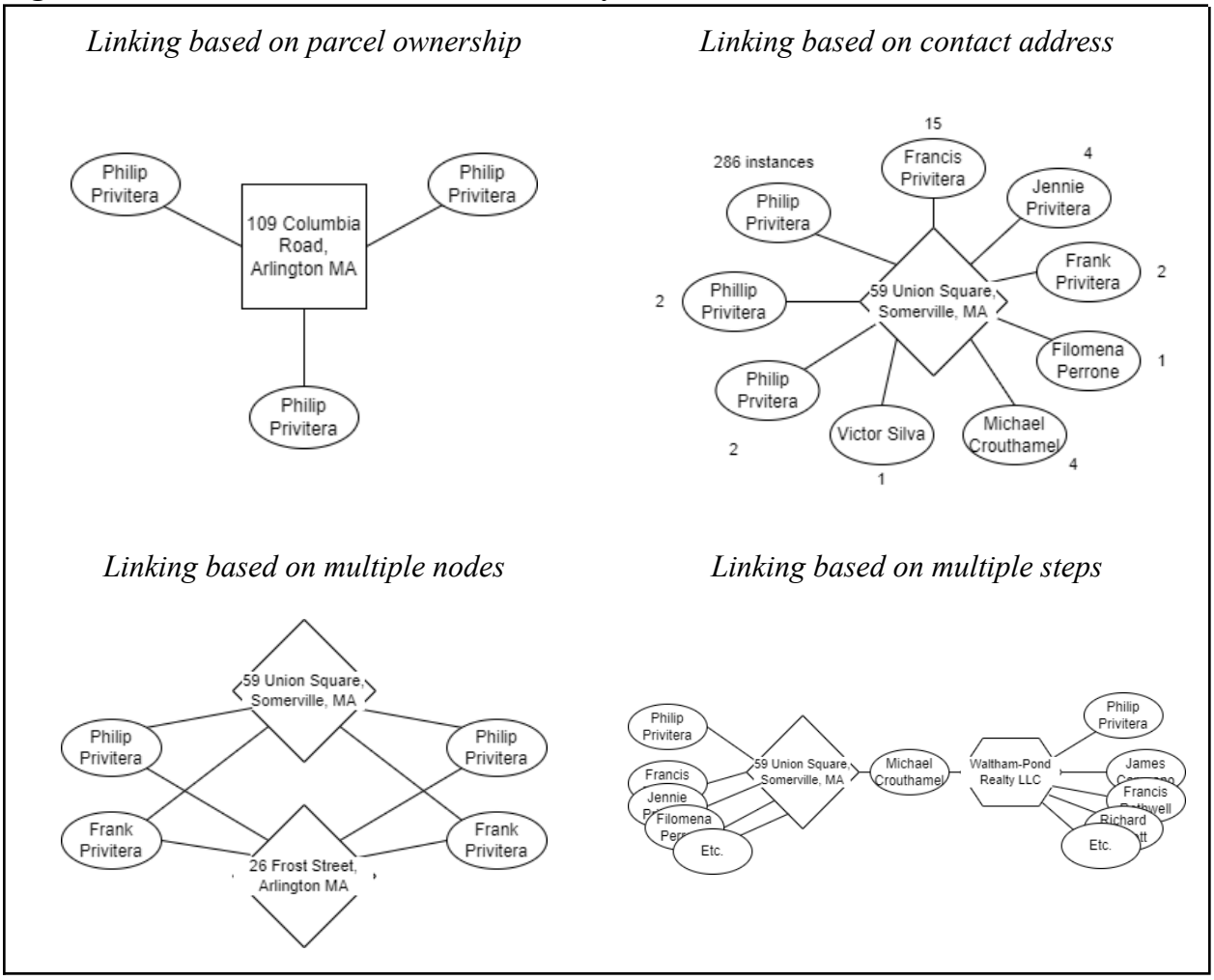
My network-based entity reconciliation method uses the connections between entities, such as owning the same parcel, being an officer for the same company, or listing the same contact address, to identify likely matches.. For example, Figure 2, Plot 1 shows that Philip Privitera owned a parcel at 109 Columbia Road, Arlington MA, which, in the two other years of data, was also owned by a person named Philip Privitera. The connection, in this case, was owning the same parcel, and the presence of three entities with the same name constituted strong evidence that these three were the same person. However, in some cases the nodes connecting entities were less discerning, which made the matches less certain. For example, Figure 2 Plot 2 shows that Philip Privitera also listed 59 Union Square, Somerville, MA as a contact address, but so did Francis Privitera, Victor Silva, and several other people. Because the number of connected entities was higher, the probability that two Philip Priviteras based on this connection were actually different people was higher as well.

Links between people can also consist of multiple shared nodes. Figure 2, Plot 3 shows that although many entities listed 59 Union Square, only Philip Privitera and Frank Privitera listed both 59 Union Square and 26 Frost Street as contact addresses. By reducing the number of total entities that were linked, considering two-node links allowed for greater confidence that the two Philip Priviteras linked in this way were the same.[20] Finally, multi-step links could also be used to identify matches, as detailed in Plot 4, where Philip Privitera linked to 59 Union Square, which linked to Michael Crouthamel, who linked to a company, Waltham-Pond Realty LLC, which linked to another Philip Privitera. Because this instance of Philip Privitera, the officer of

[20] Matches that involved multiple nodes often required other matches to be made first, since each initial record only had at most one contact address and one connection to a company or parcel (see Appendix F.5.).

Waltham-Pond Realty LLC, did not list 59 Union Square, he did not appear in Plot 2, and he would not have been combined with the others if not for the multi-node link.

Figure 2: Illustration of Network-Based Entity Reconciliation



Note: Figure 2 shows examples of links between entities that were used in the network-based entity reconciliation methodology. Ovals represent people; squares represent parcels; diamonds represent addresses; and hexagons represent companies.

I formalize this network-based entity reconciliation method mathematically in Equations 1 through 7. "i" and "j" index instances where two people appear in the records (whether tax assessments or business filings); M_{ij} indicates whether those two instances refer to the same person; N_{ij} indicates whether the instances' names match; and L_{ij} indicates whether the instances are linked in the network structure, in any of the ways detailed in Figure 2. Equation 1 defines

the probability that two instances are the same entity, conditional on a name match and link, using Bayes' Rule. This can be simplified, as shown in Equation 2, to an equation consisting of the probability a match is true, conditional on the two entities being linked through link "k" (β_k), and the probability that two entities that are not the same have a name match (α) (see Equations 3 and 4).[21] This formalizes the intuition that if a link is effective in subsetting to likely matches (meaning β_k is close to 1), and it is rare for two different entities to have the same name (α is close to 0), then the probability that entities are the same, conditional on them have a name match and being linked, should be high (for full details on the derivation, assumptions, and consequences of violations to the assumptions, see Appendix F).

$$P\left(M_{ij}=1\middle|N_{ij}=1,L_{ij}=1\right)=\frac{P(N_{ij}=1|M_{ij}=1,L_{ij}=1)P(M_{ij}=1|L_{ij}=1)}{P(N_{ij}=1|L_{ij}=1)} \quad 1)$$

$$=\frac{\beta_k}{\beta_k+(1-\beta_k)\alpha} \quad 2)$$

$$\beta_k = P(M_{ij}=1|L_{ij}=1) \quad 3)$$

$$\alpha = P\left(N_{ij}=1\middle|M_{ij}=0\right) \quad 4)$$

$$\beta_k=\frac{\sum_{i\neq j}\mu_k N_{ij}}{\sum_{i\neq j}1}=\mu_k A_k \quad 5)$$

[21] Several assumptions are made in this derivation. First, I assume that if two entities are truly the same person, they will have the same name and will share at least one link. I also assume that the probability of a name match between two entities that are not the same, but are linked, is the same as the probability of a name match between any two non-matching entities (Equation 3). While these assumptions are likely to be violated in a small number of cases, in Appendix F.3 I show that departures do not substantially bias my estimates. A full, detailed derivation is provided in Appendix F.2.

$$\mu_k = \frac{A_k - \alpha}{A_k - A_k\alpha} \quad (6)$$

$$A_k > \frac{\alpha}{1 - \mu_k + \mu_k\alpha} \quad (7)$$

I estimated α, the unconditional probability that two different individuals have the same name, to be 2.3 x 10^{-6}, using public datasets of first and last name frequencies (see Appendix F.4). μ_k, or how effective a link is in identifying true matches, is different for each link k, since addresses, parcels, and other nodes can be more or less discerning, as demonstrated in Figure 2. I estimated μ_k as the fraction of all connections created by link k that are name matches (A_k) times the probability a linked name match is a true match (μ_k) (see Equation 5). I was then able to calculate a closed-form solution for μ_k (Equation 6), set a minimum value for μ_k (true positive rate), and calculate a minimum value for A_k. The minimum value of A_k in order to have a true positive rate of 0.999 is approximately. 0.0023, meaning that 0.2% of all pairwise connections between instances based on link k have the same names.[22] I then calculated A_k for each link and determined which were valid for linking entities. I applied the same framework to fuzzy matching, defining N_{ij}, a name match, as a fuzzy match. This meant that α, the probability that two different people have a name match, increased, which entailed that the links needed to be more discerning, or A_k needed to be higher, in order to ensure the same true positive rate (see Appendix F.4. for details).

[22] That this threshold appears relatively low, meaning links between entities can successfully identify true matches even when most of the linked entities are not the same, reflects that it is relatively rare for two different people to have the same name.

Based on the above calculations, I linked all person entities that had a fuzzy or exact name match and were linked by an address, parcel, a connection to another entity, or a multi-edge link (like that in Figure 2 Plot 4) with an A_k threshold ensuring a true positive match rate of 0.999 or greater (for details see Appendix F.5). Based on these rules, I matched Philip Privitera to 309 other instances where he appeared in the business filing and tax assessment records.

I matched companies using an analogous procedure, drawing on networked connections and similar names. However, matching companies differed in three ways. First, companies are required to have unique names within a state, which allowed me to match any two companies that shared a full name. Second, company names were frequently misspelled, abbreviated, or written in multiple ways, which meant that matching on exact names alone would miss many matches. However, using the same types of fuzzy matches that I used for people was impeded by the difficulty of estimating α, the probability that two different companies would have similar names. Thankfully, the third difference, that all non-bank companies are required to make business filings, allowed for inexact matches between companies in the tax assessment and business filings, as long as they were unique. For example, "132 Chelsea Street" appeared in the tax records, but because it did not include a business signifier, it could refer to "132 Chelsea Street Inc." "132 Chelsea Street LLC," or "132 Chelsea Street LP," each of which could be a legally distinct entity within the business filings. However, if only "132 Chelsea Street LLC" shared an address with "132 Chelsea Street," then I could be fairly certain that this was the company referred to in the tax records (for a detailed discussion of company matching rules, see Appendix F.6.).

Table 3 shows that 78% of all companies, weighted by the rental units they owned, matched to business filing. This proportion is lowered by the presence of entities like banks and trusts that do not make filings, and among LLCs, corporations, and partnerships, all of which should have filings and collectively owned 88% of non-person-owned rental units, 87% matched to business filings.

This network-based entity reconciliation method has several benefits compared to existing approaches. First, it allows one to draw on networked information, from two entities sharing the same address to complex, multi-node links like the one in Figure 2, Plot 4. Second, it allows one to accurately model that different links convey different amounts of information. For example, a home address might be highly significant for determining matches, but a connection to a large commercial building much less so. For a discussion of how this differs from conventional approaches see Appendix F.1.

Table 3: Success of Entity Reconciliation for Property-Owning Companies

Company type	Proportion of non-person-owned units owned	Proportion of entities matched to filings (weighted by rental units)
All	1.0	0.78
LLCs, corporations, partnership	0.88	0.87
Banks, trusts	0.08	0.05
Other	0.05	0.20

Note: Table 3 shows the proportion of non-person-owned rental units that were matched to business filings for different types of owners. The first column shows the proportion of units that were owned by this type of entity and the second column shows the proportion of those units that were matched to filings. Accordingly, the second row indicates that LLCs, corporations, and partnerships own 88% of non-person-owned rental units, and 87% of that 88% were matched to business filings.

Incorporating Additional Datasets (Example: Eviction Filings)

The data construction pipeline allows easy incorporation of additional datasets, such as housing code violations, building permits, and criminal incidents. As an example, I incorporated eviction filings using the address at which the eviction was filed and the name of the plaintiff to link them to other records.

First, I included plaintiffs mentioned in the eviction filings in the entity reconciliation process, allowing plaintiffs to be linked with instances where they were mentioned in the tax assessments or business filings. As Table 7 shows, 49% of eviction plaintiffs were linked to a business filing and 44% to an entity in the tax records. In many cases, plaintiffs do not appear in the business filings because they are not companies or do not appear in the tax records because a different entity than the owner, such as a lawyer, property manager, or other shell company, filed the eviction.

I geocoded filings to the parcel dataset using three types of matches. First, I used standard geocoding techniques that matched eviction addresses to parcel addresses, based on components like street numbers, street names, and zip codes. Table 4 shows that this method successfully geocoded only 58.6% of eviction filings. This low match rate occurred because tax assessment records typically contained only a single address for each parcel, even when the parcel spanned multiple street addresses, meaning eviction filings that occurred at those addresses not listed in the tax assessment records failed to match. To address these instances, I identified properties whose owners shared the plaintiff's name. For example, one of Privitera's companies, Arlington Fremont Realty Inc., filed an eviction at 35 Fremont Street, but geocoding based on address alone failed because the eviction listed 02474 as the zip code while the parcel listed 02476. However, I identified that the parcel record had Arlington Fremont Realty Inc. as its owner,

allowing me to connect the eviction filing to the correct location. These name-based geocoding techniques matched an additional 3.2% of filings. Finally, I used the latitude and longitude in the eviction filing records to overlay them onto the parcel shapefiles, geocoding an additional 29.6% of filings, for an overall rate of 91.4%.

Table 4: Incorporation of Eviction Filings into Other Records

	Proportion of filings
Entity reconciliation	
Linked to business filing	0.492
Linked to tax records	0.438
Geocoded	0.914
Address geocoded	0.586
Name geocoded	0.032
Geographic geocoded	0.296

Note: Table 4 shows the proportion of filings where the entity was linked to another entity in the business filings or tax assessments and the proportion where the filing was geocoded to a parcel.

Identifying Landowner Conglomerates

Large-scale landowners typically consist of numerous companies and employees, and even small-scale owners often own properties through multiple shell companies (Travis 2019; Gomory 2022). Operationalizing landlord characteristics like their scales of operations and geographic locations requires identifying all of the entities that constitute these conglomerates. Studies in economics and economic sociology have identified corporate networks in order to analyze topics like interlocking board networks and corporate connections between countries (Zeitlin 1974; La Porta et al. 1999). However, these studies typically draw on filing data from the Securities and Exchange Commission (SEC), which have detailed information on subsidiaries and shareholders, but are only available for large, publicly traded corporations and not for private companies, which include the vast majority of landowners. Several landlord studies have drawn on these SEC filing data to operationalize corporate networks for a small number of very large owners

(Chilton et al. 2018; Raymond et al. 2018), but doing so for the broader population of landlords requires alternate methodologies.

I determined landowners' corporate networks by identifying a range of connections between entities that suggested they were part of the same conglomerate. These connections only "suggested" being in the same conglomerate because unlike the subsidiary relationships detailed in public filings, they were not definitive. The primary connection I used was company officership, as detailed in the business filings. However, officership alone was insufficient, first, because it was missing for trusts, banks, and companies not matched to business filings. Furthermore, important connections like being family members, using the same uncommon address, or having similar names between companies and people, were not identified using officership. Figure 3, Plot 1 shows the connections identified by officership, where the circle in the middle indicates Philip Privitera, the surrounding squares indicate companies he owned, and the edges indicate officership. The cluster of people and companies in the lower-left corner indicates entities that were part of the Privitera conglomerate but were missed when using officership alone.

I used several types of connections beyond officership to identify which entities were part of the same conglomerate. For example, 89 Forest Street Realty Trust did not link to a business filing, but it listed 59 Union Square as its contact address, the same address listed by all of Philip Privitera's companies, which suggested that Privitera owned it. However, connecting based on addresses in this way is legitimate only if the address is unique to the conglomerate and not the location of a third-party, like a resident agent, or of multiple conglomerates, like a corporate park. I identified unique addresses based on the degree of uniformity among other companies that listed the address, for example calculating the proportion of associated companies with the

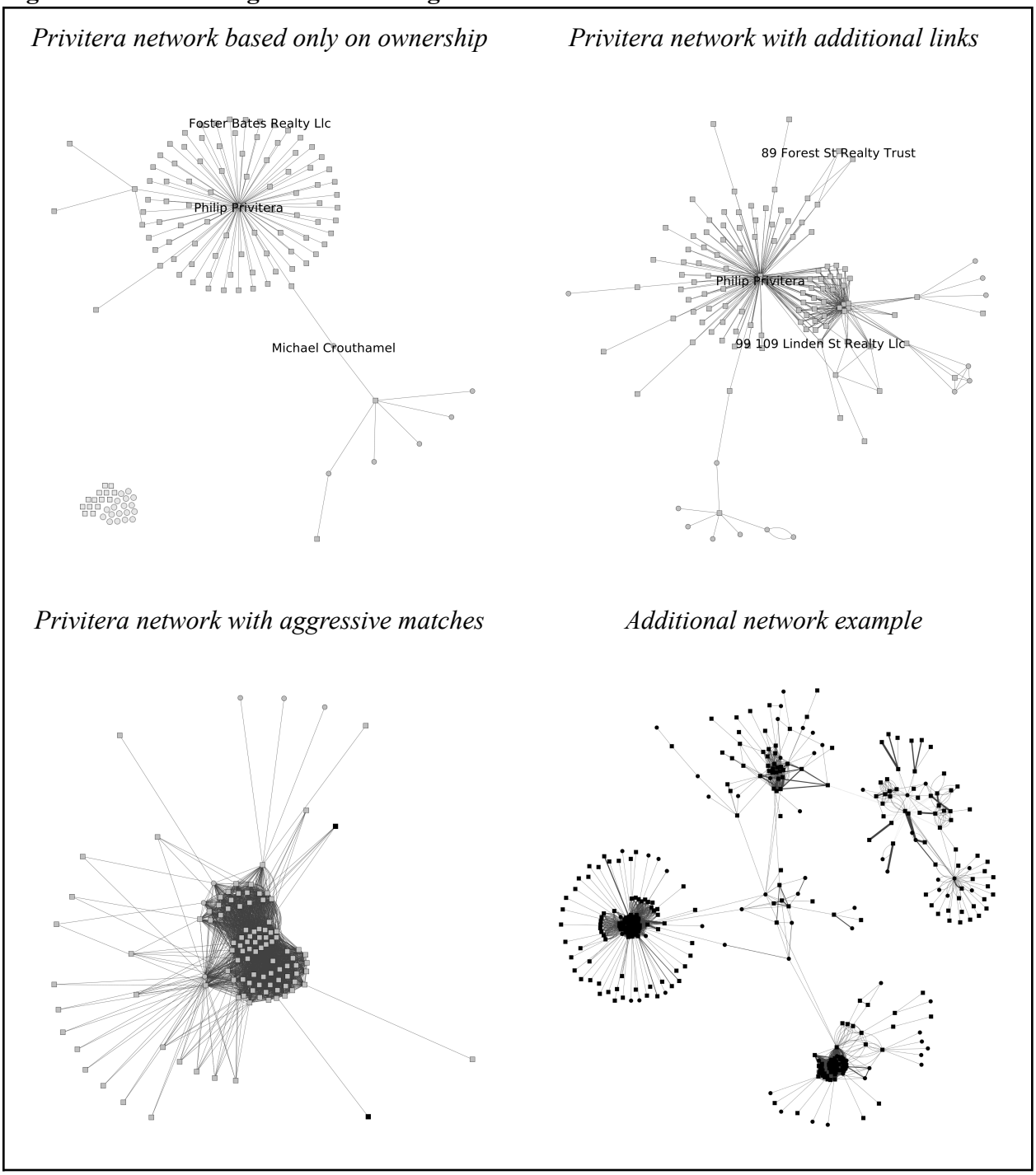
same officers.[23] Landlord conglomerates frequently contain multiple family members, and, generally, ownership by family members is often used to obscure and distribute ownership (Bacon 2013a; 2013b).[24] I identified family members by finding instances where two people shared a last name and a connection to a sufficiently discerning[25] address, parcel, or other entity. In this way, I linked Philip Privitera with Francis Privitera and Jennie Privitera. I also connected people to companies that had similar names, which was particularly useful when trusts contained the names of the beneficiaries. Finally, I identified instances where companies shared a connection to a parcel, address, or other entity and had similar names (for a detailed list of all connections used in identifying landowner conglomerates, see Appendix G.3). Figure 3, Plot 2 shows the Privitera conglomerate's full network of links.

[23] For example, I deemed 59 Union Square legitimate because, of the 74 companies that listed it as a principal address, 73 had the same officer, namely, Philip Privitera. However, some other companies in the network listed 175 Paramount Drive, Raynham, MA, which is a corporate park, as a contact address. Of the 17 companies with 175 Paramount Drive as a principal address, only 7 had the same officer, so matches based on this address were omitted. For details on determining legitimate and illegitimate addresses, see Appendix H.

[24] Many US laws consider assets owned by an immediate family member in the same way they do assets by the principal individual (Cornell).

[25] I follow analogous rules for identifying legitimate nodes on which to link families as used in entity reconciliation (see Appendices G.3 and G.4).

Figure 3: Constructing Landlord Conglomerate Networks



Note: Figure 3 shows the networks of connections out of which landowner conglomerates were constructed. Plots 1 through 3 detail steps in the construction of the Privitera conglomerate, and Plot 4 shows an additional representative example.

These networks inevitably contained false positives, or connections between entities not in the same conglomerate. For example, the lower-right corner of Plot 1 shows several

companies owned by Michael Crouthamel that were not part of the Privitera conglomerate but were connected through a company that Crouthamel and Privitera co-own. I attempted to sequester these false connections by identifying the true connections in multiple ways, such that the true connections formed dense clusters and the false connections were spindles between them (see Figure 3, Plot 4). For example, if Michael Crouthamel's companies had shared addresses with Privatera's they would have formed a denser cluster, but because they didn't, they formed a spindle. A second way I attempted to overweight true matches was by conducting aggressive matches within small partitions of the network. For example, I connected any two people within the network in Figure 3, Plot 2, who had the last name Privitera or the contact address 59 Union Square. Matching this aggressively across the entire dataset would create too many false positives, but doing so within a partition helped to consolidate conglomerates into dense clusters. Figure 3, Plot 3 shows the Privitera conglomerate after aggressive name and address matches were included, revealing a much denser cluster of entities. Finally, I identified the dense clusters that constituted conglomerates using community detection, specifically a modularity-based algorithm that maximizes within and minimizes between-cluster edges. Plot 4 shows an additional example of the networks out of which conglomerates were identified, to give a better sense of the range of corporate networks present in the data and the types of cluster-and-spindle structures I attempted to identify.

The conglomerate construction process presented a large number of analytic decisions to be made, and without a ground truth to compare results to, there was no definitive way to choose the best approach (see Table 5 and Appendices G.3 and G.4 for a more detailed discussion of these decisions). In light of this ambiguity, I followed Nelson's (2019) admonition that new computationally derived measures should be transparent and reproducible, in the following way.

I hand-checked several hundred randomly sampled cases to develop a set of reasonable possible implementations, then created separate versions of each conglomerate according to each implementation (see Appendix G.4. for a full list of implementations). Using this range of operationalizations, I can test how much different conglomerate construction decisions influence particular estimates.

This approach improves upon existing landlord conglomerate identification methodologies by leveraging a wider range of possible connections between entities (e.g., address-based matches, family member matches), using more sophisticated matching techniques (e.g., removing likely third-party addresses, conducting aggressive matches within network partitions), and implementing these matches in multiple ways to allow for a range of possible results (see Appendix G.3).

Table 5: Analytic Considerations to Be Made During Landlord Corporate Network Construction

Consideration	Description
Ownership matches	Companies unlikely to be realty-related can be removed; officers unlikely to be owners can be removed; resident agents can be considered likely officers in some circumstances.
Matches based on company name	Name matches can be made with higher and lower fuzzy thresholds and with higher and lower discernment thresholds for nodes connecting companies.
Matches based on shared addresses	Criteria for determining legitimate addresses can be more and less stringent; address matches can be used only for those companies without officer matches; only principal addresses can be used, or contact addresses can be included as well.
Family member matches	Thresholds for the connecting nodes for family members can be more or less discerning.
Aggressive matches within small partitions	More or fewer words and addresses can be used in aggressive matches, and different words and addresses can be filtered out.
Sample of entities to include in construction	Only property-owning entities, property-owning entities and likely realty-related entities, or all entities can be included.
Network resolution and weighting decisions.	Different weights can be attached to different types of links between entities and different penalization parameters can be chosen for community detection.

Note: Table 5 details seven analytic decisions to be made when constructing landowner conglomerates, such as whether to include non-property-owning companies and how large of a penalization parameter to use in community detection. For a detailed discussion of these decisions and the different implementations I used, see Appendix G.3.

Creating Measures of Landowner Characteristics

Scale

Scale of ownership has been the most common characteristics used to classify landowners, with scholars distinguishing "mom and pop" from large-scale owners and showing differences in their rent-seeking strategies (Molotch and Logan 1987), eviction practices (Gomory 2022; Balzarini and Boyd 2020), maintenance procedures (Sternlieb 1966; Stegman 1972), tenant screening practices (Rosen et al. 2021; Decker 2023), and other behaviors. Although landlord scale is rarely the direct causal reason for differences in behavior, it is still an immensely useful heuristic for distinguishing owners. I operationalized scale by calculating the number of properties, number of residential units, and total value of properties owned by each conglomerate in each year. For example, the Privitera conglomerate owned 1,251 units per year between 2005 and 2019, through 63 companies and 4 people.

Corporate Formality

Corporate status has also been a popular characteristic used to differentiate landowners, with academic and popular reports frequently discussing corporate versus non-corporate owners. However, the precise meaning and import of "corporateness" are frequently ambiguous, in some cases referring to the liability protection provided by LLCs and corporations (Travis 2019), in others the formalization of business practices (Gomory 2022), and in still others as a stand-in for large-scale or institutional owners. However, ownership through an LLC does not necessarily

convey anything other than liability protection, and operationalizing these characteristics in a more precise, transparent, and consistent way is essential.

To capture differences in legal regulations like personal liability, I first operationalized the legal type (e.g., LLC, corporation) of each non-person entity using their name and, if they had one, business filing. To operationalize corporate formality, I calculated the number of distinct officers and officer titles in each company, since informal companies are often owned by a single person and a fully articulated employee structure of president, secretary, etc. suggests greater corporate formality. I also created a distinct indicator of whether the company had a director, which indicates the presence of a board, and whether the company is a corporation (rather than LLC, partnership, etc.), since corporations are more often used by formal companies. Finally, I determined whether the company's principal address was an office, rather than a home (distinguishing homes from offices as discussed below). In many cases, LLCs and other informal corporate structures were used as intermediaries, owning properties that were in turn owned by companies that showed more signs of corporate formality. To operationalize the corporate formality across a conglomerate as a whole I identified the company that was the ultimate owner (i.e., including ownership through intermediary companies) of the greatest amount of property, labeled it the focal company, and used its formality characteristics.[26]

For example, Foster Bates Realty was an LLC with only a single officer, Philip Privitera, no distinctions between titles, and no director, but its principal address, 59 Union Square, was a commercial property. I summed these indicators of corporate formality (being a corporation, having multiple officers, having distinct titles, having a director specifically, and having a commercial primary contact address) into an aggregate corporate formality score with each

[26] Corporate formality measures can also be aggregated across all companies that constitute a conglomerate, but using the measures from a single focal company simplified estimation of possible bias (see below).

indicator scored as zero or one,[27] for which Foster Bates Realty scored a one one out of five, indicating low corporate formality. Across companies, the Privitera conglomerate did not show high formality, and its focal company also showed a score of only 1.

Geographic Variables: Identifying Home and Business Addresses

Studies frequently highlight landowners' geographic locations, distinguishing between distant, absentee owners, and local, or even co-resident owners (Mallach 2014; Gomory 2022), showing that geographic distance is associated with differences in landlords' relationships with tenants, maintenance practices, and eviction frequency (Sternlieb 1966; Travis 2019). Having a corporate office outside the metro area or state also suggests the landowner operates on a regional or national scale and likely owns more properties than just those identified in the data. Finally, identifying individuals' home addresses and neighborhoods provides insight into their socioeconomic positions and demographic characteristics. However, studies often operationalize landowner location using the contact address listed in tax assessment records, without considering all of the addresses listed by members of the conglomerate or distinguishing between addresses that refer to landlords' homes, corporate offices, rental properties, and third-parties like resident agents.

To distinguish between different types of addresses, I geocoded all contact and principal addresses[28] to the property tax records and to census blocks. I first identified as homes any residential address with an owner-occupancy tax exemption,[29] which found home addresses for 32% of rental units owned by people or by companies owned by people (see Table 6). This did

[27] I created binary measures for officers and titles assigning one to companies with more than one officer and title, respectively.
[28] In business filings, companies list a principal address, which is their primary business location, but in practice this may be their home address or a third-party address, so I treat it in much the same way I would a contact address.
[29] If multiple such addresses existed I chose the one listed most recently or most often listed as a contact address.

not fully identify home addresses, however, because an individual's home may be missing an owner-occupancy exemption because it is outside of the metro area, not geocoded to the tax assessments, or for another reason.[30] This raised the problem of distinguishing which contact addresses referred to homes and which referred to rental properties, third-parties like resident agents, and business locations. I considered an address a "likely home" if it was a small residential property, if it was listed as a contact address (rather than being an owned property not listed as a contact address), if few or no other people listed the same address as a contact address, if the person did not list other properties as a contact address, and if it was in a highly residential Census tract. A full description of the rules for determining home addresses is provided in Appendix H.

Based on these criteria, I identified likely home addresses for an additional 43% of person-owned units, and ambiguous but possible home addresses for an additional 5%. In the remaining 20% of cases,[31] no home address could be identified, typically because the person listed only office addresses, rental addresses, third-party addresses, PO boxes, or incomplete addresses. I also identified the primary business location for each conglomerate as a whole, whether that was a distinct corporate office or the home of a prominent member, by omitting all third-party and rental property addresses and identifying the modal address among the remaining.

For example, Philip Privitera listed contact addresses 310 times throughout the tax assessment and business filing records. 95% of them were 59 Union Square, which, because it was a commercial parcel, was labeled as an office. In three instances, he listed 26 Frost Street,

[30] The tax assessment unit count validation above (Figure 1) demonstrated that owner-occupancy undercounts non-rental properties, suggesting that some non-owner-occupied properties were in fact the owners' homes, and considering all properties without owner-occupancy exemptions to not be homes would likely miss many individuals' home addresses.

[31] Appendix I describes the characteristics of observations with missing data, showing that observations missing or with ambiguous home data were more likely to be non-person-owned, multi-family rental units.

which was a single-family property but not owner-occupied and so was labeled as a likely rental property. In 15 instances he listed 59 Winchester Road, an owner-occupied, single-family property that he owned, which was labeled as his home address.

Table 6: Identification and Categorization of Addresses

	Proportion of rental unit-years
Person home address identified?	
Certain home	0.32
Likely home	0.43
Ambiguous	0.05
No home identified[32]	0.20
Rental unit-years	39,496,843

Note: Table 6 shows the proportion of rental unit-years that were owned by people (either directly owned or owned through a company) for which certain homes, likely homes, ambiguous homes, and no homes could be identified. The precise rules for determining home addresses are detailed in Appendix H.

Identifying Racial Identities of Landlords

Race is central to the functioning of American housing markets. Most American cities are highly racially segregated (Massey and Denton 1993); neighborhood racial composition is a key factor in how properties are valued and managed (Taylor 2019; Robinson 2021); and the racial identities of housing market intermediaries shape their relationships and behavior, for example by affecting their access to capital (Satter 2010; Korver-Glenn 2018). Accordingly, identifying landowners' racial identities allows for a range of important analyses, including studies of discrimination in rental markets and the legacy of racialized processes like redlining.

I estimated the racial identities of landlords using Voicu's (2018) approach, which draws on an individual's first name, last name, and the racial composition of their home neighborhood. This and similar methods have been used widely and shown to have a high degree of accuracy

[32] Characteristics of observations missing and not missing homes (as well as other measures) are discussed in Appendix I.1.

(Imai and Khanna 2016). First and last names can effectively distinguish Hispanic (all other racial groups are non-Hispanic) from Asian from Black/White individuals, and first names and the racial composition of home addresses' neighborhoods are effective in distinguishing Black from White individuals. Accordingly, it was essential to ensure that the address used in this procedure was the home address, and not merely a contact address which may correspond to a rental property or business location. I used the 2010 Census surname list, which covers 90% of the US population, to estimate racial composition by surname, and I used mortgage application data collected by Tzioumis (2017) to estimate racial composition by first name. I used block-level racial composition data from the 2010 Census, where block-level identifiers were available for an individual's home, and tract-level or zip-level racial composition elsewhere.

For example, Philip Privitera's home address was on a block with 100% White residents. 92.9% of people named Philip are White, and 94.4% of people with the last name Privitera are White. As a result, the method estimated there was a greater than 99% chance that Philip Privitera was White.

Table 7 shows the distribution of posterior probabilities for the racial identity of each person, weighted by the rental unit-years they own. In 73% of cases, the resulting probability was above 0.85, and in an additional 23% it was between 0.50 and 0.85. Where home addresses could not be identified, or first or last names did not appear in the first name and surname lists, I drew on the most data that was available. For example, if no home address was identified, I used only the first and last names. Table 7 shows that 53% of cases used all three sources of data, 24% used one name and geographic data, 13% used two names, and 10% used only one piece of information.

Table 7: Criteria for Estimating Racial Identities and Resulting Posterior Probabilities

	Proportion of rental unit-years
Posterior probability	
>0.85	0.73
0.50-0.85	0.23
<0.50	0.03
Method	
First, last, geography	0.53
One name, geography	0.24
First and last	0.13
One piece of information	0.10
Unit-years	39,496,843

Note: Table 7 shows the proportion of rental unit-years that were owned by people (either directly owned or owned through a company) for which different racial imputation criteria were used and the resulting posterior probabilities.

Estimating Bias from Missing Data, Imperfect Data, and Multiple Specifications

Framework for Estimating Bias from Missing and Imperfect Data

Constructing landowner measures, which frequently involve multiple sources of information that are not uniformly available across observations, inevitably entails that some measures are missing or imperfectly constructed for some observations. For example, for conglomerates where I could not identify a person owner, landlord race data was missing, and for those where I could identify a person but not a home address, landlord race was nonmissing but created without geographic information. Even seemingly straightforward measures like scale, which was nonmissing for all owners, was only as accurate as the matching of the constituent entities within the conglomerate. Simply omitting missing and low-quality observations is unfeasible, since data quality is strongly correlated with characteristics like corporate ownership, and dropping these observations would introduce considerable bias. Frequently, studies that create measures from

administrative sources justify their resulting constructs by showing that they are the best measures possible given available data, but these best-possible measures may still be biased to an unacceptable degree. In this study, I directly quantify the prevalence of missing and imperfect data and the degree of resulting bias for different constructed measures.

Equation 8 shows how the bias for a given variable, or the expected difference between O_i, its operationalized value, and T_i, its true value, can be disaggregated between observations with high data quality, low data quality, and missing data. Q_i indicates the observation's data quality, with H, L, and M indicating high, low, and missing, respectively. I define L_i as the measure operationalized under low data quality (definitions of data quality for each measure are provided below), and I_i as the measure imputed from a random forest model drawing on a landowner's characteristics and property holdings (for details on imputation, see Appendix I.2). For low data quality observations (meaning $Q_i = L$), O_i is equal to L_i, and for missing observations (meaning $Q_i = M$), O_i is equal to I_i (see Equation 9).

I calculated the bias among low data quality observations by first identifying observations that have high data quality for that measure and re-calculating what the measure would be under low data quality ($L_i|Q_i=H$).[33] Because the expected bias may differ for those observations with high and low data quality, I re-weighted these high data quality observations such that they resembled the low data quality in terms of landlord characteristics, using random iterative method reweighting (for details see Appendix I.4). I then calculated the difference between the high and low data quality operationalizations (L_i-T_i) for the reweighted sample (see Equation 10). I calculated the bias among missing observations by subsetting to a test set of high data

[33] Methods for calculating low data quality versions differed for each measure. For landowner race, I simply re-imputed the race using less information. For scale, however, the process was more complex as I had to remove certain pieces of information, re-estimate the resulting conglomerates, and re-calculate the conglomerate scale (for full details on the construction of low data quality versions see Appendix I.3).

quality observations that were not used in fitting the random forest imputation model, reweighting them to resemble the observations missing data, and calculating the expected difference between the imputed and true data (Equation 11).

$$E[O_i - T_i]$$
$$= E[O_i - T_i|Q_i = H]P(Q_i = H) + E[O_i - T_i|Q_i = L]P(Q_i = L) + E[O_i - T_i|Q_i = M]P(Q_i = M) \quad (8)$$

$$= E[L_i - T_i|Q_i = L]P(Q_i = L) + E[I_i - T_i|Q_i = M]P(Q_i = M) \quad (9)$$

$$E[L_i - T_i|Q_i = L, X_i] = E[L_i - T_i|Q_i = H, X_i] \quad (10)$$

$$E[I_i - T_i|Q_i = M, X_i] = E[I_i - T_i|Q_i = H, X_i] \quad (11)$$

Bias from Missing and Imperfect Data

Figure 4 shows the results of these bias calculations for several key landlord measures, detailing the proportion of observations with low quality and missing data, the average bias per-unit for observations with low quality data and missing data, and the aggregate bias from both sources.

First, whether the conglomerate's focal company had multiple officers—an indicator of organizational formality—was missing for those conglomerates whose focal company was not successfully matched to a business filing, constituting about 16% of rental units. Based on the reweighted test dataset, the imputed values for those observations overstated ownership by officers with more than one unit by 0.56 per unit, creating an aggregate bias of 0.08, meaning it overstated ownership by organizationally complex landlords, by eight percentage points. However, the imputation was more accurate for predicting whether organizations had more than two officers, which had an aggregate bias of only 4 percentage points.

Landowner scale was never missing, but was labeled low quality if a conglomerate was insufficiently matched. I labeled conglomerates as having high-quality scale data if at least one of its companies was matched to a business filing,[34] and I estimated the scale measure that would arise under low data quality, for those observations with high data quality, by reconstructing their conglomerates as if they were missing all business filing data. 10.1% of units were owned by conglomerates with low quality data, and I estimated that having low quality data decreased the per-unit proportion of owners with 100 or more units by 0.40. This suggests that insufficiently linking entities decreases the aggregate estimate of owners with 100 or more units by about four percentage points.

The home location of the central[35] person-owner for a conglomerate was labeled missing if no person was identified or an identified person-owner listed only PO boxes, rental properties, third-party addresses, and commercial addresses. It was labeled low-quality if the person listed only addresses that were labeled ambiguous. 31% of units were owned by landlords with missing home data, and 5% were owned by landlords with ambiguous home data. Both sources of bias were minimal, with per-unit bias (in terms of identifying out-of-state home addresses) estimates of 0.006 and 0.001 for low-quality and imputed data, respectively, creating almost no aggregate bias.

The primary contact location for the conglomerate was missing if the conglomerate only listed rental properties, third-party addresses, or PO boxes as contact addressees, and it was labeled low-quality if the conglomerate only listed likely, rather than definitive, homes and business locations. 41% of units were owned by landlords with low-quality primary location

[34] In alternate analyses I only labeled observations with half or all of units as high quality, and I found similar results.

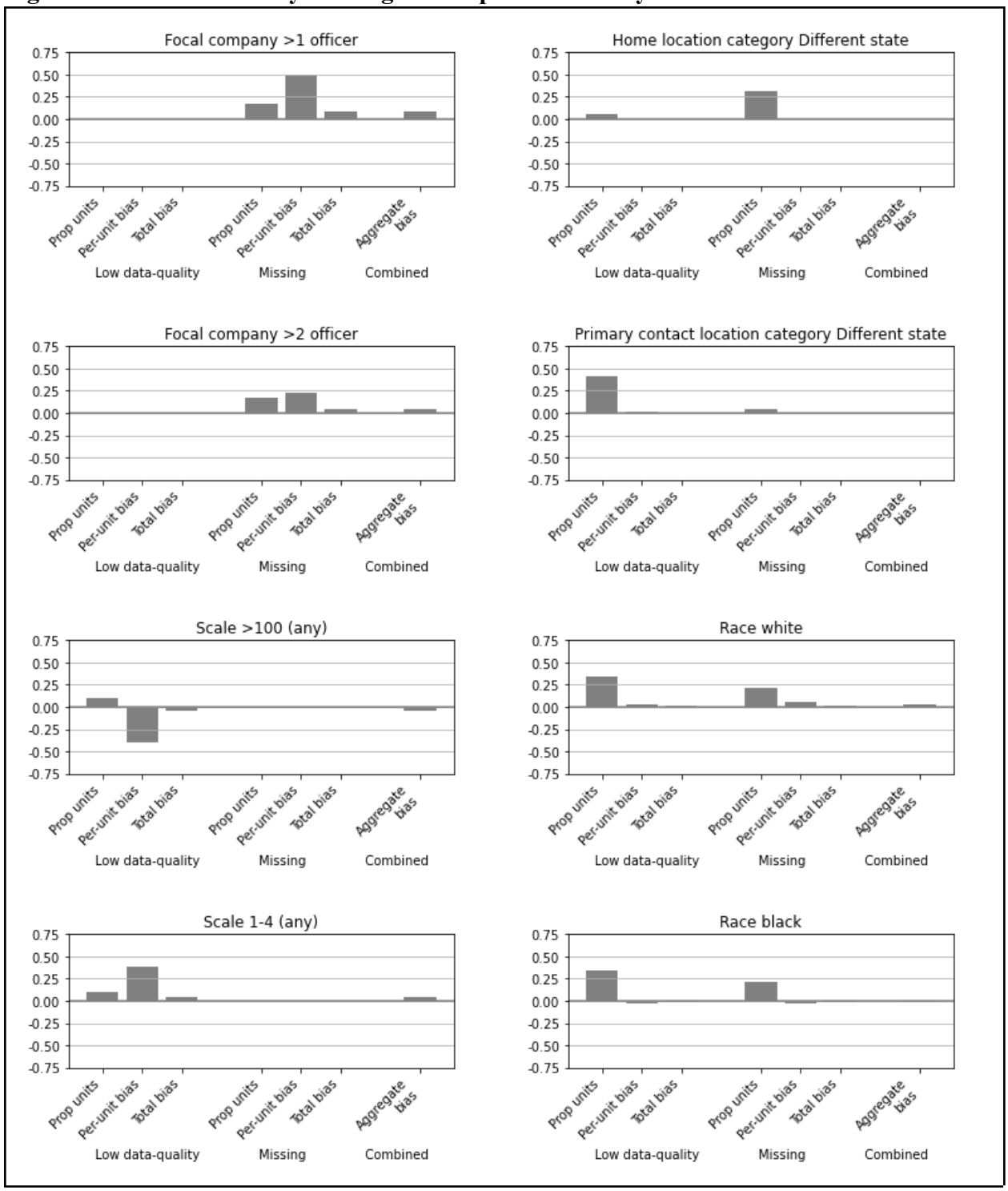
[35] Analogously to identifying a central company, I identified a central person for each conglomerate as the person who owned the most property, whether directly or indirectly through companies.

data, and on average they were 1.2 percentage points more likely to list an out-of-state address compared to their high data quality versions. 3.8% of units were owned by landlords with low-quality primary location data, and on average they were 0.00% more likely to list an out-of-state address. This created an aggregate bias of only 0.5 percentage points.

Finally, landlord race data was missing if no person-owner was identified or they had insufficient data to impute their race and was low-quality if a person was identified but the full range of information (first, last, home) was not available to impute race. 35% of units were owned by landlords with low-quality race data, and on average they were 3.3 percentage points more likely to be identified as White and 3.0 percentage points less likely to be identified as Black, per unit. 22% of units were owned by landlords with missing race data, and on average they overstated White ownership by 6.2 percentage points and understated Black ownership by 3.1 percentage points per unit. This created an aggregate bias of 2.5 percentage points for White ownership and -1.7 percentage points for Black ownership.

This analysis shows that although some variables have considerable missing or imperfect data, the low-quality measures and imputations are quite accurate, leading to a low degree of overall bias. An extreme example of this is the data from Baltimore, where the lack of officer data made it exceedingly difficult to operationalize measures like landlord race. Nevertheless, as discussed in Appendix J, where I show the results from Boston, Houston, and Miami after running the data construction pipeline without officer data, even the imputed values from Baltimore are remarkably accurate. Estimating the bias that arises when using administrative data to measure hard-to-observe constructs is an essential, though frequently neglected, part of using these data sources, allowing a determination of when data is sufficient to estimate a particular construct and when it is not.

Figure 4: Bias Produced by Missing and Imperfect Data by Measure



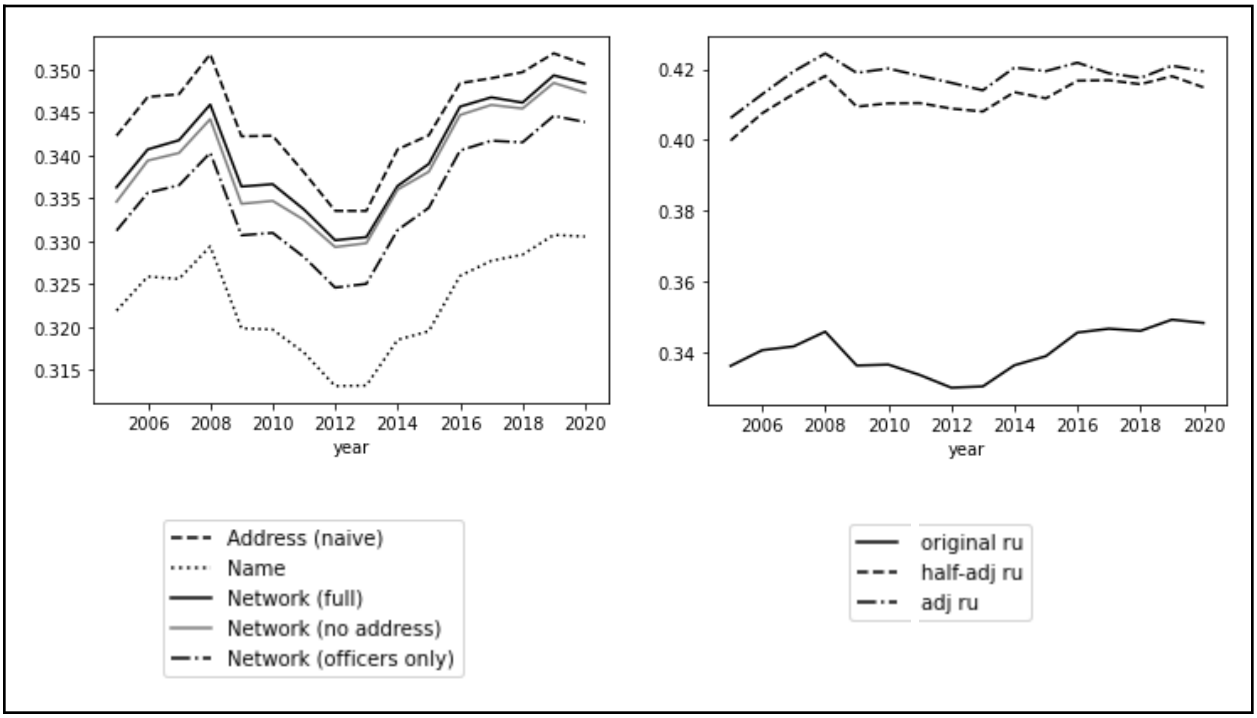
Note: Figure 4 shows estimates of the bias produced by low data quality and using imputed values for missing data for eight landlord measures. Each plot details the proportion of units with low data quality, the per-unit bias for those units, and the resulting total bias due to low data quality. It then shows the same three measures for bias created due to missingness, and the aggregate bias, which is the sum of the total bias from low data quality and from missingness.

Multiple Specifications

Aside from missing and imperfect data, another source of potential bias is that there are multiple ways to construct certain variables. For example, the choice of whether to adjust unit counts to match ACS/Census records and the array of decisions made when constructing landlord conglomerate networks could change estimates of the degree of ownership of large- and small-scale landlords. To estimate the potential variation in estimates created by these types of analytic decisions, in Figure 5 I present a range of estimates for the proportion of rental units owned by landlords with 100 or more units, by year. Plot 1 compares estimates under different conglomerate construction rules. The lowest line uses a naive linking rule, combining people with the same names and companies with the same names after removing corporate signifiers and common words. The next lowest line shows the results when only officership information is used, resulting in an increase of more than one percentage point. The next lowest includes all additional matches other than address-based matches, and finally the solid line shows the proportion when using all matches. The highest line uses another naive liking rule, combining all entities with the same address. This overstates ownership by large entities, showing the importance of using a more sophisticated set of linking rules.

Figure 5, Plot 2 compares the proportion of rental units owned by owners with 100+ units when using the original rental units estimates and those adjusted to match the Census/ACS. Adjusting to match the census leads to dramatically smaller estimates of ownership by very large landlords.

Figure 5: Multiple Estimates for the Proportion of Rental Units Owned by 100+ Unit Landlords



Note: Figure 5 shows variation in landlord measures created by different analytic decisions, using the proportion of rental units owned by landlords with 100 or more units as an example. The first plot shows variation in estimates from using different conglomerate construction rules (see Table 5). The second plot shows variation in estimates from using rental units that were or were not adjusted to match Census/ACS unit counts (see Figure 1). "Original ru" indicates the original rental unit counts, "adj ru" indicates fully adjusted to match the Census/ACS, and "half-adj ru" indicates they were adjusted halfway to resemble the Census/ACS.

Empirical Demonstrations

To demonstrate the utility of this methodology, I present several empirical analyses below. First, I examine the aggregate descriptive characteristics of landlords in the sample. Second, I examine how landlords' characteristics vary by neighborhood context. Third, I examine how characteristics have changed over time. Finally, because the preceding analyses focused on aggregate characteristics, I conclude with a brief illustrative example of the types of granular analyses that these data allow.

Basic Descriptive Characteristics of Landlords

Table 8 shows the full descriptive characteristics of the resulting datasets at the rental unit level. Across the four metro areas, the plurality of rental units, 39%, are in large multi-family properties (5+ units), followed by single-family (27%), condominiums (19%), and small multi-family properties (11%). Properties are smaller-scale in Boston (where small multi-family properties are prevalent) and Miami (where condominiums are prevalent). In contrast, Baltimore and Houston have more than 50% of their rental units in large multi-family properties, and the majority of the remaining in single-family properties. Per-unit property values are lowest in Houston, with a mean of $79,000, followed by Baltimore, at $104,000, Miami at $174,000, and Boston at $223,000. The properties are oldest in Boston, followed by Baltimore, Miami, and Houston.

About half of owners, 48%, are owned by private organizations (i.e., not people or government entities), of which about 44% (21% of the total) are LLCs. As mentioned above, 78% of these entities are matched to a business filing, meaning 88% of privately owned rental units have either a person owner or a company owner connected to a business filing. Connecting to a business filing is not synonymous with identifying a person owner, since in many cases an unconnected business was able to be connected to a likely owner, and in Baltimore no officer data existed in the business filings. A person owner was identified for 79% of all non-government-owned rental units.

Although 48% of units are company-owned, relatively few show signs of corporate formality. Only 12% of all rental units are owned by landlords whose focal companies are corporations, and 33%, 34%, 12%, and 26% by private landlords whose focal companies have more than one title, more than one officer, a director, and a corporate office, respectively.

Combining these multiple indicators into a composite variable, only 36% of rental units are owned by a landlord with an organizational formality score greater than or equal to two. Overall, landlords show the most formality in Houston, followed by Boston, then Miami.[36]

Overall, most landlords appear to be quite small-scale, with about 46% of rental units owned by landlords with four or fewer units. 35% are owned by landlords with more than 100, leaving only 11% and 8% in the middle two categories of 5-19 and 20-99 units, respectively. There is substantial variation in landlord scale between metro areas, with about half of rental units in Baltimore and Houston (45% and 52%, respectively) owned by landlords with more than 100 units, compared to only 24% and 21% in Boston and Miami.

77% of landlords' homes and company headquarters are in the central city or suburbs, with only 6% and 17% outside the metro but in the state, and outside the state, respectively. Although the home locations and company headquarters are quite similar, home locations are slightly more likely to be in the suburbs than are company headquarters. Between metro areas, Baltimore shows the most out-of-state ownership, at 24%, compared to only 9% in Boston, which has the least.

Finally, more than two-thirds, or 69% of units are owned by White landlords, 18%, 8%, and 5% owned by Hispanic, Black, and Asian landlords, respectively. Hispanic ownership is substantially higher in Miami and Houston, 30% and 14%, respectively, reflecting the larger Hispanic populations in those areas.

Table 8: Full Descriptive Characteristics of Parcels and Their Owners

Variable	All	Boston	Baltimore	Houston	Miami
Property characteristics					
Single-family	0.27	0.10	0.39	0.35	0.24
Condominium	0.19	0.11	0.06	0.05	0.38

[36] Measures of organizational formality omit data from Baltimore, for which officer data is missing. The greater formality of owners in Houston is also true among top evictors, as analyzed in Chapter 4.

Small multi-family	0.11	0.34	0.05	0.00	0.11
Large multi-family	0.39	0.36	0.50	0.53	0.25
Other residential	0.05	0.08	0.00	0.07	0.02
1 unit	0.46	0.22	0.45	0.41	0.62
2-3 unit	0.10	0.34	0.04	0.02	0.08
4-9 unit	0.05	0.09	0.02	0.03	0.06
10+ unit	0.39	0.35	0.49	0.55	0.24
Value per unit	145014	222730	103574	79366	174198
Year built	1968.4	1932.6	1950.2	1981.0	1978.4
Landowner characteristics					
Person	0.51	0.58	0.37	0.41	0.60
Government	0.02	0.05	0.02	0.00	0.01
Private organization	0.48	0.38	0.61	0.59	0.40
LLC	0.21	0.14	0.29	0.25	0.18
Non-gov owners:					
Has focal company	0.52	0.46	0.65	0.61	0.44
Owner person or linked to filing	0.88	0.85	0.84	0.91	0.90
Person owner identified	0.79	0.88	0.55	0.77	0.83
Focal company corporation	0.12	0.09	0.14	0.13	0.12
>1 distinct offices	0.33	0.40	NA	0.43	0.22
>1 distinct officers	0.34	0.37	NA	0.39	0.29
Has director	0.12	0.11	NA	0.21	0.07
Has corporate office	0.25	0.18	0.25	0.31	0.22
Org formality >2	0.36	0.38	0.00	0.48	0.28
1-4 units	0.46	0.47	0.35	0.35	0.58
5-19 units	0.11	0.16	0.11	0.07	0.12
20-99 units	0.08	0.12	0.09	0.06	0.09
100+ units	0.35	0.24	0.45	0.52	0.21
Home: Central city	0.29	0.29	0.35	0.42	0.16
Home: Suburbs	0.48	0.57	0.36	0.29	0.63
Home: Outside metro, same state	0.06	0.07	0.07	0.09	0.03
Home: Different state	0.17	0.07	0.21	0.19	0.18
PC: Central city	0.30	0.29	0.34	0.45	0.18
PC: Suburbs	0.44	0.55	0.34	0.27	0.56
PC: Outside metro, same state	0.09	0.08	0.09	0.13	0.05
PC: Different state	0.17	0.09	0.24	0.15	0.21

White	0.69	0.82	0.83	0.70	0.59
Asian	0.05	0.06	0.02	0.10	0.03
Hispanic	0.18	0.05	0.01	0.14	0.30
Black	0.08	0.07	0.14	0.06	0.08
Multiple	0.00	0.00	0.00	0.00	0.00
Native	0.00	0.00	0.00	0.00	0.00
Rental units	45,611,180	8,320,625	4,963,103	14,598,815	17,728,637

Note: Table 8 shows full descriptive characteristics for properties and landlords, weighted by rental units. Unadjusted rental unit counts were used, and using the Census/ACS-adjusted counts would increase the relative presence of large scale, corporate landlords. Property values were not adjusted for inflation and reflect the average between 2005 and 2019. Year built is the average for non-missing observations (see Table 2), as missing values were not imputed for this measure. All landlord characteristics pertain to all rental units, including observations with low data quality and observations missing data for which values were imputed (see Figure 4).

Basic Descriptive Characteristics of Landlords by Neighborhood

Figure 6 examines how landlord characteristics vary by neighborhood socioeconomic status, racial composition, and household composition. Each scatterplot shows one dot for each tract-year, with tract characteristics on the X axis and landlords characteristics on the Y axis, and best fit lines[37] shown separately for each metro.

First, higher-poverty areas have larger-scale landlords, whether defined as having 10 or more or 100 or more units. An exception to this trend is Baltimore, where, after about 15% poverty, the scale of landlords begins to decline. Nevertheless, across the four metro areas, neighborhoods with very little or no poverty have approximately 35% and 25% of units owned by landlords with 10 or more and 100 more units, respectively, while areas with about 30% poverty rates have about 65% and 45%, respectively. Higher-poverty areas also have slightly more formal owners, with about 40% of units owned by landlords with multiple officers, compared to only 30% in low-poverty areas, but, interestingly, high-poverty areas are slightly less likely to have out-of-state owners.

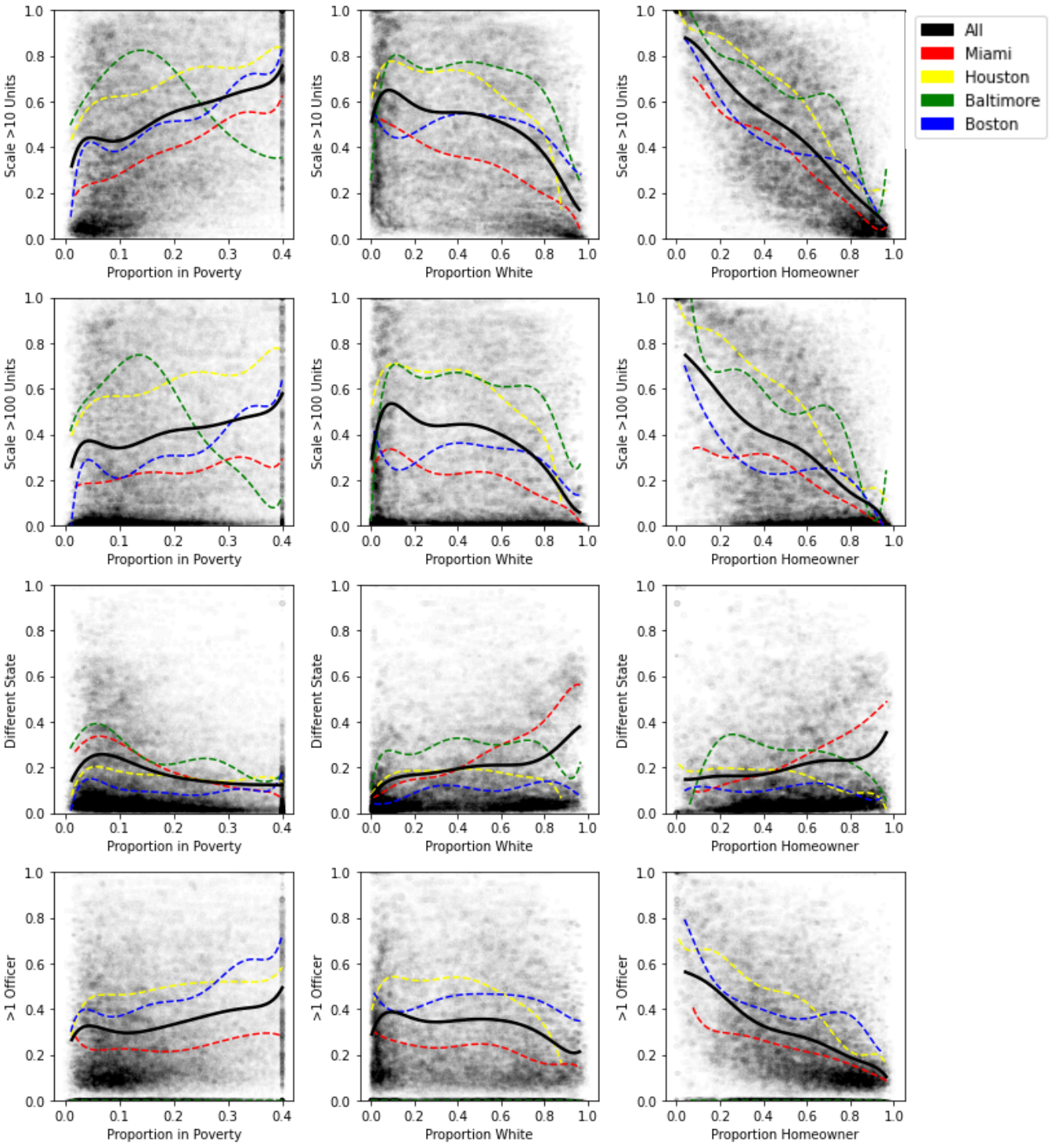
Perhaps reflecting differences in socioeconomic status, Whiter neighborhoods have much smaller-scale owners. Neighborhoods with fewer than 20% White residents have an average of

[37] Best-fit lines are weighted by the number of rental units in each neighborhood.

about 60% of units owned by landlords with more than 10 units, and about 50% by owners with more than 100, while those with more than 80% White residents have only 20% and 10% of owners with more than 10 and 100 units, respectively. Whiter areas are also more likely to have informal owners, but, surprisingly, they are also more likely to have out-of-state owners.

Finally, the proportion of homeowners in a neighborhood is very strongly associated with the scale of landlords there. Neighborhoods with fewer than 20% homeowners have approximately 75% of units owned by landlords with more than 10 units, and 65% by owners with more than 100 units. Both proportions drop to about 10% in neighborhoods with more than 80% homeowners. Out-of-state owners are more common in neighborhoods that are predominately homeowners, as are organizationally informal owners. This may reflect the presence of accidental landlords, owners who inherited rental properties or otherwise became landlords unintentionally (Shiffer-Sebba 2020), in high-homeowner neighborhoods.

Figure 6: Scatterplots of Tract and Landlord Characteristics



Note: Figure 6 shows scatterplots of neighborhood demographic and average landlord characteristics, with best fit lines presented separately for each metro area and for all four metro areas as a whole. 12 plots are presented, reflecting combinations of the four landlord characteristics on the y-axes and three tract demographics on the x-axes. These plots use unadjusted rental unit counts.

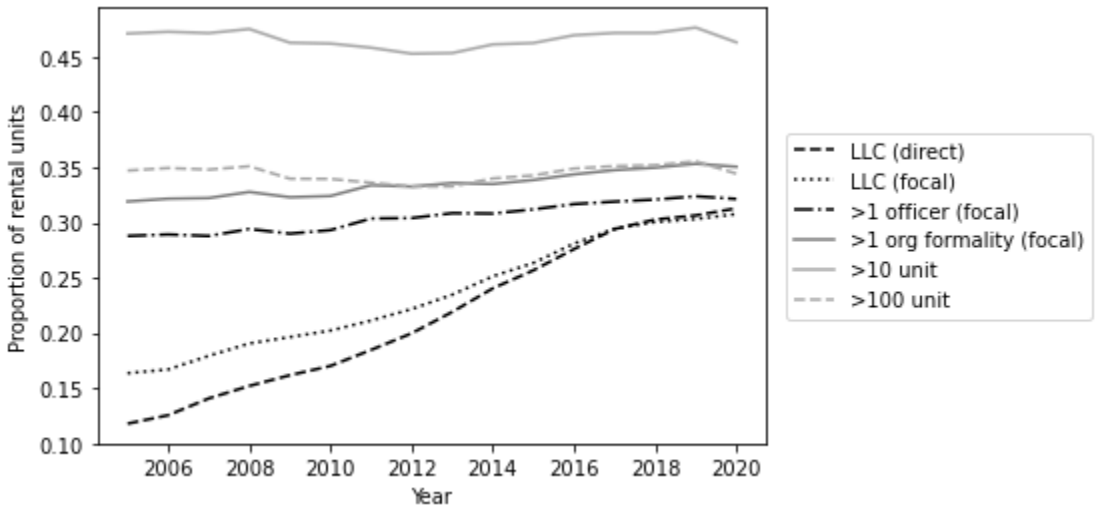
Changes in Landlord Characteristics

Numerous studies have noted that an increasing number of rental properties are owned by LLCs and other non-person entities (Travis 2019; Messamore 2024), and this trend is frequently interpreted to mean that landlords are becoming more large-scale, financialized, and organizationally formal. However, LLCs are not a direct indicator of these characteristics, and in many cases ownership through an LLC rather than a person indicates no difference in the underlying characteristics or practices of a landlord, except that they chose to own in a way that shields their liability and identity. More specific and nuanced characteristics are necessary to better understand the changing organizational characteristics of American landlords.

Figure 7 shows the proportion of rental units owned by landlords with different organizational characteristics, for each year between 2005 and 2020. First, this plot shows a clear increase in ownership by LLCs, rising from less than 12% of rental units in 2005 to nearly 30% in 2020. Likewise, when we consider the focal company within a conglomerate, (the company with the highest amount of direct and indirect ownership), we see a substantial increase in LLC ownership from about 16% to 30%. However, if we consider more nuanced measures of organizational complexity, including the proportion of units owned by landlords with more than one officer, and the proportion with an organizational formality score larger than 1, the increases are more muted. The proportion with multiple officers increased only slightly from about 29% to 31%, while the composite formality score increased from 32% to about 34%. These more nuanced indicators allow us to see that although formalization and corporatization have increased among landlords, the change has been much smaller-scale than the increase in LLC ownership.

This finding is reinforced by the trends in owner scale, which show little increase in ownership among landlords with more than 10 and with more than 100 units.

Figure 7: Changes in Landlords' Organizational Characteristics Between 2005 and 2020



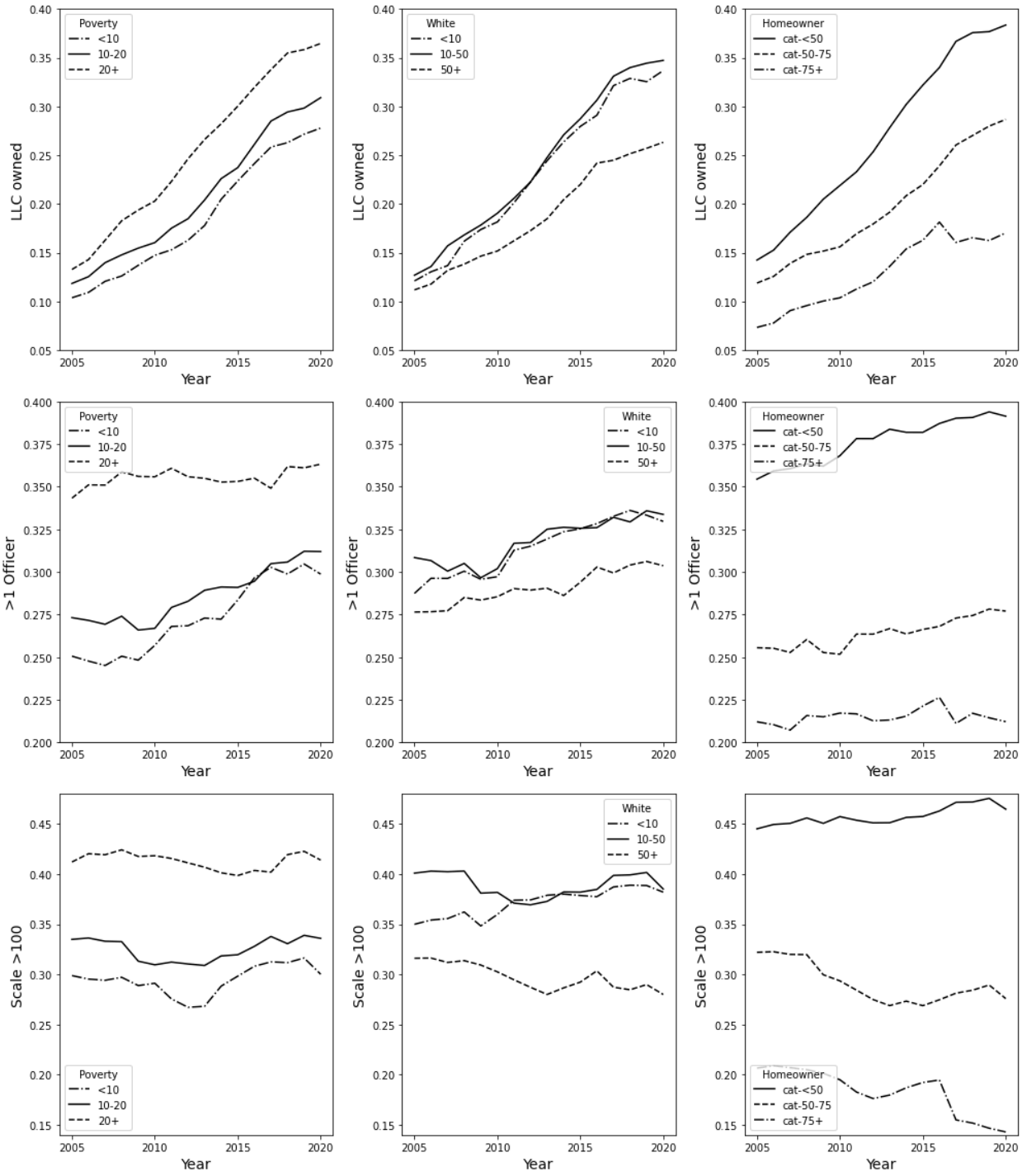
Note: Figure 7 shows changes in the composition of landlords between 2006 and 2020, using different measures of landlord scale and organizational formality. The plot uses unadjusted rental unit counts.

However, the lack of a large aggregate trend in Figure 7 does not mean that particular submarkets have not seen increases in large-scale, corporate ownership. In Figure 8, I show changes in the proportion of units owned by an LLC, by a company with more than one officer, and by an owner with more than 100 units, for subsamples with different neighborhood characteristics. In each plot, years are shown on the X axis and landlord characteristics on the Y axis, with different lines for different neighborhood types

First, row one shows that in all neighborhood types, there was a large increase in LLC ownership. However, rows two and three show that this was not necessarily reflective of changes in formality or scale. Row 2 shows that organizational formality increased across most neighborhood types, but the scale of change was much smaller than that in LLC ownership, typically fewer than five percentage points, while increases in LLC ownership were often as

much as 20 percentage points. The areas with the largest increases in organizational formality were those with few or no White residents and those with few homeowners. Few areas saw sustained increases in owner scale. Nevertheless, this analysis reinforces the earlier finding that large-scale owners are much more common in non-White areas, in high-poverty areas, and, particularly, in low homeowner areas, although these patterns have not changed substantially over time.

Figure 8: Changes in Landlords' Organizational Characteristics By Neighborhood Type

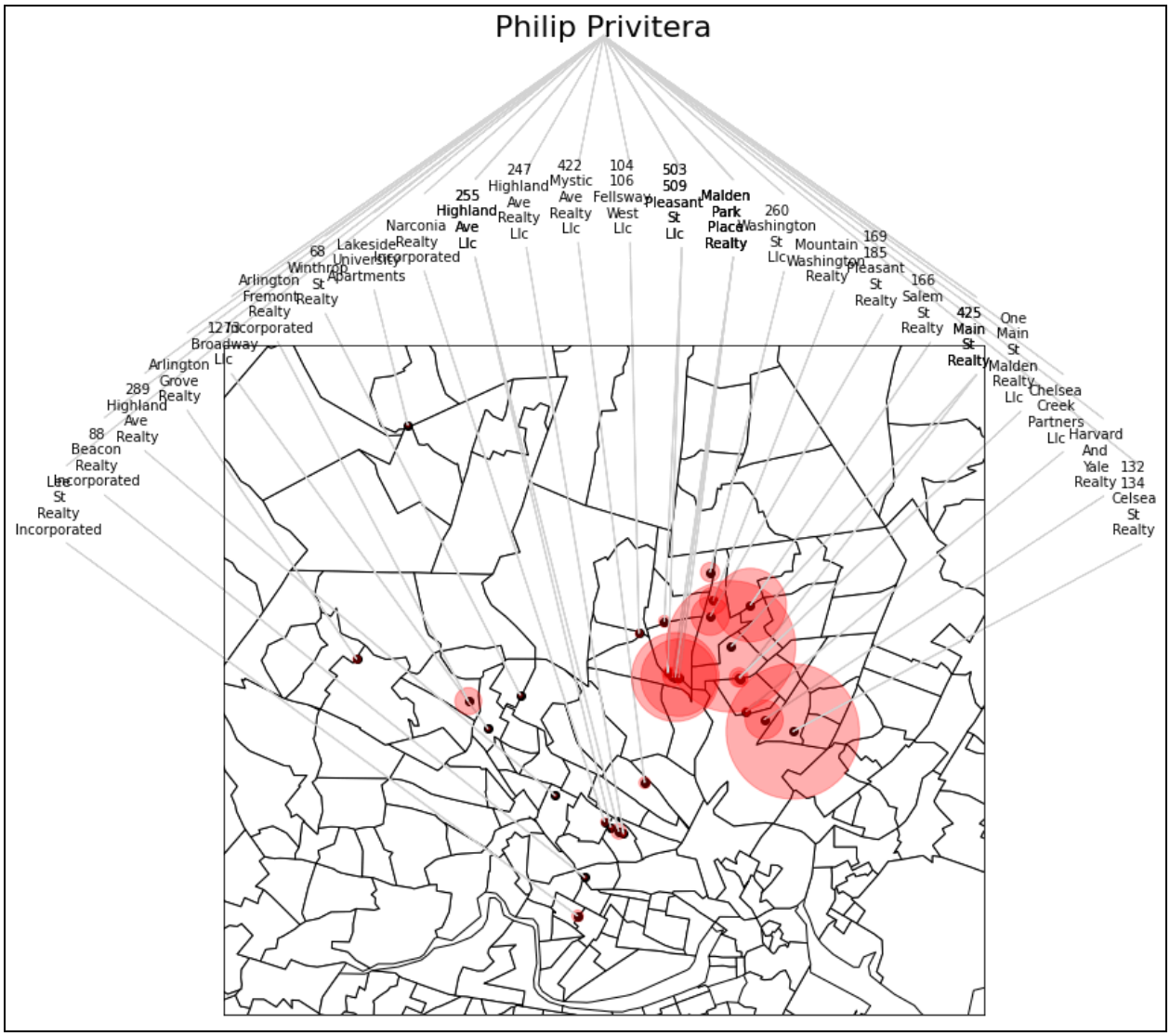


Note: Figure 8 shows changes in landlord characteristics over time for tracts with different demographic characteristics. For each plot, the x-axis shows years and the y-axis shows landlord characteristics, with different lines for each category of neighborhood demographics. For example, the upper-left plot shows changes in the proportion of LLC-owned rental units for tracts with less than 10% poverty, 10-20% poverty, and 20% or higher poverty.

Small-N Analyses of Large Owners

Although the previous analyses provide aggregate pictures of landlord characteristics, a range of other, more micro-focused analyses are also possible. One of the key advantages of these data is that they constitute a full population of landowners within each metro area, allowing fine-grained analysis of the largest owners. As a demonstration of this type of analysis, Figure 9 shows a map of the Privitera conglomerate's properties where eviction filings occurred, with circles indicating the number of eviction filings at each property and lines connecting those properties to the companies that own them. This type of data can provide a novel view of urban life and inequality, emphasizing the disproportionate effects that a small number of owners can have and thereby complementing approaches that focus on aggregate conditions and characteristics. These data can also provide a useful starting point for qualitative analyses of these owners. For example, studies could use these data to identify properties as locations for qualitative interviews or ethnographies, or as the starting point for qualitative analyses that draw on newspaper articles, websites, and other online media pertaining to these landowners.

Figure 9: Map of Philip Privitera's Companies, Properties, and Eviction Filings



Note: Figure 9 shows the properties owned by Philip Privitera in the Boston metro area at which evictions were filed, with red circles indicating the numbers of filings. A line extends from each property to the name of the company that owns it, and from there to the company's owner, Philip Privitera. Black lines outline Census tracts.

Discussion

In this article, I introduced a methodology for using tax assessments, business filings, and auxiliary datasets like eviction filing records to construct linked longitudinal datasets describing urban landowners, the properties they own, and the activities they carry out. The paper provided detailed explanations of each step in the data construction pipeline, internal and external

validations to ensure data accuracy, and several empirical examples that demonstrate the analytic potential of the resulting datasets.

This methodology contributes in several ways to ongoing efforts to produce sociologically relevant measures of landowners and rental markets, by providing solutions to four challenges that arise in these efforts: (1) I used internal validations that identified aberrant data and external harmonization with ACS/Census records to correct inaccuracies and biases in the tax assessment records. (2) I developed a network-based probabilistic method to produce accurate unique identifiers for entities with sparse but networked information. (3) I developed a framework for corporate network identification that addressed a wide range of possible links between members of the conglomerate while avoiding false positives. Finally, (4) I developed a range of sociologically relevant landowner measures, using imperfect indicators, and estimated their accuracy.

This methodology also contributes to landowner data methodologies by creating a pipeline that can be applied to data from any geographic area and that examines the potential biases arising from differences in data availability and quality from different sources. Tax assessments and business filings are widely available administrative datasets that can be obtained from local government agencies. I have made all code and data publicly available online at github.com, and once the data have been formatted and cleaned to match a standardized format, the pipeline can be run without additional changes.

The resulting datasets describing landowners and their properties opens considerable analytic possibilities. As demonstrated in the first two empirical examples, they can provide detailed descriptive characteristics of landlords across time and space. As demonstrated in the third empirical example, because it constitutes a full population, these data can be used to

identify the small number of extremely large owners who have an outsized impact on urban life. More detailed versions of these types of analyses are carried out in Chapters 3 and 4, respectively.

A range of additional studies are possible as well. Predictive analyses that connect landlord characteristics with behaviors, for example examining what types of developers pursue redevelopment strategies in gentrifying neighborhoods, could give greater insight into the behavioral differences between types of landowners. Auxiliary datasets other than eviction filings, such as building permits, housing code violations, and criminal incidents can easily be integrated into the datasets, allowing us to understand the correlates of a range of landlord behaviors. Finally, these data can also be used to show the links between advantaged owners and disadvantaged places in which they own property, contributing to a relational perspective on urban inequality.

References

August, Martine. 2020. "The Financialization of Canadian Multi-Family Rental Housing: From Trailer to Tower." *Journal of Urban Affairs* 0(0):1–23. doi: 10.1080/07352166.2019.1705846.

Bacon, Seb. 2013a. "Understanding Corporate Networks. Part 1: Control via Equity." Retrieved February 15, 2024 (https://blog.opencorporates.com/2013/10/16/understanding-corporate-networks-part-1-control-via-equity/).

Bacon, Seb. 2013b. "Understanding Corporate Networks. Part 2: Control without Voting." Retrieved June 23, 2024 (https://blog.opencorporates.com/2013/10/31/understanding-corporate-networks-part-2-control-without-voting/).

Balzarini, John, and Melody L. Boyd. 2020. "Working With Them: Small-Scale Landlord Strategies for Avoiding Evictions." *Housing Policy Debate* 0(0):1–21. doi: 10.1080/10511482.2020.1800779.

Charles, Suzanne Lanyi. 2020. "The Financialization of Single-Family Rental Housing: An Examination of Real Estate Investment Trusts' Ownership of Single-Family Houses in the Atlanta Metropolitan Area." *Journal of Urban Affairs* 42(8):1321–41. doi: 10.1080/07352166.2019.1662728.

Chilton, Ken, Robert Mark Silverman, Rabia Chaudhry, and Chihaungji Wang. 2018. "The Impact of Single-Family Rental REITs on Regional Housing Markets: A Case Study of Nashville, TN." *Societies* 8(4):93. doi: 10.3390/soc8040093.

Connelly, Roxanne, Christopher J. Playford, Vernon Gayle, and Chris Dibben. 2016. "The Role of Administrative Data in the Big Data Revolution in Social Science Research." *Social Science Research* 59:1–12. doi: 10.1016/j.ssresearch.2016.04.015.

Decker, Nathaniel. 2023. "How Landlords of Small Rental Properties Decide Who Gets Housed and Who Gets Evicted." *Urban Affairs Review* 59(1):170–99. doi: 10.1177/10780874211041513.

Desmond, Matthew. 2012. "Eviction and the Reproduction of Urban Poverty." *American Journal of Sociology* 118(1):88–133. doi: 10.1086/666082.

Desmond, Matthew. 2016. *Evicted: Poverty and Profit in the American City*. First Edition. New York: Crown Publishers, 2016.

Desmond, Matthew, and Carl Gershenson. 2016. "Housing and Employment Insecurity among the Working Poor." *Social Problems* 63(1):46–67. doi: 10.1093/socpro/spv025.

Desmond, Matthew, and Nathan Wilmers. 2019. "Do the Poor Pay More for Housing? Exploitation, Profit, and Risk in Rental Markets." *American Journal of Sociology* 124(4):1090–1124.

Elmagarmid, Ahmed K., Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. "Duplicate Record Detection: A Survey." *IEEE Transactions on Knowledge and Data Engineering* 19(1):1–16. doi: 10.1109/TKDE.2007.250581.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2018. *Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records*. *SSRN Scholarly Paper*. ID 3214172. Rochester, NY: Social Science Research Network.

Feagin, Joe R., and Robert Parker. 1990. *Building American Cities: The Urban Real Estate Game*. Englewood Cliffs, N.J.: Prentice Hall.

Fields, Desiree. 2018. "Constructing a New Asset Class: Property-Led Financial Accumulation after the Crisis." *Economic Geography* 94(2):118–40. doi: 10.1080/00130095.2017.1397492.

Fischer, Mary J., and Douglas S. Massey. 2004. "The Ecology of Racial Discrimination." *City & Community* 3(3):221–41. doi: 10.1111/j.1535-6841.2004.00079.x.

Freeman, Lance. 2006. *There Goes the 'hood: Views of Gentrification from the Ground Up*. Philadelphia, PA: Temple University Press.

Garboden, Philip ME, and Eva Rosen. 2019. "Serial Filing: How Landlords Use the Threat of Eviction." *City & Community* 18(2):638–61. doi: https://doi.org/10.1111/cico.12387.

Garboden, Philip, Eva Rosen, Meredith Greif, Stephanie DeLuca, and Kathryn Edin. 2018. "Urban Landlords and the Housing Choice Voucher Program - A Research Report." 54.

Gilderbloom, John I., and Richard P. Appelbaum. 1987. "Toward a Sociology of Rent: Are Rental Housing Markets Competitive?" *Social Problems* 34(3):261–76. doi: 10.2307/800766.

Goldenstein, Jan, and Philipp Poschmann. 2019. "Analyzing Meaning in Big Data: Performing a Map Analysis Using Grammatical Parsing and Topic Modeling." *Sociological Methodology* 49(1):83–131. doi: 10.1177/0081175019852762.

Gomory, Henry. 2022. "The Social and Institutional Contexts Underlying Landlords' Eviction Practices." *Social Forces* 100(4):1774–1805. doi: 10.1093/sf/soab063.

Gomory, Henry, and Matthew Desmond. 2023. "Neighborhoods of Last Resort: How Landlord Strategies Concentrate Violent Crime." *Criminology* 61(2):270–94. doi: 10.1111/1745-9125.12332.

Gotham, Kevin Fox. 2000. "Growth Machine Up-Links: Urban Renewal and the Rise and Fall of a Pro-Growth Coalition in a U.S. City." *Critical Sociology* 26(3):268–300. doi: 10.1177/089692050002600305 01.

Hackworth, Jason. 2006. *The Neoliberal City: Governance, Ideology, and Development in American Urbanism*. 1 edition. Ithaca: Cornell University Press.

Hangen, Forrest, and Daniel T. O'Brien. 2024. "Linking Landlords to Uncover Ownership Obscurity." *Housing Studies* 0(0):1–26. doi: 10.1080/02673037.2024.2325508.

Hwang, Jackelyn, and Nikhil Naik. 2023. "Systematic Social Observation at Scale: Using Crowdsourcing and Computer Vision to Measure Visible Neighborhood Conditions." *Sociological Methodology* 53(2):183–216. doi: 10.1177/00811750231160781.

Imai, Kosuke, and Kabir Khanna. 2016. "Improving Ecological Inference by Predicting Individual Ethnicity from Voter Registration Records." *Political Analysis* 24(2):263–72. doi: 10.1093/pan/mpw001.

Immergluck, Dan, Jeff Ernsthausen, Stephanie Earl, and Allison Powell. 2019. "Evictions, Large Owners, and Serial Filings: Findings from Atlanta." *Housing Studies* 0(0):1–22. doi: 10.1080/02673037.2019.1639635.

Jerolmack, Colin, and Shamus Khan. 2014. "Talk Is Cheap Ethnography and the Attitudinal Fallacy." *Sociological Methods & Research* 43(2):178–209. doi: 10.1177/0049124114523396.

Korver-Glenn, Elizabeth. 2018. "Compounding Inequalities: How Racial Stereotypes and Discrimination Accumulate across the Stages of Housing Exchange." *American Sociological Review* 83(4):627–56. doi: 10.1177/0003122418781774.

La Porta, Rafael, Florencio Lopez-De-Silanes, and Andrei Shleifer. 1999. "Corporate Ownership Around the World." *The Journal of Finance* 54(2):471–517. doi: 10.1111/0022-1082.00115.

Lees, Loretta. 2003. "Super-Gentrification: The Case of Brooklyn Heights, New York City." *Urban Studies (Routledge)* 40(12):2487–2509.

Leung, Lillian, Peter Hepburn, and Matthew Desmond. 2020. "Serial Eviction Filing: Civil Courts, Property Management, and the Threat of Displacement." *Social Forces* (soaa089). doi: 10.1093/sf/soaa089.

Lyon, Larry, Lawrence G. Felice, M. Ray Perryman, and E. Stephen Parker. 1981. "Community Power and Population Increase: An Empirical Test of the Growth Machine Model." *American Journal of Sociology* 86(6):1387–1400. doi: 10.1086/227389.

Mallach, Alan. 2007. ""Landlords at the Margins: Exploring the Dynamics of the One To Four Unit Rental Housing Industry."" Revisiting Rental Housing: A National Policy Summit.

Mallach, Alan. 2014. "Lessons From Las Vegas: Housing Markets, Neighborhoods, and Distressed Single-Family Property Investors." *Housing Policy Debate* 24(4):769–801. doi: 10.1080/10511482.2013.872160.

Marcuse, Peter, and David Madden. 2016. *In Defense of Housing: The Politics of Crisis*. London ; New York: Verso.

Massey, Douglas S., and Nancy A. Denton. 1993. *American Apartheid : Segregation and the Making of the Underclass*. Cambridge, Mass: Harvard University Press.

Messamore, Andrew. 2024. “The Institutionalization of Landlording: Assessing Transformations in Property Ownership Since the Great Recession.”

Molotch, Harvey, William Freudenburg, and Krista E. Paulsen. 2000. “History Repeats Itself, But How? City Character, Urban Tradition, and the Accomplishment of Place.” *American Sociological Review* 65(6):791–823. doi: 10.2307/2657514.

Molotch, Harvey Luskin, and John R. Logan. 1987. *Urban Fortunes : The Political Economy of Place*. Berkeley, CA: University of California Press.

Nelson, Laura K. 2019. “To Measure Meaning in Big Data, Don’t Give Me a Map, Give Me Transparency and Reproducibility.” *Sociological Methodology* 49(1):139–43. doi: 10.1177/0081175019863783.

O’Brien, Daniel Tumminelli, Robert J. Sampson, and Christopher Winship. 2015. “Ecometrics in the Age of Big Data: Measuring and Assessing ‘Broken Windows’ Using Large-Scale Administrative Records.” *Sociological Methodology* 45(1):101–47. doi: 10.1177/0081175015576601.

Pager, Devah, and Hana Shepherd. 2008. “The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets.” *Annual Review of Sociology* 34:181–209.

Preis, Benjamin. 2024. “Where the Landlords Are: A Network Approach to Landlord-Rental Locations.” *Annals of the American Association of Geographers* 0(0):1–12. doi: 10.1080/24694452.2023.2277810.

Raymond, Elora Lee, Richard Duckworth, Benjamin Miller, Michael Lucas, and Shiraj Pokharel. 2018. “From Foreclosure to Eviction: Housing Insecurity in Corporate-Owned Single-Family Rentals.” *Cityscape* 20(3):159–88.

Robinson, John N. 2021. “Surviving Capitalism: Affordability as a Racial ‘Wage’ in Contemporary Housing Markets.” *Social Problems* 68(2):321–39. doi: 10.1093/socpro/spaa078.

Rosen, Eva, and Philip M. E. Garboden. 2020. “Landlord Paternalism: Housing the Poor with a Velvet Glove.” *Social Problems* (spaa037). doi: 10.1093/socpro/spaa037.

Rosen, Eva, Philip M. E. Garboden, and Jennifer E. Cossyleon. 2021. “Racial Discrimination in Housing: How Landlords Use Algorithms and Home Visits to Screen Tenants.” *American Sociological Review* 86(5):787–822. doi: 10.1177/00031224211029618.

Rule, Alix, Jean-Philippe Cointet, and Peter S. Bearman. 2015. “Lexical Shifts, Substantive Changes, and Continuity in State of the Union Discourse, 1790–2014.” *Proceedings of the National Academy of Sciences* 112(35):10837–44. doi: 10.1073/pnas.1512221112.

Salganik, Matthew. 2017. *Bit by Bit Social Research in the Digital Age.* Princeton Univ Pr.

Satter, Beryl. 2010. *Family Properties: How the Struggle Over Race and Real Estate Transformed Chicago and Urban America*. First edition. New York: Picador.

Schneider, Mark. 1992. "Undermining the Growth Machine: The Missing Link between Local Economic Development and Fiscal Payoffs." *The Journal of Politics* 54(1):214–30. doi: 10.2307/2131651.

Schwirian, Kent P. 1983. "Models of Neighborhood Change." *Annual Review of Sociology* 9(1):83–102. doi: 10.1146/annurev.so.09.080183.000503.

Seymour, Eric, and Joshua Akers. 2020. "'Our Customer Is America': Housing Insecurity and Eviction in Las Vegas, Nevada's Postcrisis Rental Markets." *Housing Policy Debate* 0(0):1–24. doi: 10.1080/10511482.2020.1822903.

Seymour, Eric, and Taylor Shelton. 2023. "How Private Equity Landlords Prey on Working-Class Communities of Color." *New Labor Forum* 109579602311701. doi: 10.1177/10957960231170168.

Shiffer-Sebba, Doron. 2020. "Understanding the Divergent Logics of Landlords: Circumstantial versus Deliberate Pathways." *City & Community* 19(4):1011–37. doi: https://doi.org/10.1111/cico.12490.

Smith, Neil. 1979. "Toward a Theory of Gentrification A Back to the City Movement by Capital, Not People." *Journal of the American Planning Association* 45(4):538–48. doi: 10.1080/01944367908977002.

Stegman, Michael A. 1972. *Housing Investment in the Inner City: The Dynamics of Decline; a Study of Baltimore, Maryland, 1968-1970*. Cambridge, Mass., M.I.T. Press.

Sternlieb, George. 1966. *The Tenement Landlord*. First Edition edition. Urban Studies Center, Rutgers, State University.

Taylor, Keeanga-Yamahtta. 2019. *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership*. Chapel Hill: The University of North Carolina Press.

Travis, Adam. 2019. "The Organization of Neglect: Limited Liability Companies and Housing Disinvestment." *American Sociological Review* 0003122418821339. doi: 10.1177/0003122418821339.

Tzioumis, Konstantinos. 2018. "Demographic Aspects of First Names." *Scientific Data* 5(1):180025. doi: 10.1038/sdata.2018.25.

Voicu, Ioan. 2018. "Using First Name Information to Improve Race and Ethnicity Classification." *Statistics and Public Policy* 5(1):1–13. doi: 10.1080/2330443X.2018.1427012.

Warner, Kee, and Harvey Molotch. 1995. "Power to Build: How Development Persists Despite Local Controls." *Urban Affairs Quarterly* 30(3):378–406. doi: 10.1177/107808749503000304.

Zeitlin, Maurice. 1974. “Corporate Ownership and Control: The Large Corporation and the Capitalist Class.” *American Journal of Sociology* 79(5):1073–1119. doi: 10.1086/225672.

Zhang, Han, and Jennifer Pan. 2019. “CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media.” *Sociological Methodology* 49(1):1–57. doi: 10.1177/0081175019860244.

Appendix 1: Supplemental Materials for Chapter 2

A. Standardizing Data Sources from Multiple Sources

Tax assessment and business filing records are collected by state agencies in diverse formats, differing in the way information is organized between linked databases (e.g., in some cases owners, officers, registered agents, addresses, names, or other information are stored in distinct datasets), as well as variables and variable values. I created a standard format for both types of records, with two linked datasets each describing parcels and business filings. The precise variables, variable values, and data quality checks required (e.g., ensuring that all parcels have an owner) are detailed on the public github, and a script is provided that checks that a dataset meets the required formatting. Once the input datasets have been cleaned to match these standards, the data construction pipeline can run without further modifications.

B. Correcting Administrative Artifacts - Inconsistencies Across Years in the Parcel Data

I looked for instances where the land usage changed or the parcel ceased to exist between years. After fixing data inaccuracies, no cities had years where more than 6% of parcels changed between years or more than 10% of land usages change between years (except in cases of large-scale construction). The examples in Table B.1. below, including Princeton, FL, and Brookshire, TX, were examples of small cities where large single-family developments were created, wherein hundreds of parcels were reclassified, typically from vacant to single-family.

Table B.1: Internal Validity Checks to Identify Administrative Artifacts

Variable	Parcel-years changed	Parcels changed ever	Top city-years
Exists	0.003	0.035	Sweetwater, FL, 2013: 6.2% Melrose, MA, 2017: 6.0% Parkland, FL, 2013: 4.6%
Land usage	0.009	0.109	Princeton, FL, 2017: 35.4% Princeton, FL, 2016: 30.6% Brookshire, TX, 2019: 20.0%

C. Tax Assessment Variable Imputation

C.1. Land Usage

In Baltimore, about 3% of parcel-years in all counties were missing land usage data. These often had values in subsequent years, so I filled in these values using the land usage value listed for the parcel in the most proximate year, and I labeled both the land usages and the number of unit variables as imputed.

In Galveston, many parcels in 2009 and 2020 were missing land usage data, so I filled in their values from 2010 and 2019, respectively, leaving only a small number of parcels with missing data.

In Greater Miami, a large number of parcels were mistakenly labeled as condominiums in 2008. I filled in these values with their land usages from 2009.

C.2. Number of Units

Unit data was missing for all properties in Montgomery, Fort Bend, and Galveston counties (in greater Houston), and for all properties in Boston proper. Unit data can be inferred from land usage for many property types: single-family properties and condominiums can be assumed to be one-unit, and in some places, like Boston proper, more specific land usage designations, like two-unit and three-unit multi-family, exist. Accordingly, missing unit data is typically a problem for large, multi-family properties.

In Boston, the city publishes a dataset listing all of the street addresses associated with each parcel. For multi-unit properties, the dataset typically lists one address for the entire building and one address for each unit. For example, a three unit property at 10 Main Street would have four addresses: “10 Main Street,” “10 Main Street, Apt 1,” “10 Main Street, Apt 2,” and “10 Main Street, Apt 3.” I omitted the first addresses (which are marked with a flag) and aggregated the remaining to the parcel to calculate unit counts. The vast majority of properties missing units from the tax assessment dataset had accompanying address records, but in the small number that did not, I imputed units using their square footage, as detailed below. This method was only possible in Boston, since I could not find analogous street address datasets in other areas.

In Galveston and Fort Bend Counties, the tax assessment records provide total square footage of living area for each parcel. I estimated the number of units in multi-family properties in these areas by dividing the square footage by 900, which is a typical size for a unit in a multi-family apartment building.

In Montgomery County, square footage data is missing, so I drew on the valuation of each property. Because property value is a function of many factors other than the number of units, I could not assume a direct relationship between the two measures. Instead, I calculated the number of housing units in each block that were remaining after subtracting the single-family,

condominium, and other properties for which I could impute unit counts based on land usage, in my data. I then apportioned the remaining units to the remaining properties, proportional to their property value. This assumes that, within a Census block, property value is largely proportional to property size.

I validated the street address-based unit imputation method by comparing its results to true values for those parcel for which unit counts could be inferred based on land usage. In Boston, this consists of large numbers of two-unit and three-unit properties, making this a more stringent test than if the only properties for which unit data could be inferred were single-family and condominium properties. The address-based estimates are correlated 0.502 with the unit estimates from land usage, and the square footage based estimates are correlated 0.717 with the true values. Among multi-family properties, for which true unit values were unknown, the address- and square footage-based methods were correlated with one another 0.434.

I validated the square footage and valuation/census imputation methods using data from Brazoria, which is the county most similar to Montgomery, Fort Bend, and Galveston, but that still has nonmissing unit data (Harris County was not used because it contains Houston). In Brazoria, among those properties for which unit data cannot be determined and would need to be imputed (not single-family or condominium), the square-footage based imputation has a correlation of 0.53 with the true value, and the valuation/Census based imputation has a correlation of 0.55. The two measures have a correlation of 0.71 with one another. While not perfect, this indicates a fairly high degree of accuracy. In Galveston, the two measures had a correlation of 0.71, but in Fort Bend, their correlation was -0.09. (Correlations of each measure with the true unit counts was not possible, since, by definition, they lack true unit data. Correlations between the two imputation methods were also not possible for Montgomery, where only valuation data was available).

In Boston, for those observations missing unit data, unit counts created using street address data were correlated 0.43 with counts calculated using square footage.

C.3. Owner-Occupancy

Owner-occupancy data was missing for all observations in the cities surrounding Boston. I imputed owner-occupancy by labeling as owner-occupied any properties owned by a person (rather than a company) and who listed the property itself as their contact address. Company-owned properties are not eligible for homeowners exemptions in Massachusetts, and listing a distinct contact address would suggest the property is a rental property, a second home, or something other than the owner's primary home. This method has been used in past research (Gomory 2021; Messamore 2023), but further analysis should examine its accuracy. Nevertheless, if this approach under or over represents single-family rentals, much of the

resulting bias should be obviated by using rental unit counts that are adjusted to match property-type-specific Census counts.

D. Comparison of Tax Assessment Data to ACS/Census Data

To identify the types of properties that are driving miscounts of rental units, I estimated regressions predicting the number of total units and rental units in a tract-year, according to the Census, using the total and rental unit counts from different types of properties in my tax assessment records. If the coefficient for, say, single-family rental units is 1.0, it would suggest that my data is an accurate representation of rental units in single-family properties, but if it were 1.5 or 0.5, it would suggest that the tax records are understating or overstating units in single-family properties, respectively.

The regression shows that the tax assessment estimates for total units are largely accurate for all property types (the coefficient is above 0.80 for all). Likesie, rental units in small and large multi-family properties are largely accurate (coefficients are 0.92 and 0.86, respectively), but the tax assessments overcount rental units in single-family, condominium, and other residential properties (coefficients are 0.29, 0.24, and 0.30, respectively). These results accord with those from Figure 1 in the main text.

Table D1: OLS Model Predicting Census Tract-Level Rental Unit Counts

	Census total unit counts	Census rental units counts
Variable		
Single-family units (total or rental)	0.818	0.293
	(0.002)	(0.005)
Condominium units (total or rental)	0.884	0.237
	(0.003)	(0.003)
Small multi-family units (2-4) (total or rental)	0.936	0.920
	(0.013)	(0.009)
Large multi-family units (5+) (total or rental)	0.919	0.862
	(0.004)	(0.003)
Other residential units (total or rental)	1.020	0.303
	(0.013)	(0.011)
Intercept	401.8	192.1
	(4.0)	(2.3)

E. Network-Based Name and Address Cleaning

I used the networked nature of the data to clean missing and incorrect information in listed addresses. Since individuals and companies may list the same address tens, hundreds, or even thousands of times, it is not uncommon that a single address appears in multiple forms, even after cleaning. For example, "10 Main Street Apt 2, Boston, 02131" may appear with information omitted, as "10 Main Street, Boston 02131" or "10 Main Street, Apt 2, Boston," or with misspellings, as "10 Mian Street, Apt 2, Boston 02131." In these instances, we cannot determine the true address simply by cleaning the faulty one, but by comparing it to other addresses linked to the same person we can identify the correct address. Accordingly, I identified instances where a person (or conglomerate) listed multiple addresses with the same number and street name (e.g., 10 Main), and filled in the street directionals, city, state, zip, (and in some cases unit), based on values from other similar addresses.

F. Entity Reconciliation

F.1. Discussion of Other Approaches and Reasons for Network-Based Approach

Typically, in probabilistic entity reconciliation methods, forms of information agreement (e.g., having the same or a similar name, address, birthdate, or gender) are parameterized with a single variable representing a degree and type of agreement (e.g., having a perfect birthdate match, or a 90% name match) (Elmagarmid et al. 2014; Enamorado 2018). The networked information could be fit into this framework by including variables indicating "owned the same parcel," "were officers of the same company," or even "owned a parcel that was owned by a company that has the latter as an officer." However a single parameter is fit for each *type* of agreement, rather than for the specific networked connection. This loses valuable information, since being linked to one parcel may be a much stronger indication of being the same entity than being linked to another. The network-based entity reconciliation accounts for variation in the import of different types of links, which more effectively leverages the information in the data.

F.2. Full Derivation of True Positive Rate

In Equation 1, I use Bayes' Rule to write the probability of two entities being the same, conditional on them having a name match and a link match, as the probability that they have a name and link match, conditional on being the same, time the probability of being the same, divided by the unconditional probability of having a name and link match. In Equations 2 and 3, I rewrite the numerator using conditional probability. In Equation 4, I divide the unconditional probability of having a link match from both sides of the fraction.

$$P(M_{ij}=1|N_{ij}=1,L_{ij}=1)=\frac{P(N_{ij}=1,L_{ij}=1|M_{ij}=1)P(M_{ij}=1)}{P(N_{ij}=1,L_{ij}=1)} \quad (1)$$

$$=\frac{P(N_{ij}=1|L_{ij}=1,M_{ij}=1)P(L_{ij}=1|M_{ij}=1)P(M_{ij}=1)}{P(N_{ij}=1|L_{ij}=1)P(L_{ij}=1)} \quad (2)$$

$$=\frac{P(N_{ij}=1|L_{ij}=1,M_{ij}=1)P(M_{ij}=1|L_{ij}=1)P(L_{ij}=1)}{P(N_{ij}=1|L_{ij}=1)P(L_{ij}=1)} \quad (3)$$

$$=\frac{P(N_{ij}=1|L_{ij}=1,M_{ij}=1)P(M_{ij}=1|L_{ij}=1)}{P(N_{ij}=1|L_{ij}=1)} \quad (4)$$

I make my first assumption at this point, that the probability of having a name match, conditional on being the same entity and having a link match, is 1, which simplifies the expression to that in Equation 5. (I discuss departures from this assumption below). I then use the law of total probability to decompose the denominator (the probability of a name link match on a link match), into analogous probabilities conditional on being the same and not being the same entity (Equation 6). I then use the same assumption to simplify to Equation 7.

Assume $P(N_{ij}=1|L_{ij}=1,M_{ij}=1)=1$

$$=\frac{P(M_{ij}=1|L_{ij}=1)}{P(N_{ij}=1|L_{ij}=1)} \quad (5)$$

$$=\frac{P(M_{ij}=1|L_{ij}=1)}{P(N_{ij}=1|M_{ij}=1,L_{ij}=1)P(M_{ij}=1|L_{ij}=1)+\ P(N_{ij}=1|M_{ij}=0,L_{ij}=1)P(M_{ij}=0|L_{ij}=1)} \quad (6)$$

Assume $P(N_{ij}=1|L_{ij}=1,M_{ij}=1)=1$

$$=\frac{P(M_{ij}=1|L_{ij}=1)}{P(M_{ij}=1|L_{ij}=1)+\ P(N_{ij}=1|M_{ij}=0,L_{ij}=1)P(M_{ij}=0|L_{ij}=1)} \quad (7)$$

I then make the assumption that the probability two entities who are not the same have the same name, conditional on them having a link match, is the same as the unconditional probability that two different entities have the same name. This allows me to simplify to Equation 8, and then I rewrite in order to have like terms in Equation 9.

Assume $P(N_{ij} = 1 | M_{ij} = 0, L_{ij} = 1) = P(N_{ij} = 1 | M_{ij} = 0)$

$$= \frac{P(M_{ij} = 1 | L_{ij} = 1)}{P(M_{ij} = 1 | L_{ij} = 1) + P(N_{ij} = 1 | M_{ij} = 0) P(M_{ij} = 0 | L_{ij} = 1)} \quad (8)$$

$$= \frac{P(M_{ij} = 1 | L_{ij} = 1)}{P(M_{ij} = 1 | L_{ij} = 1) + P(N_{ij} = 1 | M_{ij} = 0)(1 - P(M_{ij} = 1 | L_{ij} = 1))} \quad (9)$$

I then define α as the probability that two entities that are not the same have a name match. I define β_k as the probability of a match, conditional on a link (Equation 10), and I estimate this as the proportion of connected entities that share a name match (A_k) times the probability that a name match and a link match imply a true match, (μ_k), which is the value we are originally try to estimate in Equation 1.

$$\alpha = P(N_{ij} = 1 | M_{ij} = 0) \quad (10)$$

$$\beta_k = P(M_{ij} = 1 | L_{ij} = 1) \quad (11)$$

$$\beta_k \approx \frac{\sum_{i \neq j} \mu_k N_{ij}}{\sum_{i \neq j} 1} = \mu_k A_k \quad (12)$$

I then insert β_k and α into Equation 9, producing Equation. I apply Equation 12 (the estimate for β_k) to produce Equation 14. I then rearrange the terms to isolate μ_k in Equation 15, and I rearrange to isolate A_k in Equation 16.

$$\mu_k = \frac{\beta_k}{\beta_k + (1 - \beta_k)\alpha} \quad (13)$$

$$\mu_k = \frac{\mu_k A_k}{\mu_k A_k + (1 - \mu_k A_k)\alpha} \quad (14)$$

$$\mu_k = \frac{A_k - \alpha}{A_k - A_k \alpha} \quad (15)$$

$$A_k = \frac{\alpha}{1 - \mu_k + \mu_k \alpha} \quad (16)$$

In equation 17, I insert my estimate for α, or the probability of a name match conditional on being different entities (see derivation below), and I apply a minimum threshold of 0.999 for the true positive match rate. I then estimate the value of A_k, or the proportion of linked entities that

share a name, that is necessary for that true positive rate, and calculate 0.00237. I then use that threshold in my matching rules (see Appendix F.5).

$$A_k \geq \frac{2.373 * 10^{-6}}{1 - (0.999) + (0.999) * (2.373 * 10^{-6})} \quad (17)$$

$$A_k \geq .00237 \quad (18)$$

F.3. Effects of Violations to Assumptions

Two assumptions are made in the above derivation: (1) that true matches imply name matches and (2) that the probability that two *different* entities have the same name is the same unconditionally as when they have a link match. Below, I consider the degree to which departures from these violations would affect the estimate of what A_k is required to obtain a threshold true match rate.

First, Equation 19 defines the probability of a name match, conditional on a link match and being a true match, as X rather than 1, and Equation 20 rewrites Equation 7 with that definition. Equations 21 through 23 then show the same derivations as Equations 14 through 16.

If we cannot assume $P(N_{ij} = 1 | L_{ij} = 1, M_{ij} = 1) = 1$

$$P(N_{ij} = 1 | L_{ij} = 1, M_{ij} = 1) = X \quad (19)$$

$$= \frac{P(M_{ij} = 1 | L_{ij} = 1) * X}{P(M_{ij} = 1 | L_{ij} = 1) * X + P(N_{ij} = 1 | M_{ij} = 0, L_{ij} = 1)(1 - P(M_{ij} = 1 | L_{ij} = 1))} \quad (20)$$

$$\mu_k = \frac{\mu_k A_k X}{\mu_k A_k X + (1 - \mu_k A_k)\alpha} \quad (21)$$

$$\mu_k = \frac{A_k X - \alpha}{A_k X - A_k \alpha} \quad (22)$$

$$A_k > \frac{\alpha}{X - \mu_k X + \mu_k \alpha} \quad (23)$$

Violations to the assumption would mean that true matches do not necessarily imply name matches, that two entities that are really the same may have different names in some contexts. This is undoubtedly true, but the departure is likely to be relatively small. For example, if a person writes their name as “John Smith” 999 times, and “Jonh Smith” 1 time, and all 1000 are connected to the same link, then the probability of a name match, conditional on them being

linked and being the same, is about 0.999. If we assume a very conservative value of 0.90, and use the original values for α and μ_k of $2.373*10^{-6}$ and 0.999, we obtain a minimum value for A_k of 0.00263, rather than 0.00237. It is slightly higher, indicating that the link matches need to be more discerning to obtain the same minimum true positive rate, but they are very close. Even if the probability of a name match conditional on a link match and the two entities being the same were only 0.50, the required threshold would only increase to 0.00472.

Departures to the second assumption mean that the probability of a name match, conditional on two entities being different and having a link match, is not the same as the unconditional probability that two different entities have the same name. This is likely true in some cases, since two people that share a connection to an address may be family members, which increases the likelihood that they have a fuzzy name match, or they may be parent and child with the same names. Also, when we consider companies below, it is likely that distinct companies attached to the same address have similar names.

If we cannot make this assumption, then we can define the probability of a name match as Y times greater when there is a link match, as shown in Equation 24. We can then insert that value into Equation 7 above, producing Equation 25. We can then recalculate the estimates of true positive rate (Equation 26) and of the threshold link quality (A_k) necessary to obtain a particular true positive rate (Equation 27).

$$P\left(N_{ij}=1\middle|M_{ij}=0,L_{ij}=1\right)=Y*P\left(N_{ij}=1\middle|M_{ij}=0\right) \quad (24)$$

$$=\frac{P\left(M_{ij}=1|L_{ij}=1\right)}{P\left(M_{ij}=1\middle|L_{ij}=1\right)+Y*\ P\left(N_{ij}=1\middle|M_{ij}=0\right)P\left(M_{ij}=0\middle|L_{ij}=1\right)} \quad (25)$$

$$\mu_k=\frac{A_k-\alpha Y}{A_k-A_k\alpha Y} \quad (26)$$

$$A_k>\frac{\alpha Y}{1-\mu_k+\ \mu_k\alpha Y} \quad (27)$$

Violations to this assumption inflate both the numerator and denominator of this estimate, mitigating its effect. Accordingly, even if name matches are two times more likely (for distinct entities), conditional on having a name match (and we assume the same true positive threshold and unconditional probability of a false name match), the required link quality threshold (A_k) only increases from 0.00263 to 0.00472.

These analyses demonstrate that although violations to the assumptions are likely, they would have to be unrealistically large to substantially affect our estimates of the true positive rate.

F.4 Estimating α

From *WOA_functions/other_tasks/miscellaneous/calculate_probability_same_name.R* on Coale.

I estimated α, or the probability that two random distinct people share a name, using datasets detailing first name and surname frequency from the Social Security Administration and US Census, respectively. The first name dataset (https://www.ssa.gov/OACT/babynames/limits.html), details first name frequency, by year, excluding names with fewer than 5 occurrences in "any geographic area." I chose to use frequencies from 1970, which, since my data spans 2005 to 2020, pertains to people aged 35 to 50 during my study period. The dataset contains information for 3,606,829 people, but, according to public data sources, closer to 3.7 million people were born in 1970. I account for the additional 100,000 by assuming that each name had only 4 observations total. I calculated the probability of a first name match by calculated the sum of squared name frequencies divided by the total number of names squared.

I calculated the probability of two different individuals having the same name using data from the US Census (https://www.census.gov/topics/population/genealogy/data.html), which detailed the surname frequency for all Census respondents in 2010, censoring surnames with fewer than 100 observations. This provided surnames for 265 million individuals, but omitted names for an additional 29 million, for whom I assumed a frequency of 99, each. I then calculated surname frequency, analogously, by dividing the sum of squared surname frequencies by the total number of surnames, squared.

Based on this methodology, I calculated the probability that two people who share a first name as 0.00421, and the probability they would share a last name as 0.000563, and the probability of a shared full name as their product, 2.373*10-6. One issue is that this assumes independence between first and last names, which is unlikely.

I also estimate alpha values for fuzzy matching, or the probability that two people would have similar names. Since fuzzy matches are made between full names, a closed form solution would require me to consider every combination of first and last name, and their Levenshtein distance from every other combination, which would be computationally unfeasible. Accordingly, I sampled 5,000 first and last names, based on each's probability, and combined them pairwise, creating a random sample of 5,000 first and last name combinations. I then calculated the Levenshtein distance between each of the 25,000,000 combinations of the first and last names, and calculated the proportion that were within different thresholds of similarity. I repeatedly sampled groups of 5,000 (or 25 million combinations), until the estimates converged. My estimates of α for fuzzy matches with a 0.95 and 0.90 Levenshtein similarity proportion were 1.50 and 7.80 times larger than those for exact matches.

F.5 Details about Matching Rules

I matched any two person entities that:

- Were connected to the same parcel (owned the parcel, was an officer or registered agent for a company that owned the parcel, or filed an eviction at the parcel) that had a A_k value above 0.0023, and had the same full name.
- Were connected to the same person/company (the focal person was an officer of the company, registered agent of the company, or co-officers with the person) that had a A_k value above 0.0023, and shared a full name.
- Were connected to the same address (listed the address as a contact address, listed the address as a principal address, is officer or registered agent for a company that lists this address as a contact or principal address, owns a property at the address, is officer or registered agent for a company that owns a property at the address, filed an eviction or works for a company that filed an eviction at the address), that had a A_k value above 0.0023, and had the same full name.
- Were connected to the same parcel, entity, or address (same connections as above) that had a A_k value above 0.0179, and had the same full name. I adjusted the alpha to account for the greater probability that two people would have similar names (than would have the same name) as discussed above in "Estimating alpha."
- Person 1 was connected to an address (same connections as above) that was connected to an entity (same connections as above) that was connected to Person 2, where the two-node link had a A_k value above 0.0023, and the two people had the same full name. This includes connections like Person A listing a contact address that was also listed by a company, one of whose officers is Person B.
- Person 1 was connected to an entity that was connected to an address that was connected to an entity that was connected to Person 2, where the two-node link had a A_k value above 0.0023, and the two people had the same full name. This includes connections like Person A working for a company that listed a contact address that was also listed by another company that has person B as an officer.

After each match, I combined entities that had been matched, to reduce computational complexity and allow for the multi-node matches described above. (The multi-node matches rely on individuals having multiple connections to other objects, such as owning a parcel *and* working for a company. Having multiple connections is only possible after combining entities, since initially entities typically have only one or two connections (i.e., owning a parcel and having a contact address, or working for a company and having a contact address). After recombining, I kept distinct which connections generated from which initial entities, in order to accurately estimate A_k for each node-based link.

F.6 Matching Setup for Businesses

As discussed in the main text, matching businesses differs in several ways from matching people, but some of the same intuition for drawing on networked links and evaluating how well they identify matches, can be retained. I define X_k as all businesses deriving from the tax records whose names indicate they should have a business filing (meaning they are not trusts or banks) and I define those companies deriving from business filings as Y_l. Because the elements of X_k are not trusts, I assume that for all k, there exists some l such that there is a match (M_{kl}=1). I then create n comparisons C_{nkl}, such that if two entities are a match (M_{kl}=1), then the comparison should be true (C_{nkl} = 1). For example, one comparison could be that they have the same name, omitting business signifiers. If, for a given business in the tax records X_k, there exists only one Y_l for which C_{mkl} is true, then M_{kl} must be true as well. Based on this approach, I link entities where there is a unique match in the business filings based on a Levenshtein ratio above 0.90 and a shared contact address. I also match when there is a unique match in the business filings based on the name without business signifiers and without common words like "properties." These unique matches can be considered a special case of the name-and-link matches used for person entities. I use thresholds for the nodes, similar to those discussed in the person matching, to limit potential bias that could arise if the assumption that M_{kl} implies C_{mkl}. The precise logic of the approach is summarized below.

$$\text{For all k, there exists at least one l such that } M_{kl} = 1 \quad (28)$$

$$\text{If } M_{kl} = 1, \text{ then } C_{kl} = 1 \quad (29)$$

$$\text{If } C_{kl} = 1 \text{ for only one k for a given l, then } M_{kl} = 1 \quad (30)$$

Based on this setup, I matched any two businesses that:

- Had the same exact full name (including business type signifiers like LLC, Inc.)

And I linked any non-trust companies from the tax assessment records that were not linked to the business filings, to a company in the business filings, if it was the only company in the business filings that matched in one of the following ways:

- They were linked to the same address or parcel (with A_k threshold above .00345) and their full names had a Levenshtein ratio of 0.95 or higher.
- They were linked to the same address or parcel (with A_k threshold above .01794) and their full names had a Levenshtein ratio of 0.90 or higher. (Doing these two matches separately is necessary because the former may identify a unique match that the latter doesn't, and the latter identifies many matches that the former doesn't.)
- They shared the same short name (full name minus business signifiers like LLC, Inc., etc.)

- They shared the same "proper name" (short name minus common words like properties, holdings, etc.)
- They shared the same alphabetized "proper name."

I used A_k thresholds identical to those used in the person matching, even though they do not apply directly to company matches and, technically, the company matches should be legitimate if they are unique, regardless of the quality of the matching node, just to ensure that the nodes being matched on were not extremely common. This is particularly important if the assumption that two entities that are the same will necessarily have a true value for each comparison (29) is violated. In such a case, it would be possible to identify false matches as unique matches, since the true match would not have occurred.

The condition that the matches be unique is essential in this analysis. For example, if there were a company in the tax records named "Smith Properties" that didn't match to any business filings, then linking to business filings based on the proper name "Smith" is likely to produce a large number of matches. Since none of them would be unique, none of them would be made. However, a company with a more unusual name, like, "Michael Martin Smith's Lovely Properties," whose proper name would be "Michael Martin Smith Lovely," would match to a business filing named "Michael Martin Smith Lovely Apartments" if it were the only matching business filing.

G. Conglomerate Construction

G.1 Exact Information Contained in Business Filings and Possibility of Officers Not Being Owners

Officership data in the business filings is an imperfect indicator of ownership. Texas, Florida, and Massachusetts all require LLCs to detail their managers in their business filings, but in theory, these managers do not need to be the owners of the LLC (also called members). Filings for corporations typically list the president, secretary, and treasurer of the corporation, who are also not necessarily owners, but often are. Partnerships and associations typically list the owning partners, however, in many cases there are limited partners whose identities are not revealed.

Nevertheless, assuming officers are owners appears to be very accurate. I reached this determination in two ways, for smaller and larger landlords separately, through hand-checking hundreds of landlord companies.

Among small-scale owners who use companies, LLCs are by far the most common choice of corporate form. Typically, a single person or two business partners are managers for a range of LLCs, often one for each property they own. Several factors suggest that in these types of corporate configurations, the LLC managers are owners. First, if they were not owners, it would

mean that they were managing properties on behalf of a range of true owners, and we would expect to see variation between the multiple companies reflecting their distinct ownership. For example, we might expect the companies to have different contact addresses, principal addresses, or officers from one another, linking them to their true owners. These types of ownership configuration do exist in some cases, in which an officer of a company is the central node surrounded by spokes connecting to companies that are otherwise disconnected from one another. However, these types of ownership configurations are rare, and much more common are instances where the multiple companies managed by an individual share many pieces of overlapping information. Typically, the multiple companies that an individual manages have the same principal addresses, contact addresses, and often similar names, and this density of overlapping information suggests that these owners are not front-facing managers for multiple hidden owners. In contrast, business filings for entities like nonprofit associations, school boards, and other civic associations where officership does not convey ownership, frequently show this hub-and-spoke configuration. Although these configurations are rare among landlords, I also include a conglomerate matching condition in which those instances where officers are connected to a range of disconnected companies are removed. Second, in dozens of cases I drew on the original business filing images, which in some cases detail the names of the original members. In nearly every case, those members were identical to the managers listed in the digitized records. Finally, for the largest of these informal owners (e.g., Fred Starikov and Stephen Whelan in Boston), their identities as landlords can be verified by Googling them.

Among large-scale owners, corporations are frequently used in addition to LLCs, allowing identification of the presidents, treasurers, and other executives. I hand-checked dozens of these conglomerates, googling them and examining the profiles of their top executives on their corporate websites, where possible. In nearly all cases, the central officers identified through the business filings were top executives on the websites—frequently the founders, presidents, and chief counsels. Of course, executives are managers and not necessarily owners, but in real estate, frequently even large corporate landlords have identical owners and managers. For example, family firms and founder-managed companies are extremely common in real estate. One exception is publicly traded companies, where an identity between ownership and management clearly does not exist. However, publicly traded landlord companies, or Real Estate Investment Trusts, are relatively rare.

Finally, even in those rare instances where officership does not accurately identify owners (such as REITs), the officership information is still very likely to accurately identify links between the members of a conglomerate. This is most common in large, formal landlord organizations, where the top executives listed in the business filings may not own the company in its entirety. In these cases, linking based on these executives is nevertheless legitimate, since even if they do not fully own the company, the multiple companies for which they are officers are almost always part of the same conglomerate.

G.2 Determining Non-Unique Addresses

Addresses can be an extremely useful piece of information for linking the entities that constitute a single conglomerate. Landlords frequently list the same contact address for multiple companies, whether informal owners listing their home address for different shell companies, or formal companies listing a corporate headquarters. Accordingly, using addresses to link entities can help to clarify the dense clusters within otherwise sparse corporate networks and can often be the only way to identify connections for companies that did not connect to business filings. However, landlords also may list third-party property managers or lawyers' officers as their contact addresses, or the addresses listed for their corporate headquarters may refer ambiguously to a corporate park containing multiple companies. In both cases, using addresses to link entities would introduce many false positives.

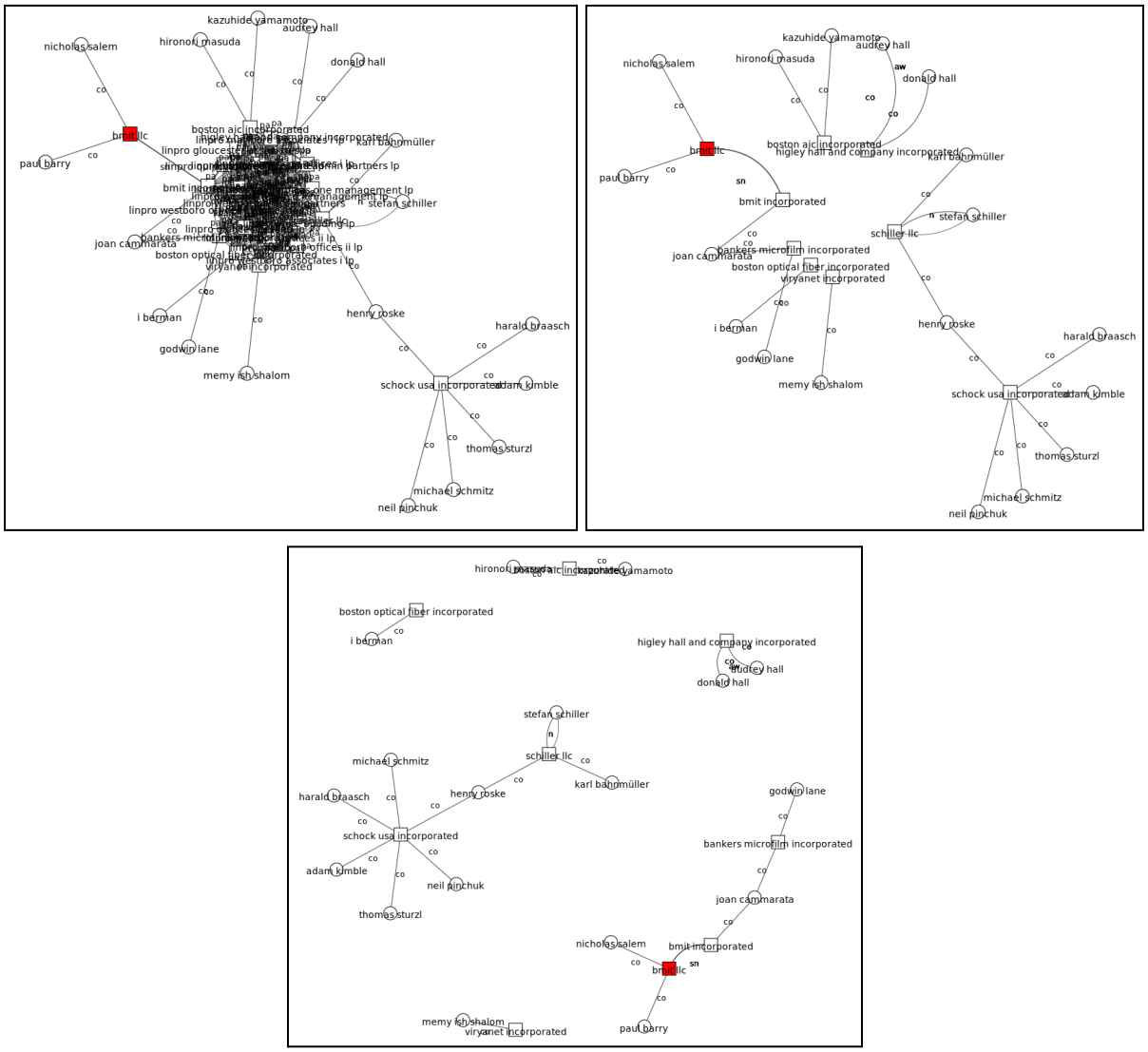
To distinguish between addresses that are unique to a single conglomerate and should be used to link entities from those that are ambiguous and should not, I constructed several measures. (These measures were needed to make conglomerate determinations and thus could not draw on whether entities were in the same conglomerate). First, to determine when principal addresses were unique to a conglomerate, I calculated, for each address, what proportion of those companies that listed it as a principal address shared an officer. Accordingly, if 25 companies listed a single address, and one individual was an officer for 20 of them, I gave the address a "principal address-officer-closure score" of 0.80. I created analogous measures for name and contact address closure (i.e., the proportion of companies that listed a principal address that had similar names, and the proportion that had the same contact address).

Because some companies did not connect to business filings and thus do not list principal addresses, I created analogous measures based on contact address. Specifically I created, for each address, the proportion of companies that listed that address as a contact address that had the same principal address, that had the same officers, and that had similar names.

These measures allowed me to determine whether to link a company to other companies based on a shared address. For example, if I had a company that was missing business filing information (and thus could not be linked based on officers), but had an accompanying address, I would look at whether the other companies, which did have officer data, had the same officers. If they did, I would connect the original company to them, and if not, I would not connect it. I hand-checked many random samples of these addresses and looked them up to determine whether this method was accurate. It appeared to be highly accurate, identifying addresses pertaining to lawyers' offices, corporate service companies, third-party property managers, and corporate parks that were home to multiple companies as invalid addresses for matching.

An example of this process is demonstrated below in Figure G1. The first plot shows a network of potential conglomerate connections including a dense cluster produced by address matches. This address was deemed to be invalid for matching, and the second plot shows the remaining network after those links are removed. As can be seen, the groups of entities that were previously connected are now largely disconnected, suggesting the address was the only source of their connections. Finally, in Plot 3, the entities are rearranged to reflect their clustered without the address matches.

Figure G1: Example of Omitting Faulty Address Matches



I also constructed similar measures for officers of companies and registered agents, determining how much their companies resembled one another, to determine whether they were third-party managers rather than the true owners. This was particularly useful for determining the rare instances where registered agents could be assumed to be owners. If a registered agent was only

registered agents for companies that shared similar names, addresses, and officers, it suggested that they were members of the conglomerate themselves.

G.3 Discussion of Types of Matches and Resulting Options for Conglomerate Construction

Below I detail the types of matches used to create the network of links out of which conglomerates are identified. For each, I discuss what the link means, its potential for creating false negatives and false positives, and the different options for using the link.

G.3.A. Officer Matches

The most obvious way to link entities to the companies for which they are officers. By extension, we can also create links between companies that share officers and between officers of the same companies.

- *False negatives:* This can fail to identify some matches because companies that do not make state level business filings, such as banks and trusts, and those companies that could not be matched to business filings, lack officership information. Among companies for which we have business filings, in some cases the ownership data is missing (Baltimore), and in some cases whether the officers do not appear to be owners and are removed.
- *False positives:* Two entities may be officers of the same company, who are not actually part of the same conglomerate, as in the case of joint ventures, and the two officers should not be matched. Some companies, such as civic associations, non-realty commercial establishments, and many others, are irrelevant to our purposes. However, being overly conservative in removing matches (e.g., removing all non-property-owning companies) may result in the loss of important information (e.g., if two people own a company that owns a property, the fact that they co-own multiple other companies together, even if those companies are not property-owning, suggests that they are part of the same conglomerate.)
- *Different options*:
    - Whether to omit certain officers if they do not appear to be affiliated with the company: We can remove matches if officers appear to be connected to a wide range of dissimilar companies, suggesting that they are a third-party manager or lawyer for these companies.
    - Which companies to include: Including companies pertaining to organizations like community associations, churches, educational associations, hardware stores, and others can introduce irrelevant links that obscure the true conglomerate clusters.
- Notes

- Question of whether to add connections between co-officers and same officers - i think yes because otherwise it's a 2 step match which isn't really right - but have thresholds on these, it's not just any that share officers

G.3.B. Registered Agent Matches

Registered agents are frequently third-party entities, such as lawyers or corporate service companies, that should not be considered to be part of the same conglomerate as the attached company and that should not be used to link companies. However, in many cases they are members of the conglomerate, often the officers themselves or, for larger conglomerates, the chief counsel of the firm, and in instances where officer data is unavailable (i.e., Baltimore), such information is indispensable.

- *False positives:* Registered agents are frequently third-party individuals who should not be included in conglomerate matching.
- *Different options:*
    - When to consider a registered agent a part of a conglomerate: We can use information about the companies associated with a conglomerate, as well as its name and the addresses it is associated with, to identify third-party registered agents.

G.3.C. Name Similarity Between Companies

In many cases, companies that make up the same conglomerate have similar names. For example, several of the companies that make up Beacon Properties are named: "Beacon Properties LLC" "Beacon Properties Inc." "Beacon Properties Acquisitions Corporation" "Beacon Properties Acquisition LLC" "Beacon Properties Corporation" and "Beacon Properties Limited Partnership." This information can be very useful, particularly for matching companies that do not have ownership information.

- *False negatives*: Not all companies that are in the same conglomerate have similar names. For example, the same person might name each LLC they own after the street address of the property it owns, e.g., 10 Main Street LLC, 24 Beacon Street LLC, etc.
- *False positives*: It is possible to match on overly common words, so when using this type of match, one has to be careful to avoid generic words that do not really convey being in the same conglomerate. In some cases this is obvious. For example, we should obviously not match on words like "Properties" or "Boston." However, in some cases, highly distinctive names, like "Sunset Vista Homes" may be shared by two companies because it refers to a property development that each company owned at a different time. These are more difficult to identify.
- *Different options*
    - Only match based on name if there is another node connecting the companies
    - Only match based on name if there are multiple nodes connecting the companies
    - Only match if the node has a value of A_k above a certain threshold
    - Match based on full name, short name, or proper name

- Use exact matching or fuzzy matching

G.3.D. Companies Share an Address

Often a conglomerate will have a single address that most or all of its businesses list as their principal or contact address, and this can identify matches that would otherwise be invisible, particularly in cases where businesses' owners cannot be identified. Methods that connect landlords based only on tax assessments frequently rely on these types of links (An et al. 2022; Gomory and Desmond 2023).

- *False negatives*: Businesses do not always use a single address however. In some cases landlords will list as their principal address the property that the LLC owns, meaning that each of their shell companies is linked to a different principal address.
- *False positives*: Businesses will often list a third-party address, such as a lawyer or property manager, and linking on these addresses would combine entities that should not be part of the same conglomerate. Even in cases where the listed address is the primary location of the conglomerate, there are cases where the address refers to a property that is the primary location of multiple businesses, such as a large commercial building or corporate park, and linking on this principal address would create false positive matches.
- *Different options*:
    - Never link on principal addresses
    - Link only on principal addresses that do not appear to be third-party locations or large commercial properties, using information about associated companies
    - Only use principal address links if there is no officer information
    - Whether to use principal addresses, contact addresses, or both

G.3.E. Name Similarity Between a Person and a Company

In some cases people name their companies after themselves, for example James Smith might own a company called James Smith Properties LLC. Identifying these cases can be very useful for connecting companies without ownership information to their owners. In particular, this is very useful for identifying the beneficiaries of trusts, because trusts do not have filing information but are often named after their beneficiaries. For example, James Smith might create a trust called "Smith Family Revocable Trust."

- *False negatives*: This fails to identify any match where a person does not include their name in the company.
- *False positives*: False positives are relatively rare for this type of match, as long as the company and person are connected in some other way in addition to the name, such as listing the same contact address or having owned the same parcel.
- *Different options:*
    - How discerning of a node to require the person and company share, in addition to the name similarity.

G.3.F. Family Members

Landlord companies frequently consist of multiple members of the same family, and even outside of real estate, putting companies and properties in family members' name is a common way that conglomerates obfuscate their holdings. By identifying and connecting family members we can ensure that we identify these connections, which might be missed by looking just at ownership.

- *False negatives*: There are many instances where landlord companies do not contain any family members, and, moreover, there are many types of links within conglomerates that are not family member links.
- *False positives*: There may be cases where family members are not actually part of the same conglomerate and their holdings should not be joined. These instances are unlikely to bias the resulting landlord conglomerate characteristics too much, however, as non-landlord family members likely hold only one or two properties, at most. There is also the possibility of linking two people who have the same last name but are not actually part of the same family.
- *Different options*
    - How discerning (A_k) must the node be, through which they share a link?

G.3.G. Aggressive Matches

Often networks of links can be greatly clarified by enacting aggressive matching within small partitions. This means adopting matching rules that, if applied to the entire dataset, would produce too many false positives, but when applied on small partitions, are effective. For example, these matches can identify all of the companies that have "Beacon" in their names or that are attached to a corporate headquarters. I apply aggressive matching based on any shared words in names and any shared connections to addresses. Since these only apply to entities that are already linked to one another indirectly, this is aimed at clarifying the underlying dense clusters that represent conglomerates rather than identifying new,

- *Different options*:
    - How big of a partition to apply the aggressive matching over
    - How many, and how common of, words and addresses to match on (typically matching on all words and addresses would create too many false positives, and we want to subset to only the most common words in a partition, which are most likely to be conglomerate-identifying words, like "Beacon")

G.3.H. Network Resolution Methods

I resolve the network of links into distinct conglomerates using community detection, specifically the modularity-maximizing Leiden algorithm as implemented by Python igraph.

- *Different options*:
    - What resolution parameter to use and when to consider a conglomerate successfully identified. Using a single resolution parameter tends to create too-large conglomerates for some and too-small conglomerates from others.

Accordingly, I identify communities iteratively, by first identifying communities with a low resolution parameter, checking whether they are sufficiently dense, then repartition those that are not, with a higher resolution parameter.

G.4. Full List of Conglomerate Construction Implementations

The considerations summarized above can be combined in many ways to produce a wide range of potential conglomerate implementations. Below, I list all of the implementations (or sets of decisions), I used when implementing conglomerates. Figure 7 in the main text shows the different estimates in scale that result from these decisions.

Often matches were limited for unimportant companies. Frequently I included only those companies that own properties, have names that indicate they are realty companies (e.g., contain "properties," "apartments," etc.), or could not be categorized (as non-profit, civic association, religious, educational, other commercial, etc.). I call these "owning, realty, and ambiguous" companies.

Base implementation

- Officer matches:
    - Only used for owning, realty, and ambiguous companies.
- Companies have same officers match
    - Only used for owning, realty, and ambiguous companies.
    - Matched if companies have at least ½ overlap of officers.
- Officers are affiliated with same companies match
    - Only used for owning, realty, and ambiguous companies.
    - Matched if officers have at least ½ overlap of companies.
- Registered agent matches:
    - Registered agent data was only used for owning, realty, and ambiguous companies.
    - Registered agent matches were only used if officership data was unavailable for the company (typically this means it was from Baltimore)
    - Registered agent matches were dropped if the agent was a known third party agent (e.g., CT Corporation Systems, but also more locally specific ones) or did not appear to be unique to a particular conglomerate, specifically, if it was an RA for multiple companies, that listed at least three distinct principal addresses, and the number of distinct principal addresses those companies listed were more than one-quarter of the total number of companies.
- Name and node
    - Only used for owning, realty, and ambiguous companies.

        - Matches were made based on proper name, with a fuzzy match Levenshtein ratio threshold of.
        - Matches were only made if the link discernment estimate was above 0.02 for exact matches and above 0.15 for fuzzy matches.
        - Matches were only made if the companies connected based on two distinct nodes (e.g., two different shared addresses, a shared address and a shared officer).
- Principal address matches
    - Only used for owning, realty, and ambiguous companies.
    - Only used if officership data was unavailable.
    - Address was only used if the officer closure value (see address categorization below) was above 0.75.
- Contact address matches
    - Only used for owning, realty, and ambiguous companies.
    - Only used if officership data and principal address data was unavailable.
    - Address was only used if the officer closure value (see address categorization below) was above 0.75.
- Family member matches
    - Only used if link discernment estimate was above 0.02.
- Person-company name similarity
    - Only used for owning, realty, and ambiguous companies.
    - Only used when fewer than 100 comparisons were made for a given node.
- Co-ownership of properties match
    - Only used if entities co-owned at least ¾ of their parcel-years.
- Aggressive matches
    - Matches based on both words and addresses
    - 8 most common words were used, as long as each constituted at least 5% of all words
    - 5 most common addresses were used, as long as each constituted at least 10% of all addresses
    - Aggressive matches were only made between entities that were not otherwise linked directly (by definition they were linked indirectly).
- Weighting
    - Use the number of instances where a relationship appeared in the data as the weight, with a maximum weight of 5
- Network resolution
    - Approve conglomerate clusters once they have a one-step density of at least 0.3 and a two-step density of at least 0.6.

No suspicious officers

- Identical to base implementation, but omits all officer matches for officers who are officers for at least five companies, where those companies do not share principal addresses at least half of the time.

No address match

- Identical to base implementation, but omits all contact address and principal address matches.

No aggressive matches

- Identical to base implementation, but omits all aggressive matches.

No family matches

- Identical to base implementation, but omits all family matches.

Officers only

- Uses only officer matches, as implemented in base implementation.

Base implementation - .35, .65

- Identical to base implementation, but it requires one-step density of 0.35 and two-step density of 0.65.

Base implementation - .4, .7

- Identical to base implementation, but it requires one-step density of 0.4 and two-step density of 0.7.

Base implementation - .5, .8

- Identical to base implementation, but it requires one-step density of 0.5 and two-step density of 0.8.

Base implementation - .55, .85

- Identical to base implementation, but it requires one-step density of 0.55 and two-step density of 0.85.

Base implementation - .6, .85

- Identical to base implementation, but it requires one-step density of 0.6 and two-step density of 0.85.

Base implementation - .7, .9

- Identical to base implementation, but it requires one-step density of 0.7 and two-step density of 0.9.

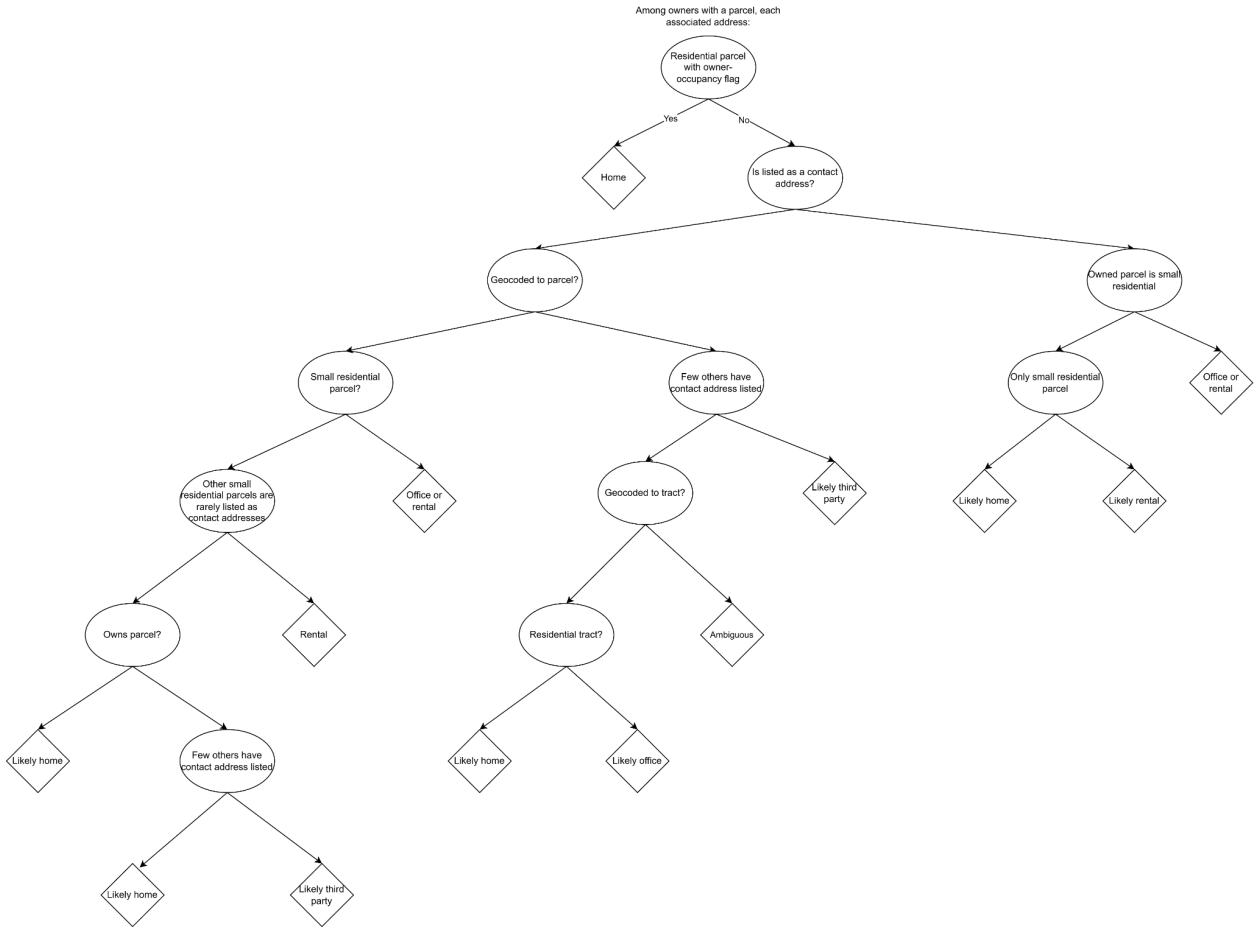
H. Address Categorization

People and companies listed contact addresses in the tax records and business filings (when officers), and businesses listed principal addresses in the business filings. I categorized these different addresses as (1) homes, (2) likely homes, (3) offices, (4) likely offices, (5) rental properties, (6) third-party locations, (7) PO boxes, or (8) ambiguous. To do so, I first geocoded all contact addresses and principal addresses to the parcel level (if they were located within a metro area, meaning I had appropriate parcel data) or to the census block level if they were not (using the Census geocoder's free online API).

If first removed PO box addresses and labeled them as such. I then labeled partial addresses as ambiguous. I chose not to drop these entirely because they still sometimes contain useful information, such as the street name, city name, or zip code of the address. I categorized the remaining addresses according to the following rules (visualized in Figure H1 below):

I considered addresses linked to small residential parcels to be much more likely to be home addresses than those linked to large multi-family properties, and (obviously) commercial or other addresses. I considered addresses linked to small parcels to be more likely to be homes if the owner rarely (<25% of the time) listed other small properties as contact addresses, in order to identify cases where an owner lists their rental properties as contact addresses. I considered addresses to be more likely to be third-party addresses if multiple conglomerates listed the same address. I considered addresses to be more likely to be homes if they were located in highly residential tracts, with more than 75% single-family properties or more than 75% homeowner households.

Figure H1: Flowchart of Address Categorization Rules



I. Missing and Imperfect Data

I.1. Correlates of Missing and Imperfect Data

Table I1 shows basic descriptive characteristics of parcels for which different owner characteristics are missing or imperfectly constructed. Overall, observations with imperfect or missing data are more likely to be those owned by companies and to be in large multi-family properties. These observations are also more likely to be in Baltimore, where officer data is missing from the business filings. Low data-quality observations resemble the full sample to a greater degree than do missing observations.

Table I1: Descriptive Characteristics of Observations Missing Data, Weighed by Rental Units

	All	Num officers Missing	Scale Low DQ	Home address Missing	Home address Low DQ
Large multi-family property	0.38	0.59	0.61	0.70	0.63
Metro: Boston	0.18	0.13	0.20	0.12	0.19
Metro: Baltimore	0.11	0.42	0.14	0.17	0.10
Metro: Miami	0.39	0.21	0.38	0.32	0.32
Metro: Houston	0.32	0.24	0.27	0.38	0.39
Direct owner is not a person	0.48	0.82	1.00	0.89	0.75
Rental units	44903464	11496055	4558264	13920949	2330167
Prop of all units	1.00	0.26	0.10	0.31	0.05

	All	Primary contact Missing	Primary contact Low DQ	Imputed race Missing	Imputed race Low DQ
Large multi-family property	0.38	0.81	0.56	0.74	0.35
Metro: Boston	0.18	0.16	0.13	0.11	0.18
Metro: Baltimore	0.11	0.00	0.16	0.23	0.07
Metro: Miami	0.39	0.26	0.35	0.31	0.46
Metro: Houston	0.32	0.59	0.36	0.35	0.30
Direct owner is not a person	0.48	0.80	0.70	0.98	0.43
Rental units	44903464	1743458	18464206	9744467	15610984
Prop of all units	1.00	0.04	0.41	0.22	0.35

I.2. Imputing Missing Data

I imputed missing data using a random forest estimated at the conglomerate level, drawing on a range of information about each owner, including: their owned properties' average land usages, property values, tract racial compositions, tract socioeconomic positions, and owner-occupancy rates; their home address's location category (central city, in-metro, etc.) and tract characteristics; analogous variables for their primary contact location; their scale of ownership; and other miscellaneous characteristics. Models were estimated only for non-government owners and owners who could feasibly be missing data (e.g., the number of company officers was imputed only for owners with focal companies, not for person-owners, who by definition have no officers). 20% of the non-missing set was omitted from the training data and was used as a test dataset.

I.3. Creating Low Data-Quality Versions

Observations were deemed to have low data-quality for particular variables if those variables were constructed under suboptimal conditions. To test the degree to which these values differed from true values, I estimated low data-quality versions of variables for observations that had high data quality.

I.3.A. Scale

Scale is never missing, but it may be inaccurate if the entities that constitute a conglomerate were combined incorrectly. Typically, this is a problem of under combining, rather than over combining. Because there is no ground truth to compare to, I cannot perfectly estimate when landowners are imperfectly combine, but observations that were not connected to business filings are more likely to be imperfectly combined. They may still have been able to be linked to their other constituent entities, through their contact address, parcel holdings, name, or other information, but because they lack data on officers, registered agents, and principal addresses, they are more likely to be under matched. Accordingly, I labeled any owners for whom none of their owning companies were linked to a business filing and labeled them as low data quality. To estimate the value for landlord scale that would have been produced under this low data-quality condition, I re-constructed landlord conglomerates for all observations without using principal address, officer, or registered agent data, and calculated the resulting scale values. This allowed me to see, for those observations that were connected to business filings and thus were more likely to be accurately combined, how much omitting that business filing data affects their estimates of scale.

I.3.B. Home Address

As shown in Table 6 in the main text, and discussed in greater detail in Appendix H, I labeled the contact and principal addresses that entities listed as homes, likely homes, offices, likely offices, likely third-party locations, rental properties, PO boxes, or ambiguous addresses. I considered home data to be missing for a person if they did not list a contact address that was labeled as a home, likely home, or ambiguous address, and those that only had an ambiguous address I

labeled low data-quality. I then identified those owners who listed a likely home or home (high data-quality), as well as an ambiguous address (low data-quality), and compared the latter to the former.

I.3.C. Primary Contact Address

Primary contact address was missing for those landowners who listed only addresses that were likely third-party locations, likely rental properties, or PO boxes (leaving homes, likely homes, offices, likely offices, and ambiguous addresses). I labeled those that were ambiguous addresses as low data quality, and I identified those owners who listed both a high quality primary contact address, as well as a low quality one, and I compared the characteristics of the two operationalizations.

I.3.D. Race Data

Race data was deemed low quality if only one or two of the three pieces of information (first, last, and geography) were used to impute race. I created low data quality measures for those with high data quality by re-estimating their racial identities using only their last and first names, and I compared these low quality versions to ones that used all three pieces of information.

I.4. Reweighting to Approximate Missing and Low Data Quality Observations

When estimating the per-rental-unit bias created by using imputed and low quality versions of observations, I reweighted the test sample to resemble those observations that were missing data or had low data quality. I did this because it is quite likely that the bias is heterogeneous. For example, low data quality versions of scale are, by definition, only different from the high data quality versions for those observations missing business filing data, and imputed values of home address are likely to be less accurate for those observations where we are missing data on any person in the conglomerate compared to those for which we have identified the person owner but not their home. Accordingly, I reweighted test samples by the proportion of units the conglomerate owns using non-person entities (their degree of corporate ownership), since that is the main vector along which data quality and bias heterogeneity are likely to vary.

J. Baltimore as a Special Case of Missing Data

Business filings from Maryland do not detail company officers, constituting a key source of missing data, but also providing a stringent test of the degree of accuracy this methodology can produce even when data quality is low. For some variables, the degree of bias produced by this missing data can be measured using data from Baltimore alone. For example, the accuracy of owners' home addresses and racial information can still be estimated, since there are still conglomerates for which likely person owners have been identified through other means (e.g., by identifying uncommon shared addresses and similar names). The degree to which imputations are accurate for this non missing (albeit small) group can then be used to estimate the aggregate accuracy of these imputations.

However, as a further test, I re-ran my data construction pipeline for Boston, Houston, and Miami, omitting officer data from the business filings (thus mimicking the data availability in Baltimore). I then compared the estimates of landlord characteristics created in this low data quality condition to those that included business filings, in Table J1. First, by definition, the "NO," meaning "No officer" samples have no pertaining to officers, directors, or the aggregate corporate formality score, and their parcel data is identical to that in their corresponding "Officer" samples. However, among other variables there is a very high degree of similarity between the "Officer" and "No officer" datasets. The proportion of rental units owned by White landlords for the "Officer" and "No officer" samples Boston samples are 0.82 and 0.82, respectively, and the corresponding values for Houston are 0.70 and 0.73, and for Miami 0.59 and 0.61. In the "No officer" data, White ownership is overstated, but the bias is slight, as was indicated by the bias analysis in Figure 4 in the main text. Likewise, the proportion so units owned by landlords of different scales in the "No officer" data are within two percentage points of the proportions in the "Officer data," with the "No officer" data slightly understanding large-scale ownership. The home and primary contact proportions are within four percentage points in the two datasets, with the "No officer" data typically overstating local ownership (with the one exception that the "No officer" data in Houston overstates out-of-state primary contact addresses). Altogether, constructing datasets without officer information for Miami, Houston, and Boston confirm that the data construction in Baltimore is largely accurate.

Table J1: Descriptive Statistics for Metro Areas Without Officer Data

Variable	Baltimore (NO)	Boston (O)	Boston (NO)	Houston (O)	Houston (NO)	Miami (O)	Miami (NO)
Single-family	0.39	0.10	0.10	0.35	0.35	0.24	0.24
Condominium	0.06	0.11	0.11	0.05	0.05	0.38	0.38
Small multi-family	0.05	0.34	0.34	0.00	0.00	0.11	0.11
Large multi-family	0.50	0.36	0.36	0.53	0.53	0.25	0.25
Other residential	0.00	0.08	0.08	0.07	0.07	0.02	0.02
1 unit	0.45	0.22	0.22	0.41	0.41	0.62	0.62
2-3 unit	0.04	0.34	0.34	0.02	0.02	0.08	0.08
4-9 unit	0.02	0.09	0.09	0.03	0.03	0.06	0.06
10+ unit	0.49	0.35	0.35	0.55	0.55	0.24	0.24
Value per unit	103574	222730	222730	79366	79366	174198	174198
Year built	1950.15	1932.61	1932.61	1981.00	1981.00	1978.48	1978.48
Person	0.37	0.58	0.58	0.41	0.41	0.60	0.60
Government	0.02	0.05	0.05	0.00	0.00	0.01	0.01
Private organization	0.61	0.38	0.38	0.59	0.58	0.40	0.40
LLC	0.29	0.14	0.14	0.25	0.23	0.18	0.18
Direct owner is person	0.38	0.60	0.60	0.41	0.41	0.60	0.60
Direct owner linked to filing	0.46	0.25	0.25	0.50	0.49	0.30	0.30
DO person or linked to filing	0.84	0.85	0.86	0.91	0.90	0.90	0.90
Direct owner traced to person	0.00	0.23	0.00	0.40	0.00	0.27	0.00
DO person or traced to person	0.38	0.84	0.60	0.82	0.41	0.87	0.60
CG owner contains person	0.55	0.87	0.74	0.77	0.41	0.83	0.60
CG owner contains FC	0.65	0.46	0.44	0.61	0.58	0.44	0.40
CG FC linked to filing	0.50	0.28	0.28	0.43	0.49	0.30	0.30
Corporation	0.14	0.09	0.05	0.13	0.04	0.12	0.09
Num distinct offices		0.11		0.12		0.06	
Num distinct officers		0.11		0.12		0.07	
>1 distinct offices		0.40		0.43		0.22	
>1 distinct officers		0.37		0.39		0.29	
Has director		0.11		0.21		0.07	
Has corporate office	0.25	0.18	0.19	0.31	0.11	0.22	0.11
Org formality >2	0.00	0.38	0.00	0.48	0.00	0.28	0.00
1-4 units	0.37	0.51	0.51	0.36	0.37	0.60	0.62
5-19 units	0.09	0.13	0.14	0.06	0.06	0.10	0.10
20-99 units	0.09	0.12	0.13	0.06	0.06	0.09	0.09
100+ units	0.45	0.24	0.22	0.52	0.51	0.21	0.19
Home: Central city	0.35	0.29	0.30	0.42	0.44	0.16	0.17
Home: Suburbs	0.36	0.57	0.59	0.29	0.29	0.63	0.63

Home: Outside metro, same state	0.07	0.07	0.06	0.09	0.11	0.03	0.06
Home: Different state	0.21	0.07	0.06	0.19	0.16	0.18	0.14
PC: Central city	0.34	0.29	0.30	0.45	0.43	0.18	0.17
PC: Suburbs	0.34	0.55	0.57	0.27	0.26	0.56	0.57
PC: Outside metro, same state	0.09	0.08	0.07	0.13	0.09	0.05	0.06
PC: Different state	0.24	0.09	0.06	0.15	0.22	0.21	0.20
White	0.83	0.82	0.82	0.70	0.73	0.59	0.61
Asian	0.02	0.06	0.06	0.10	0.06	0.03	0.02
Hispanic	0.01	0.05	0.06	0.14	0.15	0.30	0.29
Black	0.14	0.07	0.06	0.06	0.06	0.08	0.08
Multiple	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Native	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Rental units (1000s)	4963	8320	8320	14598	14598	17728	17728