Linking Landlords to Uncover Ownership Obscurity

Preprint at SocArXiv

Forrest Hangen [1,2] ✉ **and Daniel T. O'Brien**[1,2,3]

[1]School of Public Policy and Urban Affairs, Northeastern University - Boston, USA; [2]Boston Area Research Initiative, Northeastern & Harvard Universities - Boston, USA; [3]School of Criminology and Criminal Justice, Northeastern University - Boston, USA

✉ **For correspondence:** hangen.f@northeastern.edu

Code availability: Code for cleaning, fuzzy-matching and linking available at: https://github.com/forrest-h/linking-landlords.

Data availability: All data except Corporation Records available at the Boston Area Research Initiative's Data Portal. MA Corporation Records available upon request.

Funding: This work was supported by an NSF Graduate Research Fellowship, (grant number NSF GRFP DGE 1419118).

Conflicting interests: The author declare no conflicting interests.

Work in Progress: This preprint is a work in progress and the final version may differ from the current work.

Abstract

Defining the ownership of rental housing can be a difficult task. In recent years there has been an increasing obscurity of ownership in administrative records as more property owners use Limited Liability Companies (LLCs) on deeds and in tax assessment records. In many cases, this obscures the nature and scale of ownership and makes research into property ownership, investors, and landlords more challenging. To overcome these challenges, we compare different text-matching methods within property tax assessment records in Boston, MA from 2004-2019. We show that the source of the difficulty in creating an accurate knowledge of landlords and their portfolios of properties has shifted in the past decade from the scale of data and the messy nature of administrative data to an intentional strategy of obscurity through LLCs. To do so, we incorporate linking to corporate records to uncover intentional ownership obscurity. We assess the prevalence of obscurity among landlords as well as the extent to which it is undermining our ability to observe patterns in rental housing in ways that cannot be accounted for solely with text-matching. These include how obscurity hides not only an increasing consolidation of property ownership in the past decade, but also concentrations of disorder and evictions. In doing so, we demonstrate a comprehensive method for uncovering this obscurity and show how this representation of property ownership can form the basis for understanding inequities in rental housing and the effects of property consolidation.

Introduction

In his 1966 book, The Tenement Landlord, George Sternlieb posed an important question enroute to understanding the state of urban housing and the possibilities of urban renewal: Who owns the slums? To answer this deceptively simple question for even a small sample of tenements in Newark, NJ, Sternlieb and his team of researchers manually compiled a directory of real estate owners, real estate transactions, and title deeds to determine the plausible owners of 566 parcels. The effort of this search led Sternlieb to state: "Defining the ownership of slum tenements is a far from easy task" (Sternlieb 1966, 54). As researchers today continue to ask about landlord strategies and their impact on housing conditions, eviction practices, and the overall equity of rental housing, we still must answer the question: Who owns rental housing? And the process of answering this deceptively simple question continues to be no easy task.

Recent research has elevated and highlighted the significant role landlords play in the physical and social conditions tenants experience (e.g., Desmond and Wilmers 2019; Garboden and Rosen 2019; Gomory 2021; Immergluck et al. 2020; Seymour and Akers 2020; Travis 2019). Central to this emerging understanding is a need to have an accurate knowledge of the characteristics of landlords and their portfolio of rental properties. An accurate depiction of inequities in rental housing, differences in landlord-tenant relationships, and the effects of property consolidation (i.e., more properties owned by fewer landlords) should be grounded in an accurate representation of property ownership. In the decades since 1966 there have been two significant developments affecting the process of creating an accurate representation of property ownership. First, there have been many advances in digitizing records, computing power, and open access to administrative data (Thakuriah, Tilahun, and Zellner 2017). In many jurisdictions, tax assessment data containing owners' names and addresses can be easily obtained. One might reasonably assume that this would make discovering the ownership of 566 parcels or even the parcels of a whole city a trivial task. However, there are complications for this assumption, especially when identifying consolidation. Although a single entity might own many properties, its name or contact information might not be entered the same way in every record – creating errors that needed to be corrected. Fortunately, advances in text matching methods have significantly increased our ability to deal with the issues of messy data (Tahamont et al. 2021).

The new availability of data and ability to deal with messy administrative data has not been matched by a new ease in understanding ownership because of the second significant development: an increase in the obscurity of ownership. Ownership obscurity is the result of using multiple corporate entities to make property ownership untraceable solely within tax assessment or property transactions data. Through a desire to limit personal liability for issues and property conditions, landlords have been increasingly turning to the Limited Liability Company (LLC) as the preferred form of property ownership (Travis 2019; Gomory 2021). The use of LLCs to limit liability also creates a new form of ownership obscurity: identifiable and linkable information is replaced by a corporate entity. This can effectively negate the previous advances in linking property owners – even the most advanced and accurate fuzzy-matching method will never overcome the challenge of linking two differently-named LLCs owned by the same individual. Thus, although we have better access to ownership data than ever before, ownership has been increasingly obscured even to these modern resources through the use of corporate entities.

In the current study, we show that corporate ownership obscurity can be uncovered and that a more accurate answer to the question: "Who owns rental housing?" is possible. We utilize the significant advances in data availability and text-matching methods as well as combining data on property ownership with corporate records to quantify consolidation and uncover ownership obscurity. We compare text-matching methods within property ownership data and incorporate corporate records to reveal the consolidation of property ownership that has been obscured. In doing so, we assess the prevalence of obscurity among landlords as well as the extent to which it is undermining our ability to observe patterns in rental housing in ways that cannot be accounted for solely with text-matching. We further show that this obscurity hides not only an increasing consolidation of property ownership in the past decade, but also concentrations of disorder and evictions. In sum, we show that the source of the difficulty in creating an accurate knowledge of landlords and their portfolios of properties has shifted in the past decade from the scale of data and the messy nature of administrative data to an intentional strategy of obscurity through LLCs and other corporate entities. In doing so, we demonstrate a comprehensive method for uncovering this obscurity and show how this more accurate representation of property ownership can form the basis for understanding inequities in rental housing, differences in landlord-tenant relationships, and the effects of property consolidation.

Quantifying Consolidation

Landlords with different sized portfolios of properties are likely to exhibit different property management strategies (e.g., Gomory 2021; Raymond et al. 2018). Creating an accurate understanding of the sizes of landlords' portfolios and their degree of consolidation is the first step to uncovering these strategies and their effect on a local rental housing. The first step, and traditionally the only step, in discovering the set of properties owned by each landlord is linking named entities within tax assessment records or property transactions data (e.g., Immergluck et al. 2020; Sternlieb 1966). Ideally, one would be able to simply group properties by owners with the same names and addresses to arrive at a quantification of consolidation. In practice, spelling mistakes, abbreviations, and other errors can prevent the direct linking of entities. Multiple methods are useful for correcting for these errors such as automated text cleaning and fuzzy-matching methods.

The process of linking entities together is typically known as record linkage (Herzog, Scheuren, and Winkler 2010). When referring to links within a dataset, this is more commonly known as deduplication. Record linkage has received much attention in recent years with advances in computational techniques for linking data by matching on identifying information (Tahamont et al. 2021). Record linkage typically takes one of two approaches: deterministic linkage or probabilistic linkage (Enamorado, Fifield, and Imai 2019). In deterministic linkage, a set of linking rules are established and matches are made if the criteria of the linking rules are met. The most common example of this approach is exact matching, in which the sole criterion is that all relevant identifying information is exactly the same across both entities (e.g., matching two properties both owned by exactly the same entity: 32 Greystone LLC). If stringent criteria are used—with exact matching being the most extreme example—deterministic linking will minimize incorrectly linking two entities (false positives) at the cost of increasing the likelihood of missing links between two entities that are true matches (false negatives).

Probabilistic linkage, also known as 'fuzzy matching,' estimates a probability of a match between two entities based on an underlying theoretical model (e.g., comparisons between vectors based on identifying information, phonetic comparisons, etc.; Fellegi and Sunter 1969; Newcombe et al. 1959). This often requires the researcher to set some threshold for accepting or rejecting matches – typically through manually reviewing a sample of the matches generated. If properly used, probabilistic methods can have higher overall higher rates of accuracy than deterministic methods (e.g., Tahamont et al. 2021; Tromp et al. 2011). However, this can come at the cost of introducing more false positives.

Emerging work has already shown promise in addressing the specific record linkage challenge of identifying landlords and their properties. For example, Immergluck et al. (2020) used a probabilistic method, specifically a semi-supervised learning algorithm called Dedupe to link ownership clusters within tax assessment and sales records. This process allowed them to identify probable portfolios of properties owned by large landlords in Atlanta to understand some of the effects of property ownership consolidation. In another approach, An et al. (2022) used another semi-supervised method called OpenRefine to link owners in parcel records for Fulton County, GA. Much of this work demonstrates that landlords and investors differ based on the size and composition of their portfolios and this necessarily depends on an accurate understanding of a landlord's portfolio of properties (e.g., Raymond et al. 2018; Seymour and Akers 2021). While linking solely within property records might have historically been enough to quantify consolidation, the rise of LLCs complicates record linkage and requires new strategies to uncover ownership obscurity. As we elaborate in the next section, LLCs render text-matching within tax assessments or property transactions data insufficient for completely uncovering ownership and consolidation.

Ownership Obscurity

Since the 1990s, the LLC has become increasing popular as a form of ownership over properties (Travis 2019). The use of corporate entities does not automatically make the process of understanding property ownership more difficult – 10 properties owned by Jamie Clark are just as easily grouped as 10 properties owned by Jamie Clark LLC. However, this is rarely how LLCs and other forms of corporate ownership are organized. In many cases, those ten properties would each be held by uniquely named LLCs, such as 12 Grey St LCC, 56 Green St LCC, etc. An owner with 10 properties under 10 different LLCs is at a lower risk level than an owner with 10 properties under a single LLC. Both are protecting their personal assets, but the use of multiple LLCs protects single properties or groups of properties from the potential lawsuits at other properties. In addition to lower risk of liability, owners can maintain a relative degree of anonymity. For example, any violations issued against the property or taxes unpaid will refer to the LLC, not the owners of that LLC (Demond, 2016). Both the anonymity and the liability benefits make owning properties under LLCs an adaptive strategy for landlords. It is this obscurity, where properties owned by an individual or set of individuals are recorded under multiple corporate entities or LLCs, that hinders the process of understanding who owns what despite the advent of fuzzy matching.

Ownership obscurity creates two main issues for understanding the landscape of rental housing. First, the responsibility for an individual property is hidden. This can create issues with building collective power among tenants with obscure owners. Two individuals sitting in eviction court might not know that they have a shared landlord – each having only dealt with a property manager or seen the impersonal LLC written on their eviction notice. The logistical challenges of forming tenant unions or collectively bargaining within a 100-unit building are already greatly compounded when trying to unionize across 100 single-family rental homes. This only becomes more difficult and even impossible if those 100 tenants are unaware that all of their houses are owned by one landlord. Furthermore, municipalities looking to work with landlords to mitigate disorder and crime are at a disadvantage in understanding the scale of problems and the scale of potential cooperation if they are unable to correctly determine the ownership of properties (O'Brien et al. 2022).

The second issue stemming from obscurity is that research into the role of landlords in the state of rental housing could generate an inaccurate picture if obscurity is not uncovered. Under a degree of ownership obscurity, properties that are owned by an entity are represented as being owned by separate entities, reducing the sizes of landlords and increasing the number of separate landlords. This then causes an issue in any aggregated understanding of landlord characteristics, such as the distribution of landlord sizes. Without a way to uncover the obscurity, the number of owners and the scale of their property portfolios could misleadingly show a lack of consolidation. 100 properties under 100 different LLCs could mistakenly show a diversity of ownership and lack of coordinated management over these properties when in reality those 100 LLCs and 100 properties could be managed by the same entity. This can also have implications for downstream analyses of landlords and their properties. For example, examining the eviction rates of different sizes of landlords could be skewed if landlords are not properly grouped by size because of their obscurity.

Fortunately, there is a potential solution to the obscurity created by LLCs, one that requires us to move beyond linking within a single database to linking within corporate records. In corporation records, individuals have to report their identities as the owners of LLCs, allowing us to connect 12 Grey St LLC and 56 Green St LLC as both being owned by the same individual. Linking landlords to their portfolio of properties comes with a few additional challenges that make record linkage more difficult. First, there is generally no ground truth (i.e., a set of known matches between owners and their properties) to establish accuracy rates. This makes it more difficult to determine if the linkage process has successfully identified true positives while avoiding both false positives and false negatives. Second, there are multiple possible stages of record linkage. Properties owners can be linked within tax assessment records. LLCs and other corporate entities can be linked to corporation records. Corporate entities can be linked within corporation records. Gomory (2021) was the first to take a comprehensive approach to solving this problem, using both deterministic and probabilistic methods to match entities within property tax records and within corporate records. Others have used other sources, like SEC filings to uncover ownership structures (e.g., Raymond et al. 2018). This emerging body of work shows the potential that linking landlords has for understanding strategies of property management and the effects on housing conditions, markets, and the experiences of tenants. To aid future research in this area, we will provide a comparison of linking methods and their various stages to directly assess the need for corporate data linking to supplement fuzzy matching to accurately quantify consolidation and uncover obscurity and, in turn, alter the inferences drawn about housing conditions and equity.

The Current Study

In the current study, we show the efficacy of text-matching and demonstrate the need to incorporate corporate records into the linking process to overcome the two significant methodological challenges of understanding property ownership: linking named entities and uncovering intentional ownership obscurity. First, we compare different text-matching methods in linking named entities within property tax assessment records in Boston, MA from 2004-2019 to show the variety of matches generated by these different record linkage processes. Second, we show that the nature of the difficulty in creating an accurate representation of property ownership has shifted in the past decade to ownership obscurity caused by the rise in LLC usage. To do so, we incorporate linking to corporate records to demonstrate the added accuracy and ability this method has to uncover ownership obscurity. Ownership obscurity is both a methodological challenge in understanding property ownership and an intentional strategy worthy of further study. We show the first quantification of ownership obscurity and show that this phenomenon is becoming increasingly more common and obscures increasingly higher levels of property consolidation. We further show how ownership obscurity also covers concentrations of disorder and evictions – complicating the study of rental housing quality and landlord-tenant relationships. In sum, we demonstrate a comprehensive method for uncovering ownership obscurity and show the utility of using this method in understanding the state of rental housing.

Methods & Data

Data

We utilize a large-scale set of administrative data from Boston, Massachusetts between 2004-2019. For the current study, we utilize two main sources of data to understand property ownership: 1) Tax Assessment records from 2004 to 2019, 2) Corporate Filings from the Massachusetts Corporate Database. We further utilize eviction records from 2015-2016 and 311 reports from 2010 to 2019 to show the effects ownership obscurity and record linkage have on downstream results. These data are organized through the Boston Area Research Initiative's Geographical Infrastructure, a database that geographically links administrative records for Boston at 17 nested geographic scales (Zoorob et al. 2021).

Ownership Data

We use property-level Tax Assessment records from the City of Boston from 2004 – 2019 as the base ownership data (Shields et al. 2019). Relevant to the current analyses, tax assessment records contain the property owner's name and address as well as a parcel-level identifier. While these data encompass all properties in Boston, we utilize only residential rental parcels in our analyses to focus on linking landlords.[1] These include 700,675 parcels over the 15-year span. To uncover ownership obscurity, we use corporate filing records obtained from the Massachusetts Corporate Database.[2] These filings contain the name of a corporate entity (e.g., Eagle Properties LLC) and the names, addresses, and unique identification numbers for each individual involved in the ownership and management of the entity. These data are stored as separate databases. First is a database of corporations, each with a unique id. Second is a database of individuals and the ids of the corporations in which they are involved. When linked to one another by these ids, we can link LLCs in the tax assessments based on their shared individual members.

[1] We define residential rental parcels as those containing a single property with a residential property type and one that either is not owner occupied or is owner occupied but has more than 1 unit. This excludes condominiums that might be rentals as multiple condominium properties can be contained within a single parcel.

[2] Obtained by request to the MA Secretary of the Commonwealth Corporations Division.

Outcomes

We examine two property-level outcomes to demonstrate how obscurity can generate misleading conclusions and how record linkage can help clarify the distribution of outcomes across landlords and their properties. First, we look at 311 reports related to housing issues per unit at each parcel-year from 2010-2019 as a measure of issues experienced at parcels (e.g., unsatisfactory living conditions, pest infestation, etc.; O'Brien, Hangen, and Ristea 2022). Second, we examine eviction court filings for Boston per parcel-year from 2015-2016. Eviction filings were collected and digitized by the Office of Housing Stability in the Department of Neighborhood Development for the City of Boston.

Linking Process

We use both deterministic and probabilistic record linkage methods to link landlords together. As mentioned above, record linkage among property owners comes with multiple challenges that we seek to overcome through our linkage process. The ideal would be to compare the results of our matching process to some ground truth, however no such ground truth exits. As such, we have to make subjective decisions throughout this process as to what constitutes a match. As much as possible, we adopt a conservative approach and prioritize minimizing false positives. As ownership can change between years, we only make within-year matches. The full code for this process can be found at: https://github.com/forrest-h/linking-landlords. Our process is as follows, with examples of links of both individuals and corporations at each step to illustrate[3]:

[3] While all owner names are obtained from public records, we have chosen to preserve individual anonymity in this publication. To do so, individual names and their associated addresses in examples are fictitious but representative of real individual matches. Corporations and their addresses are provided are not anonymized.While all owner names are obtained from public records, we have chosen to preserve individual anonymity in this publication. To do so, individual names and their associated addresses in examples are fictitious but representative of real individual matches. Corporations and their addresses are provided are not anonymized.

Deterministic Linking

Deterministic links are links that follow a set of prescribed rules. The most basic form of deterministic links are exact matches. When we link owners based on exact matches between uncleaned name and addresses from the tax assessment data, 72,703 parcels are matched, corresponding to 23,640 (3.63%) owners. This highlights the need for more in-depth record linkage methods as it is highly unlikely that the other 627,972 owner entities in the tax assessment records are truly unique. This step also involves our first subjective decision – the use of names vs. names and addresses. For example, when using just names, two individuals named Cameron Robinson would be linked together, even if their addresses are in Boston, MA and San Diego, CA. Given this issue, we chose to utilize both owner names and address for exact matches. While using just uncleaned names does reduce the unique entities from 651,612 to 609,939 owners, we think it is likely many of these matches are incorrect – especially for individuals.

To improve exact match linking, we subject each dataset to a rigorous text cleaning and standardizing process. As with most administrative data, mistakes and misspellings can occur (e.g., Boston vs. Bston). In addition, abbreviations may not be standardized (e.g., corp. vs. corporation). All of these errors can affect the potential matches and links generated through both deterministic and probabilistic linking methods. We clean and standardize the text by removing any punctuation, extra spaces, converting all numbers to Arabic numerals, and adjust common misspellings (e.g., CORP, CRP, CP, and CORPORAITON are all corrected to CORPORATION). The full code and process for cleaning can be found in the online supplemental materials. We then make exact matches based on the cleaned name and address (e.g., Indiv: Ruby Coleman, 55 Huntington St. & Ruby Coleman, 55 Huntington St.; Corp: Trust Land Trust LLC, 404 S Huntington Ave. & Trust Land Trust LLC, 404 S Huntington Ave.). This reduces the unique entities from 651,612 to 640,215. However, there likely are still many missed matches.

Probabilistic Linking

After all deterministic links were made, we used two different probabilistic methods to identify possible matches. First, we used the Dedupe python library, which uses a semi-supervised machine learning algorithm to identify matches of potential duplicates (Gregg and Eder, n.d.). Generally, Dedupe is predicting possible matches based on a string similarity metric and uses human-coded pairs of matches or unrelated pairs to improve the weighting of this similarity metric. This process involves three steps. First is an unsupervised training stage in which Dedupe provides pairs of possible matches to be rated by an individual as matches or not. Also known as active learning, this stage helps improve the accuracy of the weights. Second, using these identified pairs as training data, Dedupe classifies and matches possible matches, assigning each potential match a probabilistic confidence of matching (0 being not a possible match and 1 being an exact match). Lastly, a decision needs to be made on the cut-off threshold. Using a low threshold will result in many false positives, such as Jerry Rodriguez and Martin Rodriguez being paired together. We used the training data to generate performance metrics and chose the threshold using the lowest value that returned a false positive rate of less than 1%.[4]

[4] While generally we limited matching to within-year, doing this beforehand in Dedupe would mean generating a different set of weights for each year, resulting in potentially different sets of matches across years. We therefore chose to use all years in the same round of Dedupe and limited matches to within-year matches after the classification stage. This results in consistent matches across years.

We tried two versions of Dedupe. The first matched just on names, while the second matched on names and addresses. While most of the matches based solely on names retained a high degree of face validity, we found that Dedupe consistently mismatched names that were similar but clearly not matches. For example, Elmer H. Gill and Elmer G. Gill would be linked together. When raising the threshold Dedupe would then miss many matches that had a high degree of face validity. While the pairs of linked landlords added is consistently higher across years, these pairs tend to be ones that are textually similar but have low face validity. When linking by names and addresses, Dedupe tended to miss matches with high face validity but with different addresses. Given these errors we chose to not use Dedupe to identify matches.[5]

[5] We did find Dedupe useful for identifying potential additions to our text standardization process as looking at the matched pairs can help to identify patterns that can be coded into text cleaning rules. For example, as seen in Figure 4, there is a large spike in added pairs provided by Dedupe after 2008. This spike is caused by foreclosures and links between banks-owned properties in the wake of the 2008 financial crisis. Initially, our cleaned names missed many variations used by banks (as their address may vary based on branches and their names are often slightly varied – for example, Deutsche Bank had 30 separate uncleaned names and 302 separate uncleaned names and addresses) and using Dedupe linking solely on names we identified these patterns and then included a separate text cleaning for banks. We therefore recommend the use of Dedupe and other out-of-the box algorithms not as a production-level identifier of matches but as a check in the process for identifying ways to improve deterministic matches.We did find Dedupe useful for identifying potential additions to our text standardization process as looking at the matched pairs can help to identify patterns that can be coded into text cleaning rules. For example, as seen in Figure 4, there is a large spike in added pairs provided by Dedupe after 2008. This spike is caused by foreclosures and links between banks-owned properties in the wake of the 2008 financial crisis. Initially, our cleaned names missed many variations used by banks (as their address may vary based on branches and their names are often slightly varied – for example, Deutsche Bank had 30 separate uncleaned names and 302 separate uncleaned names and addresses) and using Dedupe linking solely on names we identified these patterns and then included a separate text cleaning for banks. We therefore recommend the use of Dedupe and other out-of-the box algorithms not as a production-level identifier of matches but as a check in the process for identifying ways to improve deterministic matches.

We chose to use a more customized form of probabilistic matching to minimize the false positives generated. Using the cleaned names and addresses, we matched based on the 3-gram cosine similarity. First, we generated the 3-grams of the combined names and addresses for each year (e.g., the 3-grams of ' text ' are [' te', 'tex', 'ext', 'xt ']). We then generated the term frequency-inverse document frequency matrix for the 3-grams (TF-IDF). This step converts each name and address to a vector representing the product of the term-frequency (i.e., how frequently a 3-gram appears in the corpus) and the inverse document frequency (i.e., the commonness or rarity of the 3-gram across all names and addresses). We then computed the cosine similarity of these vectors and found the 3 closest matches for each unique cleaned name and address. We then decided on a threshold that minimizes false positives. We used a threshold of 0.85. We found this resulted in matches with a high degree of face validity (e.g., Melissa G King, 15 Danvers St. & Melissa King 15 Danvers St.) and minimized false positives. These fuzzy-matched names provide more added information than just cleaned names, but without the increased addition of low face validity pairs generated by Dedupe. This reduced the number of unique owners from 640,215 after the cleaning step to 616,929 unique owners.

Linking Corporations

At this point we have linked landlords deterministically by cleaning and standardizing their names and probabilistically using a text similarity metric. However, there are still many potential pairs that are missed by only linking based on text similarity. For example, two LLCs with vastly different names (e.g., 123 Land Trust LLC and Rental Properties LLC) could be owned and operated by the same individual. Without added information, no form of text matching will link these two entities together. This is why we use corporate records to link corporate entities together.

We first linked the names of probable corporations to the database of corporations, using the same deterministic and probabilistic linking processes as above. This provides the id of each corporation in the tax assessment records (16,121 entities are linked). We then limited the database of individuals to those associated with the linked ids and use the same deterministic and probabilistic matching process to deduplicate individuals in the data based on their names and both residential and business addresses.[6] Next, we generated networks of linked corporations based on their shared individual members with corporations as nodes and edges representing a shared individual between both corporations. For example, if Christopher Long is involved in 10 corporate entities with 5 other individuals, the corporations of those other 5 individuals are all linked together. After pruning highly connected individuals,[7] we use the connected components of these networks to create unique ids for each set of corporations. Combined with the text-based matches of names, this is the final set of linked landlords. In the end, the linking process reduced the number of unique owners from 651,612 to 605,731.

[6] Some have residential, others business, this reduces false positives.

Results

Linking Process

In the uncleaned tax assessment, there are an average of 40,726 owners per year from 2004 to 2019. Even at this stage, there is evidence of increasing consolidation, as the number of unique uncleaned owners (defined by both a named entity and the associated address) decreased from 42,466 in 2004 to 39,502 in 2019 while at the same time the number of rental units increased from 159,038 to 175,366. Figure 1 illustrates the entire linking process from uncleaned tax assessment records to a final set of linked owners and properties (percentages shown for 2019 only). In the base uncleaned tax assessment data, 4.8% of owners of individual parcels are linked together – in other words in the base data, only 4.8% of owners own multiple parcels (see Step 0 in Figure 1). In Step 1: Data Cleaning of Tax Assessment, 2.3% of owners had changes that linked them to other owners in 2019 for a cumulative 7.1% of owners, 15.6% of parcels, and 27.8% of units with matches. After a customized stage of probabilistic matching (Step 2: Fuzzy-matching within Tax Assessment in Figure 1), 12.3% of owners in 2019 were linked to other owners. Finally, when corporate entities in the tax assessment records are fuzzy-matched to corporation records and then linked to one another based on common individuals involved in the ownership and management of corporations, 16.7% of owners in 2019 are linked to other owners. While 16.7% of owners being linked to other owners might initially seem like a small group, they have an outsized effect on the rental housing landscape, owning 24.3% of parcels and 42.0% of rental units in 2019. This begins to indicate increased consolidation of rental property ownership and the obscurity that covers this consolidation.

Dedupe vs. Custom Fuzzy-matching

While an out-of-the-box algorithm like Dedupe has the benefits of being relatively easy to use, we found this simplicity actually complicated the process of getting accurate links. As seen in Figure 2, the two Dedupe iterations and the customized fuzzy-matching step share only 2,142 matches (53.2% of the customized fuzzy-matches). These are the most likely true positives as all three methods identified these matches. An additional 1,374 matches were shared between customized fuzzy-matching and one of the Dedupe iterations meaning that only 510 matches were unique to the customized-fuzzy matching. This convergence suggests that the customized fuzzy-matching process aligns with our conservative approach to limiting false positives.

In contrast, both Dedupe iterations have over 4,000 unique matches each – which suggests large numbers of likely false positives. When we used Dedupe on names only (DNO; see Dedupe: Name Only in Figure 2), we found that Dedupe linked 22.7% of owners in 2019 -an increase of over 10% compared to the customized fuzzy-matching stage. However, many of these additional matches would best be classified as false positives. DNO has 5,163 unique pairs when compared to Dedupe using names and addresses (DN&A) and our customized probabilistic fuzzy-matching stage. For reference, here are three random matches[8]:1 1) Curtis Gray P, 27 Green St, Boston MA with Curtis Gray P JR, 10 Blue St, Boston MA. 2) OXBOW URBAN LLC 45 Red Ave, Dorchester MA with OXBOW URBAN LLC 165 Green Ave, Dorchester MA. 3) Glenn Ross 4 East St, Boston MA with Glenn Ross 87 Blue Ave, Boston MA. In the first and last example, it is likely that these refer to different people as either they have common names but are at different addresses or they have the potential to be a father and son pair. In the second example, this is a likely true positive. While the fuzzy-matched stage won't link these entities, the corporations data will link these two instances of the same LLC together. While DN&A generated only a slightly higher proportion of owners as the fuzzy-matching stage (16.6% in 2019) these 4,406 unique pairs can also be best classified as false positives and show similar patterns to the above examples for DNO.

As further evidence that Dedupe consistently over-estimates matches between owners, we look at the similarities and differences to the final set of corporation links (see Figure 3). If Dedupe is capturing an accurate links above the custom fuzzy-matching approach, we would expect to see

[7] For many individuals, the sets of corporations of their co-owners are the same or very similar to their own sets (e.g., Brenda Flores shares the same 3 corporations with Jose Scott). However, for some well-connected individuals, their shared set of corporations might be very different from the majority of their own corporations. If left unresolved, this can generate very large networks that are likely not acting together to own and manage the set of properties owned by all involved. For example, Timothy Parker is part of a group of 10 individuals who own 10 corporations. They also have one shared corporation with Jennifer Moore who is a part of a different group of individuals who own a different set of corporations. To account for this, we use an iterative process of pruning in which we remove these articulation points and their edges and add back in the nodes to their most common set of corporations. In the above example, the shared corporation between Timothy Parker and Jennifer Moore would be removed from the overall network add added back into the connected component that has the larger number of corporations in common. This reduces the number of false positive links between corporations.

[8] For privacy, individual names and addresses have been changed but the differences between names have been preserved. For example, if the original pair was John H, 1 First St. Boston, MA with John G, 12 First St, Medford, MA then the privacy-changed pair would be Ben M, 5 Green St, Boston, MA with Ben N, 52 Green St, Medford, MA.

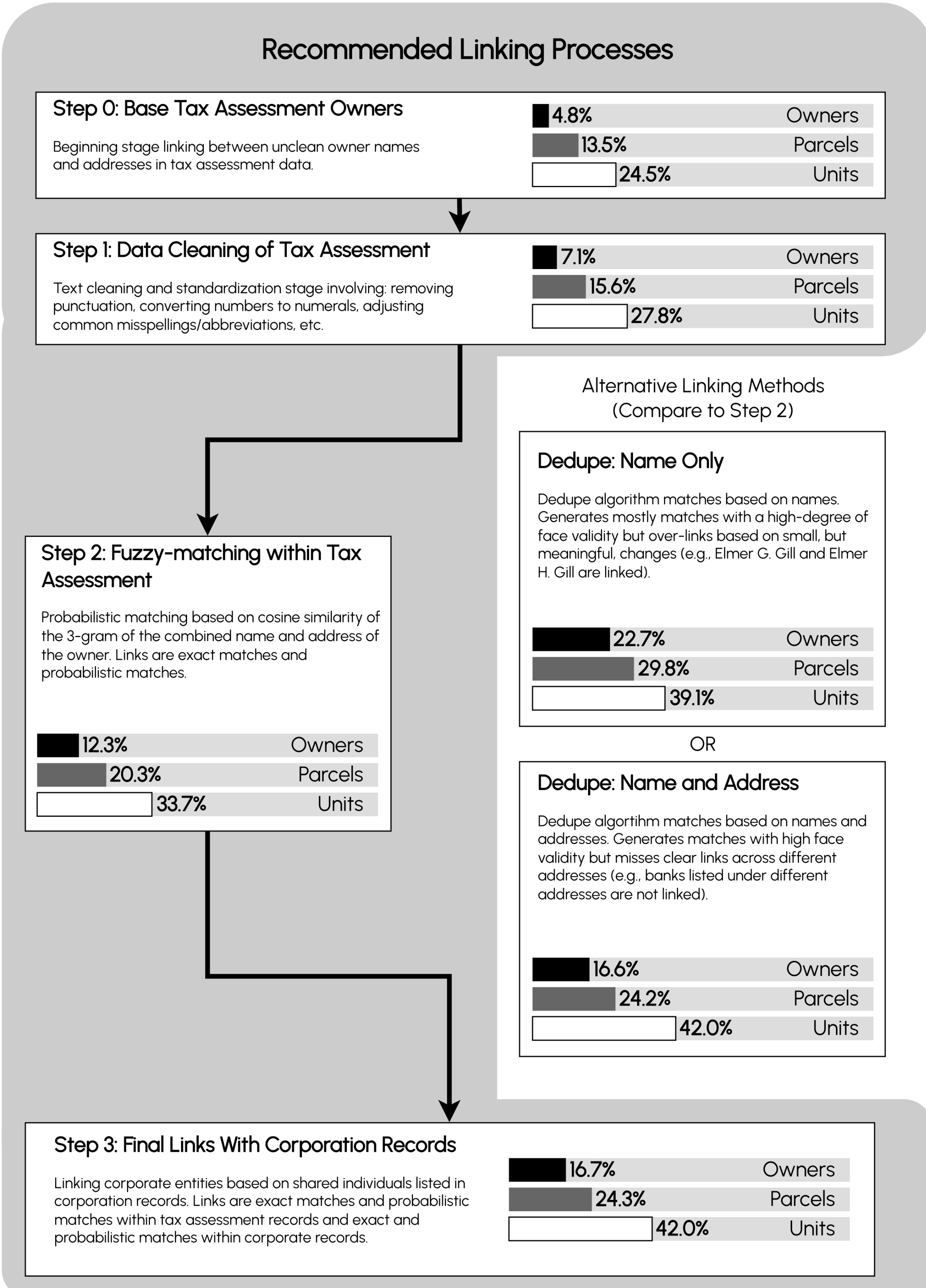

Figure 1. Recommended Linking Process
Note: Owners, Parcels, and Units are for 2019 only.

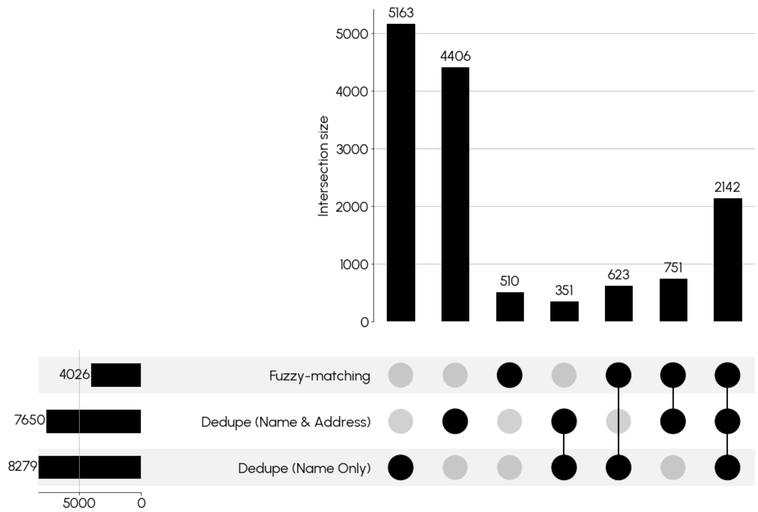

Figure 2. Set Overlaps Between Dedupe and Custom Fuzzy-Matching Methods
Note: Black dots indicate which method(s) are involved in the above bar. For example, the first black dot from the left indicates that only Dedupe (Name Only) has 5,163 unique pairs of original owners, while the last connected black dots on the right indicate that all 3 methods share 2,142 pairs.

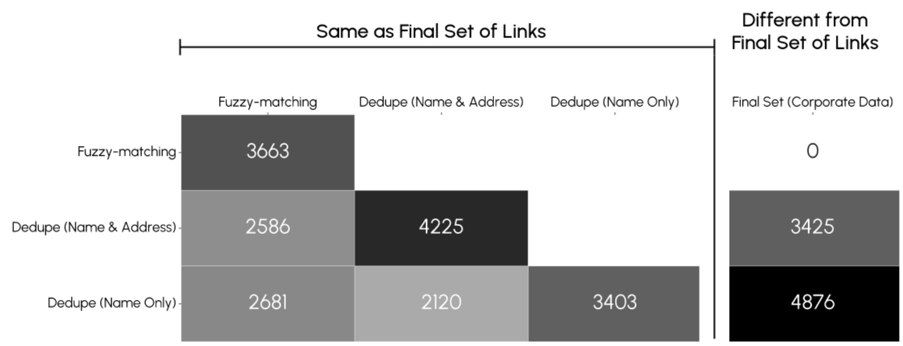

Figure 3. Probabilistic Linking Methods Compared to Final Set of Matches

---

these matches overall significantly with the final set of matches using corporate data. While there is significant overlap between these matches, there are also significant amounts of differences, with both version of Dedupe having over 3,400 matches that are not in the final set of matches.

Ownership Obscurity over Time

We find a significant increase in ownership obscurity since 2010. As seen in Figure 4, the pairs of linked landlords generated by the addition of corporate data has greatly increased since 2012. By 2019, the number of pairs added is over 4 times that of the fuzzy-matching process. This highlights both the increasing obscurity of corporate ownership and the increasing need for linking methods that incorporate corporate data. While the cleaning and fuzzy-matching process indicates there are 37,328 landlords in 2019, the addition of corporate data shows that there are more likely 35,639 landlords – a reduction of 1,689 (4.5%) fewer landlords.

As noted above, while Dedupe Name Only generated more pairs in 2004-2013 than any other method, many of these can best be described as false positives. Furthermore, DN&A failed to pick up on the consolidation of bank-owned properties in the wake of the 2008 financial crisis as seen by a lack of a spike in pairs of connected parcels in Figure 4. This is largely due to banks that list their branches at different addresses. Thus, while we tested the use of a semi-supervised learning algorithm (Dedupe) into the process of linking landlords, we ultimately decided a more customized

Figure 4. Ownership Obscurity over Time

---

probabilistic method was better suited to the current use case.

Effect on Downstream Results

To demonstrate the effect of obscurity on our understanding of landlords and their properties across linking methods, we look at 3 relevant outcomes: the consolidation of units, evictions, and housing-related property issues. For each outcome of interest, we look at the Herfindahl-Hirschman Index (HHI) as a measure of the consolidation of the relevant outcome, here calculated as:

$$HHI = \sum_{i=1}^{N} S_i^2$$

where S = a landlord's overall share of an outcome

Conceptually, as the HHI approaches 1, this indicates a higher level of overall consolidation where fewer owners are responsible for the relevant outcomes. For example, a higher HHI of rental units indicates that some owners own an outsized share of units.

First, we look at the consolidation of rental units over time (see Figure 5). While the 3 steps prior to linking with corporate data show some increases over time in the consolidation of rental units over time (a higher HHI in 2019 vs. 2004), using corporate data consistently shows both a higher overall level of consolidation across years and a steeper increase in consolidation over time. For context, using the final set of links with corporate data shows that while in 2004, the largest 1% of owners (n = 402) owned 29.3% of units (46,581 units), in 2019 the share of units owned by largest 1% of owners (n = 356) rose to 34.2% (59,957 units). In contrast, fuzzy-matching (without using corporate data) shows only a minor increase in the share of units owned by the largest 1% of owners: in 2004 (n = 404) they owned 28.0% of units (44,603 units) and in 2019 the largest 1% of owners (n = 373) owned 30.5% of units (53,438 units). Taken together with the results of Figure 5, this illustrates the increasing level of consolidation that is only revealed by using corporate data.

Finally, we compare linking stages across two outcomes relevant to understanding landlords and their management of properties. First, we look at the concentration of 311 reports of housing

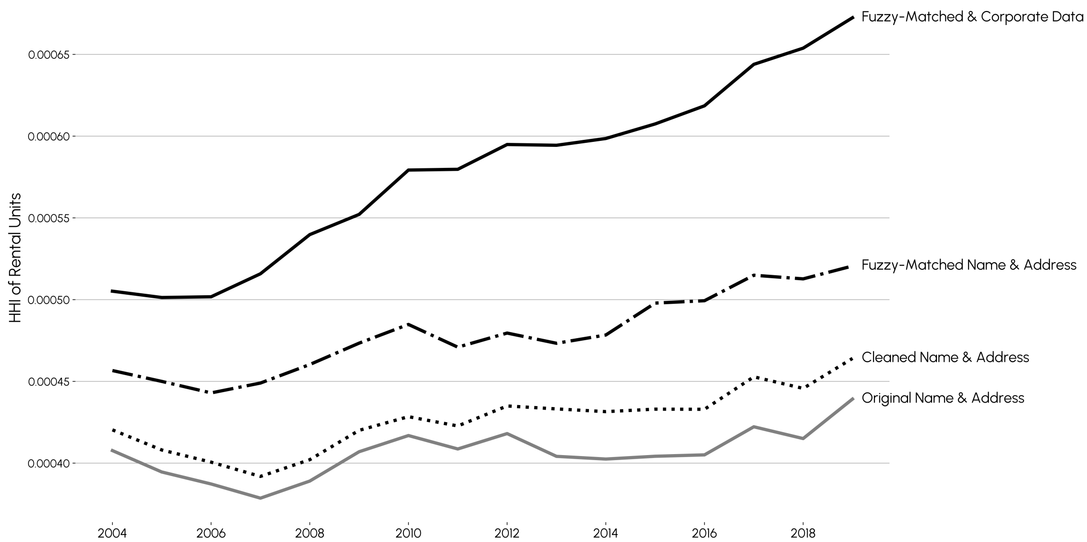

Figure 5. Consolidation of Rental Units

---

related issues (e.g., Unsatisfactory living conditions, Insufficient Heat, Lead, etc.). We look at the HHI of housing issues by calculating a landlord's share of parcels with housing issues (to control for consolidation caused by larger buildings). As seen in Figure 6, the HHI of housing issues is significantly larger after incorporating corporate data – indicating a larger concentration of housing related issues compared to other linking methods. For comparison, in 2019 the final set of links using corporate data indicates that the largest 1% of landlords (n = 356) own 36.1% of the 2,036 parcels with housing related issues while linking without corporate data only shows that the largest 1% of owners (n = 373) own 30.2% of the 2,046 parcels with 311 reports of housing issues. This highlights the dampening effect obscurity has on understanding disparities in property management strategies.

Lastly, we examine the effect of obscurity on the responsibility for eviction filings. As seen in Figure 7, the final set of links with corporate data shows the highest level of concentration of eviction filings across owners. For context, in 2016, the largest 1% of landlords (n = 366) were responsible for 48.9% of the 2,053 parcels with evictions. Whereas linking without corporate data the largest 1% of landlords in 2016 (n = 376) were responsible for 43.8% of the 2,053 parcels with evictions. These results suggest that corporate obscurity dampens the disproportionate eviction rates of larger landlords.

Discussion

The challenges to creating an accurate representation of property ownership are twofold: administrative data ripe with errors and intentional ownership obscurity. We compare the efficacy of methods to overcome these two methodological challenges and demonstrate a comprehensive method for creating an accurate answer to the question: "Who owns rental housing?" In doing so, we have shown that: 1) Ownership obscurity is an increasingly common phenomenon that dampens the disparities between landlords and the disparate outcomes of their portfolios of properties. 2) Not all linking methods are equal – only using corporate data is ownership obscurity fully uncovered. Given these two findings, we strongly recommend that future research involving landlord size, comparing outcomes of landlords, and examining the practices of landlords involve the use of corporate data and that care is taken to accurately link these data while minimizing false positives.

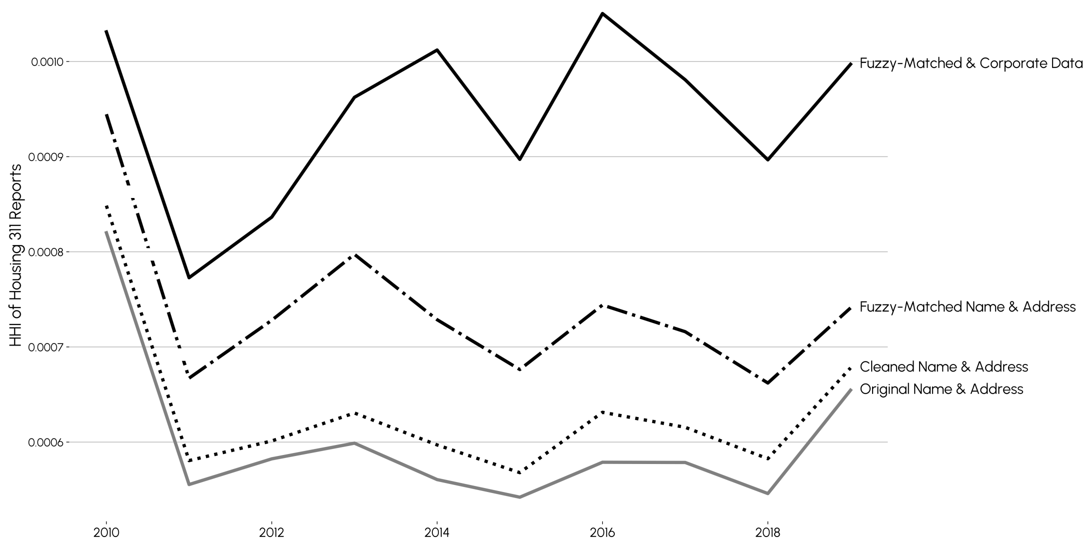

Figure 6. Consolidation of Housing Issues

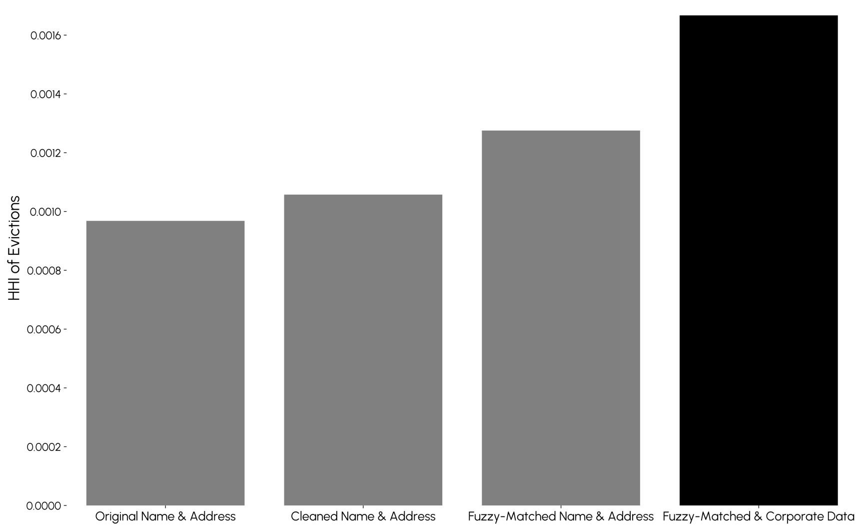

Figure 7. Consolidation of Evictions

Text-matching in Administrative Data

The first obstacle to creating an accurate representation of property ownership is record linkage. Entities in tax assessment or other administrative records need to be correctly and accurately linked. We took a conservative approach to linking – prioritizing a low potential false positive. While out-of-the-box methods, like Dedupe, are generally easier to use and do create a significant number of likely true positive matches, we found their level of probable false positives to be too high. One of the potential reasons for this is that there are some patterns an algorithm is likely to pick up on that make entities statistically similar but qualitatively very different. For example, a middle initial has significant meaning but statistically makes two entities only one letter shift away from one another. This is likely why using Dedupe on names only rendered so many likely false positives. In our more customized fuzzy-matching method, we are able to fine-tune the methods to the current use case and more accurately account for these small but significant difference in limiting false positives.

The second significant obstacle is the rise of intentional ownership obscurity. We found that in the past decade there has been an increase in the number of corporate entities used in tax assessment records and that the use of these entities significantly obscures an increasing consolidation of properties. The use of corporation records in our linking process was the key to uncover this intentional ownership obscurity. By linking LLCs whose names and addresses would never be connected by purely text-matching but who were owned and operated by the same set of individuals, we were able to overcome this significant obstacle.

Ownership Obscurity

We found increasing ownership obscurity in the past decade. While this does indicate that discovering ownership is becoming more difficult, the fact that we found this trend indicates that there is hope for uncovering this obscurity. We have demonstrated a scalable process that can uncover this obscurity while maintaining a high degree of face validity and avoiding many false positive links between landlords. Being able to uncover ownership obscurity has both research and practical applications. For research purposes, one can gain a more accurate understanding of the role landlords, their characteristics, and nature and size of their real estate portfolio can play in the conditions of properties, treatment of tenants, and the operation of the rental market and stock. In addition, by uncovering systematic error in the tax assessment records we have revealed an important process worthy of study itself (O'Brien 2018). While some ownership is initially hidden due to the error-prone data generation process, we have revealed systematic, intentional obscurity that reveals a strategy of property ownership. In other uses, tenant organizers and unions can help build collective power across buildings – helping tenants find others who live in buildings owned by their obscured landlords. A notable examples of this include the Anti-Eviction Mapping Project's Evictorbook – an online tool designed to help organizers and tenants research owners and their properties ("Evictorbook" n.d.).

Our results indicate that ownership obscurity hides a disproportionate share of rental property ownership, responsibility for disorder and disrepair at rental properties, and eviction filings. While more research is needed to explore the role between obscurity and strategies of property management, these findings begin to indicate that landlords may use obscurity to avoid responsibility for the physical disorder at their properties. While properties owned by the largest obscured owners had a disproportionate share of violations, this lack of physical upkeep was accompanied by an increased social upkeep in the form of a disproportionately large share of eviction filings and a higher eviction filing rate. More research is needed to explore the different strategies revealed as we uncover ownership obscurity.

Recommendations for Future Research

One limitation of the current work is the lack of an objective ground truth or the known true matches between LLCs and other entities. This lack of ground truth means that we cannot objectively measure performance by comparing generated matches to true matches. As such, we prioritized a conservative approach and sought to create matches with a high degree of face-validity. This means that our approach is just one approach to linking landlords and can likely be improved in future research. For example, if a source of ground truth emerges or is created (perhaps by examining a sample of landlords and their internal records of their property ownership), one could likely better fine-tune the current methods and approaches to catch errors. Ultimately, while we are confident that the current methods are generating a fairly accurate representation of property ownership, we recognize that future research could improve these methods and we recommend each. We do offer the following recommendations for future research seeking to uncover property ownership and ownership obscurity.

We found that the most accurate process to uncover ownership obscurity involves rigorous text cleaning, deterministic linking, customized probabilistic linking, and the use of corporate records. We recommend avoiding the use of 'black-box' probabilistic linking methods in the final linking process as their mistakes can create an inaccurate form of uncovered obscurity. These methods are more suited for exploring systematic errors in text cleaning that can be then incorporated into rule-based text cleaning and deterministic linking processes. In addition, we found that only by incorporating corporation records into the process were we able to uncover ownership obscurity. Given the trends in rise of LLC use, the use of corporation records is likely to be increasingly important for creating an accurate representation of property ownership. We therefore recommend investing time in developing a rigorous and scalable process that involves a customized probabilistic fuzzy-matching process and the use of corporate records. To aid in this effort, we have made all code available in the supplemental resources.

While the most accurate process to uncover obscurity involves connecting ownership to corporate records, there may be cases when corporate records are unavailable or prohibitively expensive to obtain. In these cases, even the best fuzzy-matching process will miss links between corporations. In these analyses, researchers should take care in interpreting results, especially as they relate to the size and composition of landlords' real estate portfolios. In such cases, additional manual linking may be needed to compensate for the lack of corporate information. For example, searching the internet for mentions of the largest landlords might return connections between some entities beyond their cleaned and fuzzy-matched names and addresses. This may improve the accuracy of linking but can be a large investment of time and personnel. As such we recommend obtaining corporate records whenever possible.

Conclusion

The rise of LLCs and other corporate entities as landowners has led to an increase in ownership obscurity. Ownership obscurity can misrepresent the size of a landlord's portfolio and their responsibility for property-level outcomes. While obscurity is on the rise, our understanding of the processes and data needed to uncover this obscurity are also increasing. Through combination of deterministic and probabilistic linking methods connecting ownership information to corporate records, this obscurity can be uncovered. However, the quality and accuracy of the linking process can differ based on the linking methods used. Using the methods and processes outlined in this work, we hope future work will be able to uncover obscurity and show a more accurate picture of landlords and their properties.

Acknowledgment

This preprint was created using the LaPreprint template (https://github.com/roaldarbol/lapreprint) by Mikkel Roald-Arbøl.

Author contributions

Forrest Hangen is a PhD student at in the School of Public Policy and Urban Affairs at Northeastern University. His research focuses on housing policy, landlord strategies of social and physical property management, and social inequalities.
Website: forresthangen.com
Twitter: @forresthangen
Mastodon: @forresthangen@sciences.social

Daniel T. O'Brien is an associate professor in the School of Public Policy and Urban Affairs and the School of Criminology and Criminal Justice at Northeastern University and the director of the Boston Area Research Initiative. His work focuses on the ways that researchers, policymakers, and practitioners can work together to leverage modern digital data to better understand and serve cities.

Supplementary Materials

Code for data cleaning, standardization, fuzzy-matching, and linking to corporate records is available at: https://github.com/forrest-h/linking-landlords. All data except Corporation Records available at the Boston Area Research Initiative's Data Portal. MA Corporation Records available upon request.

References

An, Brian, Andrew Jakabovics, Anthony W. Orlando, and Seva Rodnyansky. 2022. "Who Owns Urban America? A Methodology for Identifying Real Estate Owners." SSRN Scholarly Paper. Rochester, NY. https://doi.org/10.2139/ssrn.4156785.

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. https://doi.org/10.1086/701697.

Enamorado, Ted, Benjamin Fifield, and Kosuke Imai. 2019. "Using a Probabilistic Model to Assist Merging of Large-Scale Administrative Records." American Political Science Review 113 (2): 353–71. https://doi.org/10.1017/S0003055418000783.

"Evictorbook." n.d. Accessed December 2, 2022. https://evictorbook.com.

Fellegi, Ivan P., and Alan B. Sunter. 1969. "A Theory for Record Linkage." Journal of the American Statistical Association 64 (328): 1183–1210. https://doi.org/10.1080/01621459.1969.10501049.

Garboden, Philip ME, and Eva Rosen. 2019. "The Threat of Eviction: How Landlords Shape a Contingent Tenure," 38.

Gomory, Henry. 2021. "The Social and Institutional Contexts Underlying Landlords' Eviction Practices." Social Forces, June, soab063. https://doi.org/10.1093/sf/soab063.

Gregg, Forest, and Eder, Derek. n.d. "Dedupe." https://github.com/dedupeio/dedupe.

Herzog, Thomas H., Fritz Scheuren, and William E. Winkler. 2010. "Record Linkage." WIREs Computational Statistics 2 (5): 535–43. https://doi.org/10.1002/wics.108.

Immergluck, Dan, Jeff Ernsthausen, Stephanie Earl, and Allison Powell. 2020. "Evictions, Large Owners, and Serial Filings: Findings from Atlanta." Housing Studies 35 (5): 903–24. https://doi.org/10.1080/02

Newcombe, Howard B, James M Kennedy, SJ Axford, and Allison P James. 1959. "Automatic Linkage of Vital Records: Computers Can Be Used to Extract" Follow-up" Statistics of Families from Files of Routine Records." Science 130 (3381): 954–59.

O'Brien, Daniel T., Forrest Hangen, and Alina Ristea. 2022. "311 Requests." Harvard Dataverse. https://doi.org/10.7910/DVN/CVKM87.

O'Brien, Daniel T. 2018. The Urban Commons: How Data and Technology Can Rebuild Our Communities. Harvard University Press.

O'Brien, Daniel T., Alina Ristea, Forrest Hangen, and Riley Tucker. 2022. "Different Places, Different Problems: Profiles of Crime and Disorder at Residential Parcels." Crime Science 11 (1): 4. https://doi.org/10.1186/s40163-022-00165-0.

Raymond, Elora Lee, Richard Duckworth, Benjamin Miller, Michael Lucas, and Shiraj Pokharel. 2018. "From Foreclosure to Eviction" 20 (3): 33.

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, November, 1–24. https://doi.org/10.1080/10511482.2020.1822903.

Seymour, Eric, and Joshua Akers. 2021. "Building the Eviction Economy: Speculation, Precarity, and Eviction in Detroit." Urban Affairs Review 57 (1): 35–69. https://doi.org/10.1177/1078087419853388.

Shields, Michael, Saina Sheini, Justin de Benedictis-Kessner, and Daniel T. O'Brien. 2019. "Property Assessment." Harvard Dataverse. https://doi.org/10.7910/DVN/YVKZIG.

Sternlieb, G. 1966. The Tenement Landlord. Urban Studies Center, Rutgers, State University. https://books.g

Tahamont, Sarah, Zubin Jelveh, Aaron Chalfin, Shi Yan, and Benjamin Hansen. 2021. "Dude, Where's My Treatment Effect? Errors in Administrative Data Linking and the Destruction of Statistical Power in Randomized Experiments." Journal of Quantitative Criminology 37 (3): 715–49. https://doi.org/10.1007/s10940-020-09461-x.

Thakuriah, Piyushimita (Vonu), Nebiyou Y. Tilahun, and Moira Zellner. 2017. "Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery." In Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics, edited by Piyushimita (Vonu) Thakuriah, Nebiyou Tilahun, and Moira Zellner, 11–45. Springer Geography. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-40902-3_2.

Travis, Adam. 2019. "The Organization of Neglect: Limited Liability Companies and Housing Disinvestment." American Sociological Review 84 (1): 142–70. https://doi.org/10.1177/0003122418821339.

Tromp, Miranda, Anita C Ravelli, Gouke J Bonsel, Arie Hasman, and Johannes B Reitsma. 2011. "Results from Simulated Data Sets: Probabilistic Record Linkage Outperforms Deterministic Record Linkage." Journal of Clinical Epidemiology 64 (5): 565–72.

Zoorob, Michael, Alina Ristea, Saina Sheini, and Daniel T. O'Brien. 2021. "Geographical Infrastructure for the City of Boston v. 2021." Harvard Dataverse. https://doi.org/10.7910/DVN/ZHTMIW.